‘Catalytic nuclear war’ in the age of artificial intelligence & autonomy: Emerging military technology and escalation risk between nuclear-armed states

ABSTRACT

This article revisits the Cold War-era concept of ‘catalytic nuclear war,’ considered by many as unworkable, and reconceptualizes it in light of technological change, as well as improved understanding of human psychology and other factors. It argues in the modern digital era, the catalyzing chain of reaction and counter-retaliation dynamics set in motion by the deliberate action of a non-state or third-party actor is fast becoming a more accessible and plausible alternative to acquiring a nuclear weapon or manufacturing an improvised atomic device – or ‘dirty bomb.’ The article concludes that artificial intelligence (AI) technology is creating new – and exacerbating old – escalation pathways that risk catalyzing accidental nuclear confrontation between nuclear-armed powers, particularly under irrational (or sub-rational) conditions. Are existing notions of accidental and inadvertent nuclear escalation still relevant in the age of AI and autonomy?

Introduction

The main concern about ‘catalytic nuclear war’ – and ‘Nth country problem’ more broadly – in the late 1950s and early-1960s centered on the fear that a small or new nuclear power would deliberately set a major exchange in motion between the United States and the Soviet Union.[1] As its name suggests, the concept was inspired by chemical reactions where the catalyzing agent (a third-party actor) would remain unscathed by its initiated process.[2] In his seminal work on ‘catalytic nuclear war’ in 1962, Donald Kobe defines the notion of ‘catalytic war’ as the possibility that a third party’s actions induce a nuclear war between the two nuclear-armed opponents. A third-party attack on nuclear-power would be misattributed to another state, thus sparking a catalytic ‘chain reaction of retaliation and counter-retaliation.’[3] A broader definition of ‘catalytic nuclear war’ encompasses situations in which the catalyzing actor (non-state or another third party) initiates a simulated attack (e.g., a false alert) that produces the effect of an imminent attack on one or both of the nuclear-armed adversaries, for which pre-emption is considered the most expedient (or only) strategic recourse.[4] This broadened definition, and used in this article, includes any process by which a third-party actor uses its capabilities to involve nuclear-armed strategic rivals in a war intentionally or to increase the scope or intensity of a situation.[5] The paper’s originality lies in revisiting the ‘catalytic nuclear war’ concept, considered by many as unworkable, and reconceptualizing it in light of technological change, as well as improved understanding of human psychology and other factors.

Because nuclear powers must have some policy of retaliation (i.e., massive, proportional, finite, or random) in order to act as a deterrent, this chain reaction may, under certain circumstances, become a self-fulfilling prophecy. For instance, the common fear of pre-emption caused by one or both states is vulnerable to a first strike, or if a state’s command, control, and communication structures are compromised following an attack.[6] This situation would be compounded by the presence of higher numbers of nuclear-armed states – or nuclear multipolarity – with varied, and possibly conflicting, retaliation and deterrence postures.[7] Moreover, miscalculation, misperception, or an irrational act, would also increase the risk of catalytic war.[8] At the root of these conditions is human psychology, which strains the credibility of arguments that human-rationality, perceptions, and nuanced signaling ensure that deterrence will prevail.

Classical nuclear deterrence policies, premised on the assured threat of a retaliatory second strike (or mutually assured destruction (MAD)), are a fundamentally psychological-political-technical

phenomenon that has worked to reduce the chances of deliberate use of nuclear weapons.[9] This article argues that increasingly sophisticated, interdependent, and vulnerable nuclear command, control, and communications (NC3) systems – the bedrock of the mutual deterrence regime between nuclear-armed states – increases the risks associated with unintentional catalytic nuclear war.[10] The article is not a treatise on the motives or goals of non-state motives for committing such acts, but rather the evolving means at the disposal of actors to actualize their malevolent goals – or 'apocalyptic world views' – in the digital age.[11] There is a rich body of literature on the intersection of emerging technology and the nuclear enterprise – and the potential for warfare more broadly defined.[12] However, there is a relative paucity of research on how existing concepts of escalation, nuclear terrorism, and classical deterrence theories might apply (or be tested) in the digital age, which are increasingly defined by developments in artificial intelligence (AI) and autonomy perfect information and rational decision-making cannot be assumed.[13] Are existing theories of military escalation still applicable in the era of AI and autonomy?[14]

The remainder of the article proceeds as follows. First, it contextualizes the concept of 'catalytic nuclear war' with emerging security technologies, particularly AI. The article uses this analytical framework to examine how and under what circumstances the nefarious (mis)use of these technologies by non-state actors (i.e., terrorists, criminals, and state-proxy actors) might create or exacerbate escalation pathways leads to an accidental nuclear confrontation between nuclear-armed powers. Next, in the articles' empirical core, it considers four aggravating features in the interaction of emerging technology with nuclear weapons and the pathways to nuclear escalation examined in the article – information complexity and asymmetry; greater automation of NC3; disinformation, misinformation, and information manipulation; and nuclear multipolarity.[15] This article posits that AI-enhanced capabilities will likely amplify these underlying conditions (or second-order effects), thus increasing the risk of sparking unintentional escalation between strategic nuclear-armed powers – especially in crisis and conflict conditions. The next section uses four fictional scenarios – cyber false flag operation, cyber-attack on nuclear early-warning systems, drone swarming targeting ISR systems, and deepfake disinformation – to demonstrate how AI-enhanced capabilities in the hands of non-state actors might accidentally or inadvertently drag a competitive nuclear dyad into conflict. The final section considers possible ways of reducing catalytic escalation risk caused by non-state actors, particularly how militaries can maintain effective command and control of their nuclear forces in a rapidly evolving and complex conflict environment.

Catalytic nuclear war in the digital age: Evolving nuclear strategies of non-state actors

During the later 1950s and early 1960s, the notion of a 'catalytic nuclear war' centered on the possibility that a minor nuclear power would deliberately trigger a full-scale nuclear exchange between the United States and the Soviet Union. In a seminal article on the topic, Donald Kobe argued that the notion of 'catalytic war' was inspired by chemical reactions whereby the catalyzing agent remains unchanged by the process it has triggered.[16] Several scholars also raised the specter that a minor nuclear power might be tempted to catalyze a devastating exchange that would inflict maximum damage on the superpowers but leave it relatively unscathed and thus dictate conflict termination outcome.[17]

Some scholars have questioned the plausibility of this hypothesis. Even in the unlikely event that a minor power successfully concealed its role in sparking a global war, the prospect of its plot being discovered would almost guarantee decapitation, making any potential gains outweighed by the overwhelming risks thus, illusionary.[18] In the event, improvements to nuclear early-warning systems and communication channels – and broader tactic cooperation – between the United States and the Soviets did much to reduce the (albeit remote) possibility of a minor power creating the conditions inadvertently dragged the superpowers into nuclear war.[19] What if the risk and rationality considerations mattered less to non-state actors (especially terrorists) in possession of a nuclear weapon to incite chaos and other disruptive means to fulfill its apocalyptic ends?[20]

The nuclear terrorism threat discourse began in the mid-1970s. The terrorist attacks on the United States in 2001, coupled with the 2003 exposure of a secret network (led by Abdul Qadeer Khan) supplying Libya, North Korea, and Iran with the know-how for manufacturing critical nuclear equipment, which increased the sense of urgency. In particular, concerning non-state actors (especially those motivated by extreme political or religious goals), global nuclear security regimes and approaches.[21] Experts consider the probability of the successful theft of nuclear grade materials (i.e., highly enriched uranium or plutonium) and detonation of atomic weapons or manufacturing an improvised device (or a 'dirty bomb') by non-state actors as extremely low; but it cannot be entirely discounted.[22] To be sure, experts have continued to warn about the opacity of materials safety and security in smaller nuclear-armed states such as Iran, Israel, Pakistan, and most notably, North Korea.[23]

If a non-state group realized that the possession of nuclear weapons could trigger a total catalytic war between two or more major nuclear-armed states, what possible situations might set these dynamics in motion? The potential success of a catalytic nuclear detonation by a terrorist group during a crisis between two nuclear-armed states (i.e., the original target of the nefarious attack and the victim of a subsequent inadvertent retaliatory strike), when tensions are running high, and a heightened state of alert is in place – at a nuclear or sub-nuclear level – would likely be amplified. That is, a broader nuclear exchange would likely occur. Moreover, a non-state actor might exploit a tense situation by making provocative false claims that its attack had been sanctioned by the state with which the victim of the attack was already in a crisis, or worse still, nuclear brinkmanship. Against the backdrop of fear, acrimony, and paranoia – when decision-makers are more susceptible to misperception and misunderstanding – any denials or assurances would fall on deaf ears.[24] If a repeat of the 1999 Kargil crisis or the more recent Indian-Pakistani stand-off in 2019, a non-state nuclear detonation (or the credible threat of one) could spark an accidental or inadvertent nuclear exchange.[25] Thus, any confidence Indian or Pakistani has in their ability to manage a conflict below the nuclear-level would likely be unreliable in the event of a non-state nuclear attack. Absent unambiguous and incontrovertible evidence to the contrary, an unprovoked non-state attack may compound mutual suspicions about the other side's intentions and behavior, and in turn, confirm worst-case scenario assessments.[26]

In the contemporary digitized landscape, the theft of nuclear grade materials, the detonation of atomic weapons, and 'dirty bombs' by non-state actors are not the only threats that states must worry about. Emerging technologies – most notably cyber, AI technology, and drones – are rapidly creating new (and exacerbating old) low-cost and relatively easy means for non-state actors to fulfill their nefarious goals; without the need for actual physical contact with or manufacture of nuclear weapons to have the 'power to hurt' – or the power *to get others to hurt each other*.[27] One of the critical potential 'threat vectors' (i.e., a mechanism or means by which an actor can gain access to a network in order to deliver a malicious payload) is the severe consequences that would result from a non-state actors' deliberate actions to penetrate or manipulate nuclear command, control, and communication (NC3) systems with AI-enhanced conventional capabilities. Above all, cyber operations targeting early-warning satellites and radars are central to this article's central focus.[28]

The presumption of immunity or overconfidence in the robustness of NC3 networks to external (or insider) threats may increase the chances that states underestimate nuclear weapon systems' vulnerabilities to a multitude of possible attack mechanisms, *inter alia*, data manipulation, malware cyber-attacks, 'false-flag cyber operations,'[29] social media flooding, or spoofing decision-makers with disinformation and misinformation. Minuteman missile silos are, for example, considered to be particularly vulnerable to cyberattacks.[30] Nuclear-powered ballistic missile submarines (SSBNs) once believed to be air-gapped – not connected to the internet and considered 'hack-proof' – are, however, connected via various electromagnetic signals that create potential vulnerabilities to cyber-attacks. A committed non-state or third-party actor may, for example, disrupt (digital jamming), deny (denial of service attacks), and distort or destroy information (spoofing or malware attacks) used by SSBN's command and control networks to sow miscalculation and misperceptions, fan the waves of a crisis,

or trigger an accidental nuclear launch.[31] *In extremis*, third-party hackers (or 'cyber terrorists') during a crisis might use AI-augmented false flag cyber operation to mislead submarine commanders (or political decision-makers) that an SSBN was – or expected to be in the case of a 'launch on warning' policy – under-attack.[32] Echoing these concerns, a 2009 study commissioned by the International Commission on Nuclear Non-proliferation and Disarmament (ICNND) warned that non-state actors might penetrate command and control (C2) systems – and even launch an unauthorized nuclear attack – an easier and more plausible 'alternative for terrorist groups than building or acquiring a nuclear weapon or dirty bomb themselves.'[33]

Recent developments in AI-enabling technology have exacerbated these vulnerabilities and introduced additional threat vectors non-state actors might leverage, which may precipitate a catalytic nuclear war that manipulates the information landscape in which decisions about atomic weapons occur. In particular, social media manipulation and the spreading of misinformation, false memes, and fake news. For example, in 2016, a false news story appeared on the AWD News site that claimed that Israel had threatened to attack Pakistan with nuclear weapons if Islamabad interfered in Syria. The report consequently caused a tit-for-tat incendiary rhetorical exchange on Twitter – the Pakistani defense minister Khawaja Muhammad Asif warned that Israel should remember that Pakistan is also a nuclear-armed state debunked as fictitious by the Israeli Defense Ministry.[34] This incident puts a modern spin on the Cold War concept of catalytic nuclear war – in which third party actions provoke a nuclear war between the two nuclear-armed powers – and demonstrates the potentially severe damage caused by the misinformation and manipulation of information by non-state actors.

During a crisis involving a competitive strategic dyad – such as the conflict-prone South Asia or Northeast Asia – when communication is compromised, nuclear arsenals are on high-alert, decision-making timeframes are compressed, or launch authority is pre-delegated (e.g., to nuclear-armed submarine commanders), the consequences could be catastrophic.[35] Moreover, in a high pressured crisis environment with confusion and paranoia running high the risk of misperceptions of an adversary's intentions and behavior (e.g., putting nuclear arsenals on high-alert status) non-routine troop movement), the temptations for pre-emptive action increases.[36] That is, the catalyzing non-state actor produces the effect of an imminent attack on one or both of two nuclear-armed states, for which pre-emption is considered the most advantageous strategy.[37] What factors might aggravate these escalation pathways during a crisis?

Pathways to catalytic nuclear war and catalyzing aggravators

This section examines four aggravating variables – information complexity, greater automation of NC3, disinformation, and nuclear multipolarity – that coalesce in the synthesis between emerging technology and the nuclear weapon systems. In particular, how and why AI and autonomy might compound these dynamics and cause unintentional escalation between nuclear-armed powers are more likely to occur. Early warning systems are the main focus of this study because 'missiles cannot be recalled; submarine commanders may be out of touch but able to act on their own; missiles may go off accidentally.'[38] These variables do not, however, constitute mutually exclusive risk scenarios. Instead, the interplay between these conditions might allow them to feed into one another with uncertain and potentially self-reinforcing effects – or a function of the confusion and uncertainty created by the sociotechnical complexity generated in the digital age.

Cognitive psychology is relevant to all four aggravating factors described below. Humans exhibit a range of biases that can influence how they observe, collect, and process information, making them less aware of the reality of a particular situation and thus more inclined to interpret events through the lens of existing desires, preferences, and beliefs.[39] Psychological research has shown that people tend to interpret ambiguous information as consistent with their pre-existing beliefs and values – dismissing information that contradicts these views – and accept information that allows them to avoid unpleasant choices.[40] Human psychology studies also tell us that people are predisposed to making fast, intuitive, reflexive, and heuristic judgments – known by cognitive psychologists as

'System I' thinking – rather than slower, more deliberative, conceptual, and analytical – known as 'System II' thinking.[41] Fast and reflexive ('System I') thinking during a nuclear crisis would be inherently problematic if, for example, the potentially catastrophic risks and second-order effects of nuclear war (i.e., loss of human life and environmental destruction) are actively or unconsciously deprioritized in discussions dominated by national defense heuristic imperatives.[42]

Cognitive biases can also be exacerbated when information overload and unfamiliar technologies are a more prominent feature of decision-making. The exponential rise and speed data emerge from today's information ecosystem will create novel attack vectors to manipulate and propagate misinformation and disinformation during crisis times. While human cognitive biases can be intensified by information complexity in crisis decision-making, over- confirmation bias, anchoring, heuristic thinking, and confidence bias are particularly pertinent in this context.[43] In the case of 'anchoring,' perceptions of historical lessons from past conflict and crises with minimal relevance may have an outsized influence on decision-makers who seek to ground decisions in precedent and lessons from experience.[44] Humans are also considered poorly equipped to intuitively understand probability – which is essential for ranking preferences rationally for crisis decision-making.[45] People tend to misinterpret (or not recognize) randomness and non-linearity and consider the occurrence of unlikely events as virtually impossible.[46]

Three distinct – but not always separate – mechanisms can lead to nuclear escalation, namely, *deliberate* (or intentional), *inadvertent*, and *accidental* escalation (encompassing mistaken or unauthorized usage).[47] These distinctions are not, however, binary or mutually exclusive. An escalation mechanism that leads from a crisis or conflict to its outcome can involve one or more of these categories. For example, if an accidental or inadvertent escalation signal or event is triggered by a non-state actor's nefarious actions – such as a false flag cyber operation against a state's NC3 systems – which in turn leads to a deliberate escalatory response.[48] Moreover, the deliberate use of nuclear weapons that originates from a false, manipulated, or distorted assessment of a situation, or in response to an early-warning system false alarm, can quickly muddy the lines of intentionality.[49]

Accidental nuclear war – a nuclear confrontation without a deliberate and properly informed decision to use nuclear weapons on the part of the nuclear-armed state(s) involved – could be caused by a variety of accidents, most often encompassing a combination of human – and human-machine interaction failure – system errors, and procedural or organizational factors.[50] This definition of accidental escalation will be used in this article to conceptualize 'catalytic nuclear war' – triggered by the deliberate actions of a non-state or third-party actor that is not part of the adversarial nuclear dyad – with developments in AI and autonomy. Therefore, a crucial distinction is between the risk of unintentional (i.e., accidental or inadvertent) escalation and intentional escalation; the fear of deliberate escalation is generally more destabilizing.[51] Despite paying lip-service to Machiavelli's Fortuna – the role of uncertainty in international affairs – decision-makers often underestimate the importance and frequency of accidents and randomness in these interactions.[52] Ultimately, whether the impact of unintended escalation risk is stabilizing or destabilizing depends on the destabilizing force's relative strengths and the fear it instills.[53] Similar to historical cases human-machine interactions caused (or compounded) accidents with complex autonomous weapon systems, AI-enhanced systems operating at higher speeds, sophistication, and compressed decision-making timeframes will likely reduce the scope for de-escalating and contribute to future accidents.[54]

Information complexity

In contrast to previous eras of military revolution – when technologies diffused and improved specific capabilities and domains of military power – the coalescence of emerging technologies today (especially AI machine-learning) can enhance and enable a diverse range of capabilities and missions.[55] The complex interactions within the tightly enmeshed (or dual-use) and co-mingled systems that control or inform nuclear weapons systems (e.g., early-warning satellites, ISR, electronic data networks, and missile defenses), has become a critical risk vector in the digital age. Thus, making

accidents, short of a nuclear detonation, almost par for the course.[56] Further, the risks of technical and human errors that arise from this complexity and interdependency in modern nuclear systems are compounded by the prospect of cyber-attacks against early warning and C2 systems – discussed below.[57]

A key characteristic of operating NC3 systems infused with advanced technology – AI machine-learning, big data analytics, and cyber – is the vast quantities of data and information collected to inform decision-making.[58] While AI-machine learning data-mining augmented cyber-surveillance – together with AI-enabled pattern recognition technology – could significantly enhance states' ISR capabilities, the introduction of multiple data streams with varying levels of confidence could overwhelm the ability of decision-makers to determine the credibility of data, especially if the provenance and validity of the information cannot be easily verified.[59] The U.S. Air Force, for example, has characterized this phenomenon as four 'V's' – higher *volume* (collection of magnitudes more data points), *velocity* (the volume of data is acquired at rapid speed), *variety* (numerous formats of information from diverse sources), and *veracity* (the volume, velocity, and variety of data includes a substantial amount of noise and redundant data).[60] Similarly, the U.S. Navy has reported being overwhelmed by the floods of data generated from its existing information-gathering systems.[61]

The quantity and quality of information generated by the advanced technology that supports NC3 systems can increase escalation risks, *inter alia*, in three ways. First, human decision-makers' dependence on the information produced by complex and enmeshed NC3 systems can exacerbate the degradation of the quality and reliably of decision-making if these systems are compromised, especially cyber-attacks. The literature on the risks of a breakdown in NC3 systems and the limitations of human cognitive processing capacity – particularly situations that combine information overload, confusion, and compressed decision-making timeframes – focuses on the accidental escalation risks associated with the organizational reliability of complex information architectures[62] general agreement exists that these risks cannot be eliminated.[63]

Second, while information dominance does not ensure stability; juxtaposed, information inadequacy (or 'information asymmetry') can prompt decision-makers during a crisis to eschew traditional caution and the acute fear of escalation for pre-emption; thereby, increasing inadvertent escalation risks.[64] For instance, the information asymmetries by clandestine cyber operations against NC3 systems can narrow or close the crisis bargaining window, producing a trade-off between possessing counterforce capabilities to blunt an adversary's nuclear deterrent and increasing the risks of provoking nuclear warfare in the first place. In short, any new capability that increases (or is perceived to do so) information asymmetry about the balance of power and resolve can sow the seed of bargaining and deterrence failure, and ultimately war.[65] Misperception of another's capabilities, intentions, and assumptions about what information an adversary possesses (or lacks), and in turn, how it perceives a situation, is an established cause of deterrence failure in international relations.[66]

The Cold War experience demonstrated that the fear that an action or signal misinterpreted by an adversary, in the context of uncertainty and incomplete information associated with modern warfare might spark nuclear pre-emption, is a useful point of departure to consider the potential effects of AI and autonomy.[67] Moreover, the increasing use and development of advanced weapon technology such as hypersonic glide vehicles, dual-use (nuclear and conventional) missile delivery systems, anti-satellite weapons, and dual-use NC3 systems that blur the distinction between nuclear and conventional (and sub-conventional) warfare can heighten strategic ambiguity during crises. Thereby creating first-mover advantage incentives, leading states to overestimate an adversary's capabilities and strike preemptively.[68] In sum, the likelihood of escalatory crises originating in cyberspace will continue to rise because of the inherent speed, scope, and opacity surrounding cyber capabilities, coupled with the existential fear of escalatory conflict in cyberspace, the positive feedback loops associated with states' accumulation of offensive cyber capabilities, and the insensitivity of others' perceptions and intentions.[69]

A cyber-attack – dependent on deception, speed, and covertness – that undermines effective information and communication flow can, therefore, increase the rational incentives for escalation.[70] Besides, irregular or opaque communication flow between adversaries may also increase the risk of misperception and miscalculation and assume the worst of others' intentions.[71] Paradoxically, therefore, new technologies designed to enhance information (e.g., modern NC3 systems supported by 5G networks, AI machine learning, cyber, big-data analytics, and quantum computing) can also erode reliable information flow and communication, which is critical for effective deterrence.[72] Greater volumes of information emerging from the global information ecosystem, rather than reducing the likelihood of misperceptions arising between adversaries, offer motivated third-party actors unprecedented – low-cost, scalable, and fast – opportunities to manipulate decision-makers perceptions and shape public opinion about others' intentions, capabilities, and the consequences of responding to an attack that originated in the cyber domain – but might rapidly and inadvertently escalate.

Third, advanced technologies have the potential to qualitatively improve the 'always-never' criteria that NC3 systems must meet (i.e., complex failure modes, greater redundancies, and a larger attack surface)[73] the complexity and uncertainties introduced by sophisticated NC3 mechanisms – especially early-warning satellites and radars – can also cause errors, unexpected interactions, and unintended consequences, eroding deterrence and creating rational (or sub-rational) incentives) to escalate a situation.[74] The notion that sophisticated technologies, designed to improve safety, ultimately become a source of accidents and errors is a documented phenomenon.[75] This trade-off is a product of the organizational and strategic-cultural variables in the human decision-making process, creating pressures to escalate a situation, rather than viewing technology as an independent variable *per se*.[76] The assumption of rationality in a nuclear-armed adversary, or that the adversary has more information than might be the case (or 'asymmetric information'), *inter alia*, may stem from misperceptions about the nature of an opponent's technological capabilities, the nature of a crisis, or an opponent's intentions. As a result, these factors could exacerbate escalation pressures on states with less robust NC3 systems and safeguards or view its survival at stake – North Korea, Pakistan, or India. Q Furthermore, the normal (or peacetime) perception of events can shift during crises or geopolitical tension, when decision-makers are more prone to harbor worst-case scenario expectations and see things they *expect or want to see* – known as 'cognitive consistency.'[77] This logic – the way people interpret situations under pressure due to anxiety – can also become a self-reinforcing process that directly impacts the number of false alarms that occur in early warning systems. In particular, the probability of Type II errors – or 'false negatives.'[78] Moreover, as the volume of information received by NC3 systems increases during a period of tension, so do the chances of perceiving a pattern of adversarial behavior where under typical situations, no pattern would be detected.

Technological-deterministic schools of thought that conceptualize technology as the *sole* driver of strategic stability and the conduct of warfare, therefore, oversimplifies this complicated human endeavor – which can mean that rational decisions entail worse consequences than less rational alternatives. For example, if a new state-of-the-art NC3 capability such as AI provides a first-mover advantage (or perception of one) for one or both sides striking preemptively during a crisis, or irreparable damage is suffered by one side from a first strike, the risks of inadvertent nuclear war increases.[79] Besides, problems have often arisen because states tended to place 'new weapons in old bottles.' That is, regardless of how sophisticated or capable a system appears to be, these technologies will invariably be assimilated into military organizations with entrenched norms, cultures, and institutional patterns.[80] A complete understanding of the relationship between advanced technology and strategic stability requires a deeper understanding of human tendencies, not merely technical capabilities *per se*.

Greater automation of NC3 systems

Increasing automation levels in modern NC3 systems – particularly those augmented and supported by AI technology – together with false warnings and other forms of nefarious interference in

cyberspace, will likely increase the risk of accidental nuclear war.[81] Legacy vulnerabilities in C2 systems to computer malfunctions, human error (both Type I and Type II).[82] Furthermore, especially human-machine interfaces will be compounded as AI is infused into the NC3 ecosystem.[83] Moreover, a declared or *de facto* launch-on-warning posture held by either side during a crisis involving a Type II error (or 'false positive') caused by a false signal increases the accidental escalation risks a result of 'use-them-or-lose-them' pressures. Thus, it degrades nuclear forces' usability or reduces the damage limitation window to attack enemy nuclear forces if needed.[84] This skewed assessment in the context of nuclear weapons ready to launch nuclear weapons at a moment's notice could precipitate worst-case scenario thinking and trigger inadvertent escalation.[85]

The historical record demonstrates the vulnerabilities of NC3 systems to frequent false alarms, accidents, 'close calls,' and other risks associated with increasingly complex, porous, and interconnected systems, which despite their alleged 'closed' nature, may offer non-state actors multiple pathways to cause harm.[86] Because of the inherent problems attribution in cyberspace, a non-state actor might plausibly make *both states* the target of its attack, convincing each side that the other party is responsible – or a 'double-sided catalytic attack.'[87] This problem-set is compounded by the shortened response timeframes available to decision-makers, particularly where a 'launch on warning' posture is present.

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 and decide whether or not to authorize a nuclear response.[88] This compressed decision-making timeframe – for data-collection, assessment, and decision-making – could put political leaders under intense pressure to decide to escalate during a crisis, with incomplete (and possibly false) information of a situation. In the context of advanced technologies, response systems like NORAD raise particular concerns; missiles cannot be recalled, submarines on deterrence patrols may be out of touch for extended times with a high degree of C2 autonomy, and nuclear weapons may be launched accidentally.[89]

As Thomas Schelling presciently warned during the Cold War, 'when speed is critical, the victim of an accident or a false alarm is under terrible pressure.'[90] This pressure would likely be more intense for nuclear powers nearby (e.g., India and Pakistan, or China and North Korea), or where one side's NC3 systems are asymmetrically more vulnerable or less sophisticated, for example, North Korea, India, or Pakistan.[91] Further, the increasingly interdependent and commingled (or entangled) nature of states' conventional and nuclear C2 systems might exacerbate the incentives to escalation a situation to a nuclear-level once a conventional crisis or conflict begins.[92]

The decision to automate nuclear assets may also be influenced by the political stability and the threat perceptions of a nuclear-armed state. A regime that fears either a domestic-political challenge to its rule or foreign interference may elect to automate its nuclear forces, ensuring only a small number of agents are involved in the nuclear enterprise.[93] For example, China maintains strict controls on its nuclear command and control structures (i.e., separating nuclear warhead and delivery systems). Open-source evidence does not suggest Beijing has pre-delegated launch authority down the chain of command if a first strike decapitates the leadership. In short, 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 and intoxicating option to authoritarian regimes such as China.

In the context of AI and autonomy, particularly information complexity, misinformation, and manipulation, rationality-based deterrence logic appears an increasingly untenable proposition. In sum, the combination of increased speed, compressed decision-making timeframes, complex and interdependent command and control mechanisms, and increasing levels of AI-enabled NC3 automation and human-machine interactions increases the possibility of sparking a cascading set of

errors (human or machine), reinforcing feedback-loops that could ultimately lead to an accidental nuclear exchange.

Furthermore, because there is so much redundancy built into modern NC3 systems, scenarios involving accidental nuclear war as a result of a single technical error or malfunction is less threatening, compared to the simultaneous coalescence of multiple failures in a rapid and unpredictable and reverberating fashion – where malfunctions in once system are prone to cause alarms in others.[94] In particular, if they are compounded by organizational failures, ambiguous information, misperceptions, or excessive trust in technology, which may lead human operators cognitively offloading judgment to AI algorithms without fully understanding its limitations – also known as 'automation bias.'[95] Such automation bias – especially in human-machine interactions – could also mean that *both* false negatives and false positives go unnoticed (or are discarded) because the operators are overconfident in systems augmented with advanced technology such as AI.[96] This problem-set is compounded during a crisis when stress, fatigue, information overload, and commingled (nuclear and conventional) systems encounter *a priori* situations between asymmetric nuclear rivals, thickening the 'fog of war' – the inevitable uncertainties, misinformation, or even breakdown of organized units, which influences warfare – resulting in irrevocable actions, when use-or-lose-them becomes the only option.[97]

Nuclear multipolarity

In a multipolar world nuclear order with competing and contested strategic dyads – India-Pakistan, U.S.-Russia, U.S.-North Korea, U.S.-China, and perhaps India-China – with weak mechanisms for de-escalation, doctrinal opacity, and questionable attitudes towards nuclear restraint, the potential risk of nuclear catalysis increases.[98] The emergence of nuclear multipolarity in the Second Nuclear Age has created multifaceted escalation pathways to a nuclear confrontation involving an expanding number of nuclear-armed poles, compared with bipolarity during the Cold War.[99] Kenneth Waltz, the founder of structural realism, argued that while nuclear weapons served as a stabilizing force during the Cold War-era, however, 'increased numbers of actors increase levels of systemic uncertainty … rising uncertainty heightens potential miscommunication and conflict. Bipolarity [compared to multipolarity] is, therefore, the most stable form of international power distribution.'[100] This multipolarity is essential because each state will choose a different response to the new choices emerging in the digital age.[101] Competing states making decisions under the nuclear shadow will be more inclined to assume the worst of others' intentions, especially in situations where the legitimacy of the status quo is contested (i.e., maritime Asia).

Against the backdrop of existing tensions – especially a limited or proxy conflict or other crisis – between nuclear-armed adversaries (U.S.-China; U.S.-Russia; India-Pakistan; or U.S.-North Korea), therefore, an act of terrorism or other nefarious non-state action would create additional incentives for political leaders to assume the worst of the other's intentions. Besides, the victim of a non-state or terrorist attack (e.g., cyber, disinformation, or drone swarm attack[102]) would likely be under intense domestic political pressure to attribute blame swiftly and retaliate against the perpetrator – or aiders and abettors in case of a proxy actor.[103] During a tense situation, when prudent and careful planning can run aground against the fog and friction of reality, misperception and miscalculation may generate temptations for pre-emption with potentially catalytic effects.[104] According to the offensive realist scholar John Mearsheimer, 'as long as the system remains anarchic, states will be tempted to use force to alter an unacceptable status quo.'[105] The other side could easily perceive one state's efforts to enhance its strategic forces' survivability with state-of-the-art dual-use technology like AI as a potential threat to its ability to survive and respond to a nuclear first strike – or second-strike capability.[106]

Authoritarian states may perceive an adversary's intentions very differently from a democratic one.[107] The belief that a regime's political survival or legitimacy is threatened might cause leaders to consider worst-case scenario judgments and behave in a manner predicted by offensive realist

scholars.[108] Conversely, non-democratic leaders operating in closed political systems such as China, Russia, or North Korea may exhibit a higher degree of confidence (or overconfidence) in their ability to respond to perceived threats in world politics.[109] Bias assessments from a non-democratic regime's (or 'Stasi' type) intelligence services might reinforce a leader's faith – or a false sense of security – in their diplomatic skill and maneuverability.[110]

Without institutionalized structures (or general staff system) connecting the intelligence services with the military and broader political context, decisions will likely be made in vacuums, with minimal checks and balances on the political leadership (or supreme leader), and a reduction in 'bottom-up' (or a 'fact searching' organizational culture) information flow – because of the fear of contradicting the leadership.[111] This situation can reinforce a distorted (or false) sense of reality, thereby compounding the cognitive misperceptions of events that are already present. This situation can reinforce a distorted (or false) sense of reality, thus compounding the cognitive misperceptions of events that are already present. Social media – and other AI-enhanced information tools – that supply decision-makers with a continuous flow of near-real-time information might also complicate the practice of deterrence and escalation management before and during future crises.[112] It is politics and the breakdown of human bargaining, therefore, that ultimately leads to conflict; AI and other kinds of advanced technologies – drones, deepfakes, and cyber capabilities – can be viewed as the tools (or dependent variables) that complicate and exacerbate the escalation dynamics determined by human choices, biases, and interactions.[113] The amplification of false alarms or the creation of false signals by social media (i.e., false positives and false negatives) might also disrupt critical communication channels between commanders and their political leadership and between allies and adversaries during crisis or conflict.[114]

Authoritarian regimes whose political legitimacy and regime stability are conditioned or legitimized by the general acceptance of official narratives and dogma tend to become empowered when people's trust in truth (i.e., faith in what they see and hear) is undermined. The vacuum is filled by the opinions of authoritarian regimes and leaders with authoritarian inclinations.[115] Though authoritarian regimes might follow a rational logic and pattern of behavior in strategic decision-making, this 'logic' will reflect the myriad factors that influence a leader's threat perceptions – including, *inter alia*, leaders personalities, beliefs, experiences, and the importance attached to reputation and status – tethered to its (possibly distorted or misinformed) version of events.[116] Thereby confounding the expectations and calculations – notions of deterrence and escalation thresholds – of adversaries and increasing accidental escalation risk.[117] The relative importance of these factors upon states' resolve under crisis conditions in context-bound situations is difficult to answer with any certainty.[118]

Furthermore, a regime that views its second-strike capabilities – especially its NC3 systems – as vulnerable or insecure (North Korea, Pakistan, or perhaps China) may be more inclined to automate its nuclear forces and launch postures. In short, non-democratic nuclear states with relatively centralized C2 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. 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 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.[119]

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.[120] In sum, while technological factors involved in NC3 systems are of great

importance, political, psychological, strategic cultural, doctrinal, and organizational considerations are also central to understanding the nature and key to the mitigation – of accidental escalation risk. How might AI-augmented cyber capabilities create new pathways for accidental or inadvertent escalation?

Disinformation, misinformation, and information manipulation

One rapidly developing and increasingly prominent field of AI-augmented technology that can complement and force multiply existing malicious social manipulation behavior and generate campaigns of manipulation – most notably the spread of misinformation or disinformation – is the ability to generate audio and video images that fabricate events, create fictitious situations and propagate falsehoods.[121] Experts expect that it will be possible for non-state actors (or amateurs more generally) to generate photorealistic high definition video, audio, and document forgeries using machine-learning tools – in particular, generative adversarial networks (GANs) – at a low-cost and on a large scale.[122] In combination with other tools,[123] GANs have already demonstrated that producing realistic images is possible and will likely improve significantly in the next few years.[124] Examples of disinformation and manipulation, *among other things*, include distributed audio of private statements and conversations between political or military leaders (both domestically or with allies and adversaries); video of crises or conflict (see below) designed to incite public outrage and prompt pressure for retribution; or audio or video material calibrated to deflect, deceive, or otherwise distract attention from, an actor's culpability of an aggressor by producing false alternative versions of events.[125]

As AI technology advances, the quality, cost, and availability of GANs and other tools – especially AI-enhanced audio software will make it increasingly difficult to discern what is real from what is not, eroding public trust in hitherto trust-worthy information sources.[126] In 2014, for example, thousands of residents at St. Mary Parish in Louisiana received a fake text message alert via a bogus Twitter account warning of a 'toxic fume hazard' in the area. Further fanning the flames, a fake YouTube video was also posted showing a masked ISIS fighter standing next to looping footage of an explosion.[127] Thus, it is not difficult to imagine how these AI-enhanced technologies in the hands of non-state actors with nefarious goals (or 'apocalyptic' world view) might have dangerous consequences or for nuclear security and strategic stability. Deliberate malevolent information manipulation by non-state actors (terrorists, criminals, or state proxies) could destabilize implications on effective deterrence and military planning, both during peace and war. GANs generated deepfakes might also exacerbate the escalation risks by manipulating the digital information landscape, where decisions about nuclear weapons are made. It is easy to imagine unprovoked escalation caused by a malicious third-party (or state-proxy) clandestine false-flag operation in the competitive strategic environment.[128] During a crisis, a state's inability 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.[129]

AI systems will likely come under inexorable stress from nefarious attacks using counter-AI techniques (e.g., data pollution, spoofing, false alarms, or tricking a system in order to reverse engineering algorithms), which might undermine the confidence in a network, creating new vulnerabilities, errors, and unintentional escalation risks. In the emerging deepfakes arms-race – much like cybersecurity more broadly – detection software will likely lag behind advances in offensive enabling solutions – or offense-dominant ones.[130] According to computer science expert Hany Farid, there are probably 100 to 1,000 times 'more people developing the technology to manipulate content than there is to detect [it].'[131] Efforts to counter these technologies and regain the upper-hand in the war on digital fakery have been underway for several years (e.g., automated anti-virus software, encryption, and other tools to call-out manipulated images and videos). Today, more sophisticated AI-enhanced techniques coupled with research that suggests pre-existing cognitive schemes, beliefs, and attitudes rather than credulity or gullibility determines whether the public believes particular fakery is real or not – meaning fake images and videos can achieve accepted even though they can be easily debunked.[132]

AI-enhanced fake news, deepfakes, bots, and other malevolent social media campaigns could also influence public opinion – creating false narratives or amplifying false alarms – with destabilizing effects on a mass scale, especially in times of geopolitical tension and internal strife.[133] In 2017, for example, a deepfake video was circulated on Russian social media, alleging a U.S. B-52 bomber had accidentally dropped a 'dummy nuclear bomb' on a Lithuanian building.[134] A state or non-state actor could, for instance, generate an image or recording of a military commander obtained from open-sources to generate and disseminate a deepfake containing false orders, intelligence, or geospatial imagery that at best generates confusion that, in the worst-case scenario, aggravates a tense situation or crisis between rival nuclear powers – this idea is examined below. For example, in the world's first reported AI-enabled theft in 2019, non-state actors used AI voice mimicking software to generate a fake recording of a British energy executive.[135] In short, AI-augmented technology is rapidly becoming another capability in non-state actors' toolkit to wage campaigns of disinformation and deception – one that both sides may have used against them.[136]

The democratization of ever more sophisticated technology will likely amplify documented human pathologies,[137] explaining people's attraction to novel and negative information, memes, falsehoods, and filter bubbles – the so-called 'information cascade' phenomenon – which deepfakes are adapt to perpetuate.[138] In the case of cognitive availability heuristics, people tend to skew their judgments to more recent (or memorable) information, situations, or experiences, thus making new or novel opinions biased toward those that can be more easily recalled.[139] During times of high-pressure crises, decision-makers tend to interpret unusual circumstances as threatening, even if an adversary's behavior has not changed. Routine activities (e.g., troop movements) scrutinized in the context of an early-warning alert may be considered as more menacing than they might otherwise be.[140]

In 2017, for example, South Korean counterintelligence officials received fake mobile and social media alerts with orders for U.S. military and DoD personnel to evacuate the Korean Peninsula.[141] Information attacks such as this suggest that non-state actors, state proxy actors – and perhaps state actors – will inevitably attempt to use social media as a tool of war to provoke nuclear confrontation for political-ideological, religious, or other malevolent goals; and with increasing levels of sophistication, stratagem, and AI-enhanced subterfuge.[142] AI might also enable non-state actors to automate, accelerate, and scale synthetic social media accounts and content to support malevolent disinformation operations.[143]

These potentially catalytic dynamics will likely be compounded by human cognitive bias. A motivated non-state actor would be well-positioned to use AI-augmented tools (e.g., 'fake news' or deepfake generated propaganda) to exploit this psychological weakness, ensuring control of the dissemination of false narratives and opinions. Threat assessments of catalytic escalation must, therefore, include both the likelihood that a third-party can execute a particular attack and the probability that the falsehoods or fakery that transpires leads to escalation.[144] Strategic decisions made by isolated nuclear-armed authoritarian regimes under these circumstances – especially regimes that believed their survival was threatened – could trigger dangerous accidental nuclear escalation dynamics.

Fictional scenarios

The following section includes four fictional scenarios – cyber 'false flag' operation, cyber-attack on nuclear early-warning systems, drone swarming targeting ISR systems, and deepfake disinformation – to illustrated how AI-augmented capabilities in the hands of non-state actors might accidentally or inadvertently drag a competitive nuclear dyad into conflict.[145] These fictional scenarios are premised on three assumptions: (1) the involvement of a nuclear-armed adversarial dyad adversary during a crisis or period of tension; (2) the existence of information asymmetry; and (3) the technical feasibility of the operations described – either currently in existence or considered technically viable shortly. Why these capabilities and not others? These capabilities are of particular interest because they play an increasing role in how non-state actors leverage the asymmetric and low-cost benefits of

AI technologies to realize their malevolent objectives. Although multiple pathways may be activated during an actual conflict or crisis – either simultaneously or sequentially – assessing each of these potential escalatory pathways offers insight into the interplay of strategic AI technologies and stability risks. One caveat, though: in the real-world, the aggravating factors described above, would interact with each other. The participating actors' relative capabilities would vary, thus producing a potentially much larger potential universe of possible scenarios than outlined here.

Scenario (1) *Cyber 'false-flag operation'* [146]

Party A (a non-state actor) launches a false flag cyber operation – data manipulation, social media flooding, a spoofing attack, or other deception – against State B and C, which is not traced to Party A and appears to both State B and C to come from the opposing side.[147] Convinced that the other side is responsible and about to be attacked by the opposing side of the nuclear-dyad, State B retaliates against State C in a pre-emptive attack that C views as unprovoked aggression – sparking a catalytic war. During a crisis between two states, leaders would likely be predisposed to assume the worst about the other's intentions, thus making them less likely to exercise rigorous due diligence to establish high confidence attributions (or false positives) due to a cyber-attack.[148]

Scenario (2) *Cyber-terrorism vs. early warning systems*

During a period of heightened tension or crisis between State A and State B, a third-party actor or terrorist floods social media outlets and open-source crowdsourcing platforms with false information (e.g., satellite imagery, 3D models, or geospatial data) about the suspicious movement of State A's nuclear road-mobile transporter erectors-launchers (TELs).[149] Because of State B's inability to determine with confidence the veracity of this information, and with mounting public pressures to respond, State B escalates a situation on the falsehood (or false positive) it is the target of an unprovoked attack.[150] Asymmetries between adversaries NC3 systems and military capabilities would likely exacerbate the fictional scenario's escalation mechanisms.[151] Taken together, increasingly sophisticated and accessibility of deepfake technology, the inherently dual-use nature of AI, the problem of attribution in cyberspace, the increasingly complex and interdependent nature of NC3 systems, and a compressed timeframe for strategic decision-making associated with hyper-speed warfare, will continue to lower the threshold for false-flag operations.[152]

Scenario (3) *Drone swarming vs. ISR systems*

An insurgent proxy of State A uses a swarm of autonomous micro-drones over an adversary's territory as part of a clandestine ISR mission.[153] Because this type of sophisticated technology is considered beyond the reach of a lone wolf or unsponsored terrorist group,[154] State B attributes to its militarily superior adversary, and thus, views this proxy as a precursor for an attack planned by A. The use of autonomous swarms, especially during periods of tension or insensitive anti-access/area denial (A2/AD) environments, might generate incentives for pre-emption. Further, the sophistication of A2/AD capabilities such as drone swarms – and advanced non-nuclear strategic weapons including hypersonic weapons, cyber, and anti-ship ballistic missiles (ASBMs) – complicates Herman Kahn's notion of 'ladders of escalation' to achieve a superior position vis-a-vis an adversary across the range of escalation rungs – or 'escalation dominance.'[155] Where on the escalation ladder is a swarm attack against the United States' nuclear ISR? Would a Chinese cyber-attack targeting a U.S. water supply facility be more or less escalatory than an ASBM attack on a U.S. aircraft carrier in the West Pacific?[156]

Conversely, the discovery of information by State A (e.g., suspicious troop movement or missiles leaving their garrison) could prompt A to escalate a situation, motivated by the desire to capture the perceived strategic upper hand. Moreover, State A might consider the first-mover advantages of using low-cost and relatively dispensable drones as controllable at a conventional level, which would likely increase their appeal, at least in the opening stages of a conflict. In the event State A's unarmed proxy

drone swarm was attacked by B in a counterterrorist or counterinsurgent strike,[157] State A would be forced to either accept this vulnerable asset's loss or escalate a situation – through its surrogate proxy or more directly.[158]

Scenario (4) *Deepfake disinformation*

To incite conflict between two nuclear-armed rival states, State A hires proxy hackers to launch a deepfake video, depicting senior military commanders of State B conspiring to launch a pre-emptive strike on State C.[159] This footage is then deliberately leaked into C's AI-augmented intelligence collection and analysis systems, provoking C to escalate the situation with strategic consequences.[160] B, fearful of a decapitating strike and losing the first mover's advantage, swiftly escalates the situation.[161] The current underdeveloped state of 'counter-AI' capabilities and other fail-safe mechanisms (e.g., circuit breakers) to de-escalate escalation in cyberspace will make the unprovoked and unintentional escalation dynamics depicted in this scenario very challenging to anticipate and thus mitigate.[162]

Conclusion

How can militaries maintain effective command and control of nuclear forces in a rapidly evolving, uncertain, and complex conflict environment? While no amount or combination of controls, procedures, or technical enhancements can eliminate the possibility of catalytic nuclear escalation – and accidental escalation more broadly – specific measures that focus on reducing the possibility of accidental nuclear war may help to reduce some of the risks highlighted in the scenarios in this article – especially human and technical errors that occur in cyberspace.[163] Because accidental (and inadvertent) escalation dynamics are not the result of decision-makers electing to escalate, it cannot be deterred in the traditional sense.[164] Deterrence by denial measures by decision-makers, in advance of a non-state actor's attack, might contribute to escalation management, however.[165]

These measures can be broadly categorized into three areas.[166] First, enhancing the safety of nuclear weapons and harden NC3 systems and processes (e.g., safeguards, risk analysis to strengthen NC3 systems, and building more resilience into PALs).[167] NC3 infrastructure security could be strengthened further through more robust cyber defenses (e.g., deepfake detection software), identifying false information and ameliorating the impact of misinformation and disinformation.[168] Second, improving existing NC3 control protocols, and mechanisms (e.g., adding redundancies, enhancing procedures and launch protocols and authentication codes, and including back-ups into ISR operations and data-gathering).[169] Third, enhancing safeguards to contain the consequences of errors and accidents when they occur (e.g., personal training, independent sources to verify and collaborate threat and crisis assessments, collective monitoring of events, and bilateral and multilateral data exchanges in the aftermath of accidents).[170] Social media could be used, for example, to distribute and disseminate accurate rapidly and time sensitive information during a crisis or conflict.

The extent to which these measures decrease the risk of catalytic war will ultimately depend on how the various actors (state and non-state) use the cacophony of AI-augmented technology to achieve their eclectic – and often conflicting – goals. To be sure, these factors might have confounding effects in which AI and autonomy. For instance, some actors might use AI technology to decrease risk, even as others – especially non-state and third-party actors – take deliberate steps that increase escalation risks. Ultimately, the effects of catalytic escalation risk that emerge from the use of these technologies will depend on the relative strength of the destabilizing factors involved, how both sides perceive these factors, and the fear it instills in others.

Several experts have suggested that integrating AI machine learning and autonomy into NC3 systems by improving situational awareness, reducing information overload, and enhancing the speed and scope of intelligence collection and processing may strengthen nuclear safety.[171] Others contend that

the introduction of AI, and other emerging technology, could create new vulnerabilities and sources of errors, which non-state actors will inevitably seek to exploit.[172] These concerns have prompted some to propose that nuclear-states commit to retaining humans' role in decisions relating to NC3 (especially early warning systems) and the nuclear enterprise more broadly.[173] In a similar preventative approach, some have called for states to prohibit using cyber capabilities against NC3 systems. This would do little to preclude non-state (or state-sponsored) cyber interference, however.[174] In short, given the complex interactions and interdependencies of modern NC3 systems, technical solutions may well create new risks, vulnerabilities, and uncertainties that compound the existing risks of catalytic escalation.

Besides, nuclear-powers might explore expanded intelligence sharing arrangements, reduce accidental escalation risks – unauthorized use of nuclear weapons, human or technical errors, or false warning – caused by non-state actors' actions. Other more wide-ranging measures that may also 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 do today.[175] 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.[176] Strategic competition between great powers, to reap the perceived first-mover advantages of AI, will likely become a negative-sum enterprise without resolving the control and safety issues.[177]

Can arms control agreements encompass emerging technologies such as AI? How might non-proliferation look in the age of AI?[178] The long-standing Nuclear Non-Proliferation Treaty (NPT) provides a successful case study in global governance that minimized the threat posed by the weaponization of new (i.e., atomic) technologies while enabling the mutual benefits of sharing nuclear technology to strengthen strategic stability. AI is much more broadly diffused than nuclear technology; however, with the private sector's heavy involvement in the research, development, and application of AI technology. When the lines between dual-use capabilities and nuclear and non-nuclear are blurred, arms control is much more challenging, and strategic competition and arms racing are more likely to emerge.[179] In short, legacy arms control frameworks, norms, and even the notion of strategic stability itself will increasingly struggle to assimilate and respond to these fluid and interconnected trends. Another complicating issue is that today there are no standard definitions – or engineering methodology – to formulate new regulatory or arms control frameworks to ensure the safety and robustness of AI – the so-called 'AI control problem.'[180] For example, some AI experts believe that existing tools (e.g., reinforcement learning techniques) cannot resolve humans' risks from AI-augmented autonomous systems.[181] Thus, a different kind of algorithm than currently exists will probably be needed; prioritizing the safety, robustness, and interpretability of complex AI-infused NC3 systems.

As decision-makers ruminate over the ways and means in which the actions of a non-state actor might spark a catalytic nuclear exchange, they will need to consider how they would respond to such an attack and how an appropriate response might become a self-fulfilling prophecy – a chain of reaction and counter-retaliation – that leads unintentionally to nuclear war. Thus, as militaries modify operational concepts and strategic doctrines to manage escalation risk in the digital age, it behooves leaders to account for the rapidly evolving nature of the emerging threat environment, characterized by complexity, uncertainty, deception, misperception, and vulnerability. Future research would benefit from a data acquisition strategy test – such as wargaming or experimental wargaming – to the extent the effects of the aggravating factors described in this article affect human decisions to escalate a situation, potentially leading to catalytic war.

Disclosure statement

No potential conflict of interest was reported by the author.

Additional information

Notes on contributors

James Johnson

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

Notes

1 The 'Nth Country Problem' refers to the Cold War-era diffusion of nuclear weapons to additional states – through indigenous development programs, assistance from the present nuclear powers, or a combination of both – became a key area of concern in the United States, in particular, the implications of this proliferation for U.S. disarmament and nuclear assistance programs. See, Fred C. Ikle, Hans Speier, Bernard Brodie, Alexander L. George, Alice Langley Hsieh, and Arnold Kramish, *The Diffusion of Nuclear Weapons to Additional Countries: The 'Nth Country' Problem* (Santa Monica, CA: RAND Corporation, 1960).

2 Several scholars have argued that catalytic war is unlikely because the catalyzing actor might be destroyed in the resulting war. If the catalyzing country had previously determined that the war would not damage it, it might view the utility of a catalytic war as advantageous. For example, see Henry A. Kissinger, *The Necessity for Choice* (New York: Harper & Bros., 1960), and Herman Kahn, 'The Arms Race and Some of Its Hazards', *Dxdalus*, 89 (1960), pp.744–80. On 'catalytic nuclear war' theorizing also see Richard S. Leghorn et al., 'The Problem of Accidental War', *Bulletin of the Atomic Scientists*, 14 (1958), pp.205–9. Arthur L. Burns, *The Rationale of Catalytic War* (Princeton, N.J.: Center of International Studies, Princeton University, Research Monograph No. 3, 1959).

3 Donald H. Kobe, 'A Theory of Catalytic War', *The Journal of Conflict Resolution* 6, no. 2 (1962), p.125.

4 Fred Charles Ikle, 'Nth Countries and Disarmament', *Bulletin of the Atomic Scientists,* 16 (1960), pp.391–394.

5 Kahn, 'The Arms Race and Some of Its Hazards', pp.744–80.

6 Ibid., p.129.

7 Richard A. Brody, 'Deterrence Strategies: An Annotated Bibliography', *The Journal of Conflict Resolution,* 4 (1960), pp.443–57.

8 Therefore, a catalytic war would still be possible if a third-party instigator believed that the potential benefits from the war outweighed the risk of receiving a retaliatory attack. Ibid., p.128.

9 Optimal arsenal levels and force structure of nuclear weapons would balance the 'positive' role of nuclear weapons (i.e., mutual deterrence) and the 'negative' one (i.e., accidental nuclear war), allowing for a stabilizing deterrence regime alongside restructured nuclear forces and their attendant support systems to reduce the chances of accidental escalation. Michael D. Intriligator and Dagobert L. Brito, 'Minimizing the Risks of Accidental Nuclear War: An Agenda for Action', in *Inadvertent*

Nuclear War, eds., Hakan Wilberg, Ib Damgaard Petersen, and Paul Smoker (New York, NY: Pergammon Press, 1993), p.228.

10 It is beyond this article's scope to examine all of the technical vulnerabilities of NC3 systems or the potential implications of possible countermeasures. See, Amy Woolf, *Defense Primer: Command and Control of Nuclear Forces* (Washington, DC: Congressional Research Service), 11 December 2018; and Page O. Stoutland and Samantha Pitts-Kiefer, 'Nuclear Weapons in the New Cyber Age: Report off the Cyber-Nuclear Weapons Study Group' (Washington, DC: Nuclear Threat Initiative, September 2018).

11 On the nature and drivers of non-state actors (especially terrorist groups) interest in nuclear weapons see, Mark Fitzpatrick, *The World After: Proliferation, Deterrence and Disarmament if the Nuclear Taboo is Broken* (Paris: Ifri Security Studies Centre, 2009); Charles C. Ferguson and William C. Potter, *The Four Faces of Nuclear Terrorism* (Monterey, CA: Center for Nonproliferation Studies and Nuclear Threat Initiative, 2004); James Forest, 'Framework for Analyzing the Future Threat of WMD Terrorism', *Journal of Strategic Security* 5, no. 4 (2012), pp.51–68; and Graham Allison, *Nuclear Terrorism: The Ultimate Preventable Catastrophe* (New York: Owl Books, 2004).

12 For scholarship on the intersection of technology with the potential for war see, James D. Fearon, 'Cooperation, conflict, and the costs of anarchy', *International Organization*, 72:3 (2018), pp523-559; for escalation dynamics see, Caitlin Talmadge, 'Emerging technology and intra-war escalation risks: Evidence from the Cold War, implications for today', *Journal of Strategic Studies*, 42:6, (2019), pp.864–887; for arms racing and strategic stability see, Todd S. Sechser, Neil Narang, Caitlin Talmadge, 'Emerging technologies and strategic stability in peacetime, crisis, and war', *Journal of Strategic Studies*, 42:6, (2019), pp.727–735; and for deterrence and the offense-defense balance see, Ben Garfinkel & Allan Dafoe (2019) How does the offense-defense balance scale? *Journal of Strategic Studies*, 42:6, 736–763; and Joseph S. Nye, 'Deterrence and dissuasion in cyberspace', *International Security*. 41:3 (2017), pp.44–71.

13 Notable exceptions include Herbert Lin, 'Escalation Dynamics and Conflict Termination in Cyberspace' *Strategic Studies Quarterly*, Vol. 6, No. 3, (Fall 2012), pp.46–70; Erik Gartzke and Jon R. Lindsay, 'Thermonuclear Cyberwar', *Journal of Cybersecurity* 3, no. 1 (February 2017), pp.37–48; Edward Geist and Andrew Lohn, *How might artificial intelligence affect the risk of nuclear war?* (Santa Monica, CA: RAND Corporation, 2018), and Matthijs M. Maas, 'How viable is international arms control for artificial military intelligence? Three lessons from nuclear weapons', *Contemporary Security Policy* Vol. 40, Issue 3 (2019), pp.285–311; and James Johnson, 'Artificial intelligence & future warfare: Implications for international security', *Defense & Security Analysis* 35:2 (2019), pp.147–169.

14 'Artificial Intelligence' (AI) refers to a machineability to perform tasks that generally require human-level intelligence. For example, recognizing patterns, learning from experience, making predictions, or taking action – whether digitally or as smart, behind autonomous physical systems. Today's AI systems use 'narrow' AI – or task-specific intelligence. For more on definitions of AI, see Nils J. Nilsson, *The Quest for Artificial Intelligence: A History of Ideas and Achievements* (Cambridge: Cambridge University Press, 2010). 'Autonomy' is the ability of a system to respond to uncertain situations by independently composing and selecting among different courses of action to accomplish goals based on knowledge and a contextual understanding of the world, itself, and the situation. Autonomy is categorized by degrees of unsupervised behavior (i.e., degrees of autonomy) that range from fully manual to fully autonomous. United States Air Force, 'Artificial Intelligence Annex to DoD AI Strategy', Tech. Rep., United States Air Force (2019); and Andrew P. Williams and Paul D. Scharre eds., *Autonomous Systems – Issues for Defence Policy Makers* (NATO Allied Command Transformation, Norfolk, VA, 2015).

15 The concept of ‘emerging technologies’ is a loose and ill-defined one that is often used interchangeably when discussing the impact of technologies such as cyberspace, social media, autonomous weapons, quantum computing, rail guns, directed energy weapons, 3D printing, and AI. Emerging technologies can mean different things to different people and are analytically very distinct. For example, cyber capabilities can enable offensive capabilities to penetrate and compromise an adversary’s systems, while social media can permit attackers to influence perceptions, create false memes, or otherwise disrupt or upend prevailing norms. AI can amplify and enhance the potency and efficacy of these effects. This distinction also implies that there will be variation among third party non-state actors in their relative capabilities to achieve their malevolent objectives in the context of nuclear weapons, which the article will highlight.

16 Kobe, ‘A Theory of Catalytic War’, pp.125–142.

17 For example, see Arthur L. Burns, *The Rationale of Catalytic War, Research Monograph No. 3* (Princeton, NJ: Center of International Studies, Princeton University, 1959); and William T. R. Fox, “Political and Diplomatic Prerequisites of Arms Control,” *Daedalus 89* (1960), pp.1000–1014.

18 Fred C. Ikle, ‘Nth Countries and Disarmament’, *Bulletin of the Atomic Scientists* 16 (1960), p.392.

19 Herman Kahn, ‘The Arms Race and Some of Its Hazards’, *Daedalus* 89 (1960), pp.744–780.

20 For a description of the ‘chain of events’ required for a terrorist nuclear detonation, see Ferguson and Potter, *The Four Faces of Nuclear Terrorism*, p.6.

21 For example, evidence revealed that Al-Qaida actively sought nuclear weapons and expressed a clear desire to use them. Rolf Mowatt-Larssen, ‘Al Qaeda’s pursuit of weapons of mass destruction: the authoritative timeline’, *Foreign Policy*, 25 January 2010.

22 For a recent discussion of various nuclear terrorism, scenarios see Ferguson and Potter, *The Four Faces of Nuclear Terrorism*, and Matthew Bunn and Anthony Wier, ‘Terrorist nuclear weapon construction: how difficult?’ *The Annals of the American Academy*, vol. 6, no. 1 (2006), pp.133–149.

23 Erin D. Dumbacher and Page Soutland, “NTI Nuclear Security Index: Building a Framework for Assurance, Accountability, and Action: Fourth Edition,” *Nuclear Threat Initiative*, September 2018.

24 If the false state accused were able to establish the attackers’ identity unambiguously, and its false provocations, military action against them demonstrate might deliver a clear message of resolve and dis-ownership of the group involved, potentially mitigating nuclear exchange between the nuclear states. Robert Ayson, ‘After a Terrorist Nuclear Attack: Envisaging Catalytic Effects’, *Studies in Conflict & Terrorism*,” 33:7 (2010), p.589 and 593.

25 Toby Dalton and Gaurav Kalwani, ‘Might India Start the Next South Asian Crisis?’ *War on the Rocks*, 1 November 2019.

26 The broader notion that escalation can occur inadvertently (or unintentionally) is attributed to a failure to appreciate the pressures that one’s actions put on an adversary. These pressures generate perceived first-mover advantages and associated with the ‘security dilemma’ concept – when defensively motivated actions can appear and be perceived as offensive. See, Robert Jervis, ‘Cooperation under the Security Dilemma’, *World Politics* 30:2 (1978), pp.167–214.

27 Thomas C. Schelling, *Arms and Influence* (New Haven, CT, and London: Yale University Press, 1966), p.2.

28 A cyber-attack targeting a states NC3 network would require a relatively high level of technical expertise, including: reconnaissance; collecting intelligence; technical knowledge of the NC3 systems being targeted; testing and rehearsing an attack to ensure its efficacy; extensive due diligence on the targets' protocols, strategic culture and cognitive fallibilities; and knowledge of the systems redundancies and countermeasures. Therefore, such an attack would be beyond the abilities and resources of opportunistic or less sophisticated third-party or non-state actors. Innovations in the digital age, such as AI and autonomy, will likely low this threshold and create new mechanisms for non-state actors to fulfill their goals – primarily through manipulating and distortion of the information landscape. See Patricia Lewis and Beyza Unal, *Cybersecurity of nuclear weapons systems: Threats, vulnerabilities, and consequences* (London, United Kingdom: The Royal Institute of International Affairs, Chatham House).

29 A 'false flag' cyber operation in this context would occur due to a cyber-attack that convinces one side of a nuclear dyad that the other is about to attack, thus increasing incentives for pre-emption. Because of the attribution challenge in cyberspace, 'false flag' operations are particularly concerning.

30 Lewis and Unal, *Cybersecurity of nuclear weapons systems: Threats, vulnerabilities and consequences*, pp. 4 and 19; and Peter Hayes, 'Non-State Terrorism and Inadvertent Nuclear War', *Nautilus Institute for Security and Sustainability Special Reports*, 18 January 2018.

31 While SSBN's are not connected to the internet in any meaningful sense, the submarine's missiles, warheads (i.e., launch controls and targeting systems), and the various support systems rely on networked computers and software. Submarines are particularly vulnerable to malware introduced to the network during the procurement phase while the submarine and its missiles and warheads are being built, or when the submarine is in port for maintenance, refurbishment and software updates. Andrew Futter, 'Is Trident safe from cyber-attack?' *European Leadership Network*, February 5 2016, https://www.europeanleadershipnetwork.org/report/is-trident-safe-from-cyber-attack/

32 While theoretically, hackers could acquire submarine launch codes through cyber-espionage and then transmit these 'go codes' to submarine commanders, it is highly unlikely that non-state actors could compromise an SSBN in this way. More likely, however, is that hackers steal highly classified information about SSBN's stealth or navigation technology or information about its locations and patrol itinerary – which will likely be stored or backed-up on the internet. Jason Fritz, 'Hacking Nuclear Command and Control, International Commission on Nuclear Nonproliferation and Disarmament', July 2009, www.icnnd.org/Documents/Jason_Fritz_Hacking_NC2.doc

33 Ibid.

34 Russell Goldman, 'Reading fake news, Pakistani minister directs nuclear threat at Israel', *New York Times*, 24 December 2016.

35 Hannah E. Haegel and Reema Verma, 'The terrifying geography of nuclear and radiological insecurity in South Asia', *Bulletin of Atomic Scientists*, 27 January 2017.

36 Robert Jervis, *How Statesmen Think: The Psychology of International Politics*, (New Jersey, Princeton: Princeton University Press, 2017), p.222.

37 Arthur L. Burns, *The Rationale of Catalytic War*.

38 Eric Schlosser, *Command and Control* (New York, NY: Penguin Group, 2014), p.481.

39 Cordelia Fine, *A Mind of Its Own: How Your Brain Distorts and Deceives* (New York, NY: W. W. Norton & Company, 2008); and Robert Jervis, *Perception and Misperception in International Politics* (Princeton, NJ: Princeton University Press, 1976), pp.117–202.

40 Michela Del Vicario, Antonio Scala, Guido Caldarelli, Eugene H. Stanley, Walter Quattrociocchi, 'Modeling confirmation bias and polarization', *Sci Rep* (2017).

41 Daniel Kahneman, *Thinking, Fast and Slow* (New York: Farrar, Straus and Giroux, 2011).

42 Michael J. Mazarr, *Rethinking Risk in National Security: Lessons from the Financial Crisis for Risk Management* (Basingstoke, UK: Palgrave Macmillan, 2016).

43 Psychologists have demonstrated that in the case of 'anchoring, 'people tend to rely too much on the initial piece of information they receive or learn from while discounting later information. See Daniel Kahneman, *Thinking, Fast and Slow*

44 Amos Tversky and Daniel Kahneman, 'Judgment under Uncertainty: Heuristics and Biases', *Science* 185, no. 4157 (September 1974), pp. 1124–1131.

45 Nassim N. Taleb, *Fooled by Randomness: The Hidden Role of Chance in Life and the Markets*, 2nd ed, (London, UK: Penguin, 2004).

46 For recent research on the influence of specific emotions on decision-making in cyber conflict, see Rose McDermott, "Some emotional considerations in cyber conflict, *Journal of Cyber Policy*, 4:3, (2019), pp.309–325.

47 Forrest E. Morgan, Karl P. Mueller, Evan S. Medeiros, Kevin L. Pollpeter, and Roger Cliff, *Dangerous Thresholds: Managing Escalation in the 21st Century* (Santa Monica, CA: RAND Corporation, 2008), pp.19–29.

48 Lin, 'Escalation Dynamics and Conflict Termination in Cyberspace.'

49 One recent study advanced a more nuanced range of escalation categories: unauthorized use, unintended use, and intended use based on misperceptions. This treatment does not, however, consider the possibility of deliberate use of nuclear weapons. Global Zero, 'Global Zero Commission on Nuclear Risk Reduction: De-Alerting and Stabilizing the World's Nuclear Force Postures', April 2015.

50 Intriligator and Brito, 'Minimizing the Risks of Accidental Nuclear War: An Agenda for Action', pp.230–230.

51 The binary distinction between deliberate and unintentional use of nuclear weapons can be problematic, however. Sico van der Meer, 'Reducing Nuclear Weapons Risks: A Menu of 11 Policy Options', *Policy Brief*, Clingendael Netherlands Institute of International Relations, June 2018.

52 Jervis, *Perception and Misperception in International Politics*, chapter 3.

53 Miles, 'The dynamics of strategic stability and instability', p.429 and 437.

54 Military autonomous systems and inadvertent risks associated with these systems is not a new phenomenon. Examples include landmines, torpedoes, defense systems such as *Aegis* and *Patriot*), and close-in weapon systems such as *Phalanx*. See, Yuna Huh Wong et al., *deterrence in the Age of Thinking Machines* (Santa Monica, CA: RAND Corporation, 2020), pp.71–72.

55 See, Steven Metz and James Kievit, *Strategy and the Revolution in Military Affairs: From Theory to Policy* (Carlisle: Strategic Studies Institute, 1995); Stephen Biddle, *Military Power: Explaining Victory & Defeat in Modern Battle* (Princeton, New Jersey: Princeton University Press, 2004); Stephen Peter Rosen, *Winning the Next War: Innovation and the Modern Military* (Ithaca: Cornell University Press, 1991); MacGregor Knox and Williamson Murray, *The Dynamics of Military Revolution, 1300–2050* (New York: Cambridge University Press, 2001); and Michael C. Horowitz, *The Diffusion of Military Power: Causes and Consequences for International Politics* (Princeton, N.J: Princeton University Press, 2010).

56 Pavel Podvig, 'Risks of Nuclear Command and Control Accidents', in John Borrie, Tim Caughley, and Wilfred Wan (eds), *Understanding Nuclear Weapon Risks*, (Geneva, Switzerland: UNIDIR, 2017).

57 Because of the interdependencies between the various components that comprise NC3 systems, it is impossible to calculate the probability of an accidental nuclear confrontation from the risk of the failure of a particular component or system; in isolation from the system and network as a whole. Intriligator and Brito, 'Minimizing the Risks of Accidental Nuclear War: An Agenda for Action', p.230.

58 Illustrative of the data-fueled complexity and the challenge of integrating interdependent informational functions into a coherent military concept, command and control (C2) morphed into C4ISR (command, control, communications, computers, and ISR).

59 Robotics miniaturization (e.g., CubeSats and swarmed unmanned vehicles) and improvements in networking are enabling the wide deployment of formerly limited capabilities, such as aerial full-motion video, and the harvesting of open-sources (e.g., crowdsourcing, commercial satellite imagery, and geographic information systems (GIS) data), all of which adds to the information burden in the decision-making process. Shane P. Hamilton and Michael P. Kreuzer, 'The Big Data Imperative: Air Force Intelligence for the Information Age', *Air and Space Power Journal* 32, no. 1 (Spring 2018).

60 Ibid.

61 According to a RAND Corporation study, the volume of data collected by the U.S. Navy increased at an exponential rate between 2000 and 2015. Isaac R. Porche, III, *Data Flood: Helping the Navy Address the Rising Tide of Sensor Information* (Santa Monica, CA: RAND Corporation, 2014).

62. Recent studies have identified a range of cognitive and psychological effects whereby subjects tend to discard complex or conflicting information, settle for suboptimal conclusions to save time, and experience high-stress levels and other adverse psychological effects. David Bawden and Lyn Robinson, 'The Dark Side of Information: Overload, Anxiety and Other Paradoxes and Pathologies', *Journal of Information Science* 35, no. 2 (April 2009), pp.180–91.

63 Scott Sagan, *The Limits of Safety; Diane Vaughan, The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA* (Chicago: University of Chicago Press, 1996); Charles Perrow, *Normal Accidents: Living with High-Risk Technologies*, 2nd ed. (Princeton, NJ: Princeton University Press, 1999); Scott A. Snook, *Friendly Fire: The Accidental Shootdown of U.S. Black Hawks over Northern Iraq* (Princeton, NJ: Princeton University Press, 2000); Karl E. Weick and Kathleen M. Sutcliffe, *Managing the Unexpected: Sustained Performance in a Complex World*, 3rd ed. (Hoboken, New Jersey: John Wiley and Sons, 2015).

64 Jervis, *Perception and Misperception in International Politics.*

65 James D. Fearon, 'Rationalist Explanations for War', *International Organization* 49, no. 3 (1995), pp.379–414; and Erik Gartzke, 'War Is in the Error Term', *International Organization* 53, no. 03 (1999), pp.567–87.

66 Robert Jervis, 'War and Misperception', *Journal of Interdisciplinary History* 18:4, (1988), pp.675–700.

67 Stephen J. Cimbala, *The Dead Volcano: The Background and Effects of Nuclear War Complacency*,

(Westport, Connecticut: Praeger, 2002), p.147.

68 For example, China, India, and Pakistan possess multiple types of ground-launched missiles suspected of dual-use. China's DF-26 intermediate-range ballistic missile, for example, is capable of deploying conventional and nuclear payloads. James M. Acton, 'Is It a Nuke? Pre-Launch Ambiguity and Inadvertent Escalation', *Carnegie Endowment for International Peace*, 9 April 2020, https://carnegieendowment.org/2020/04/09/is-it-nuke-pre-launch-ambiguity-and-inadvertent-escalation-pub-81446

69 Ben Buchanan, *The Cybersecurity Dilemma: Hacking, Trust and Fear Between Nations* (Oxford: Oxford University Press, 2017).

70 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* 11 June 2019 https://www.tech4gs.org/leonid-ryabikhin.html

71 Robert Jervis and Mira Rapp-Hooper, 'Perception and Misperception on the Korean Peninsula', *Foreign Affairs*, 5 April 2018. For scholarship on reputation in international relations, and in particular, assuming the worst of other intentions, see Jonathan Mercer, *Reputation and International Politics* (Ithaca, NJ: Cornell University Press, 1996); Robert Jervis, *Perception and Misperception in International Politics*.

72 Gartzke and Lindsay, 'Thermonuclear Cyberwar', pp.37–48.

73 However, these systems must also ensure that states' nuclear arsenals will 'always' be available if authorized (to enhance deterrence) and 'never' be used without authorization (to enhance safety, assurance, and confidence). In this sense, early-warning systems were not designed as safety mechanisms but rather to enable launch-on-warning.

74 The historical record has shown that most international relations situations involve some sort of 'sub-rational' thinking or error, which includes misperception, confusion, and 'bounded rationality.' These scenarios also incorporate a degree of deviation from the logic of rational utility maximization. See, James M. Goldgeier and Philip E. Tetlock, 'Psychology and International Relations Theory', *Annual Review of Political Science* 4, no. 1 (2001), pp.67–92; and Janice Gross Stein, 'The Micro-Foundations of International Relations Theory: Psychology and Behavioral Economics', *International Organization* 71, no. S1 (April 2017), pp.249–63.

75 Charles Perrow, *Normal Accidents: Living with High-Risk Technologies* (Princeton, NJ: Princeton University Press, 1999).

76 Knox Macgregor and Williamson Murray, eds., *The Dynamics of Military Revolution, 1300–2050* (Cambridge, UK: Cambridge University Press, 2001).

77 Richard N. Lebow, *Nuclear Crisis Management* (Ithaca, NY: Cornell University Press, 1987), pp.67–70.

78 For example, during the 1956 Suez Crisis, misinterpretations and the tendency to view unrelated and even unconfirmed events as part of a pattern of an adversary's behavior may have compounded the anxiety and high level of tension contributed to the misinterpretation of detection data from early warnings systems, and in turn, created a crisis with a momentum of its own. Paul Bracken, *The Command and Control of Nuclear Forces* (New Haven: Yale University Press, 1983).

79 Haken Wiberg, 'Accidental Nuclear War: The Problematique', in *Inadvertent Nuclear War*, eds., Hakan Wilberg, Ib Damgaard Petersen, and Paul Smoker (New York, NY: Pergammon Press, 1993), p.17.

80 For example, during the height of the 1962 Cuban Crisis, the United States absorbed a false report (and ignored a credible one) of an imminent attack by the Soviets on the U.S. due to insufficient institutional checks and balances; and despite extensive indications and warning system. Lloyd R. Leavitt and Paul Bracken, 'Nuclear Proliferation: Neither Safe nor Stable', in *Inadvertent Nuclear War*, eds., Hakan Wilberg, Ib Damgaard Petersen, and Paul Smoker (New York, NY: Pergammon Press, 1993), pp.203–217.

81 Open sources indicate that the U.S. DoD already uses several technologies that rely on AI. These include several active programs – either under development or deployed – that could directly affect U.S. NC3 systems and the U.S. nuclear deterrence architecture. These include AI machine learning to support decision-making support systems (e.g., Deep Green); early-warning, ISR data collection analysis, and remote sensing (e.g., Project Maven, Perdix UAV swarm, and Orca XLUUV); cyber and electronic warfare; air-defense systems that use automatic target recognition (ATR) and voice-controlled support assistants in combat aircraft such as the F-15; AI-supported active radar seekers (e.g., Aegis ballistic missile defense systems); and to enable autonomous navigation in unmanned vehicles (e.g., X-47B unmanned aerial vehicle); and machine learning and autonomy to improve precision missile guidance and maneuverability (e.g., Conventional Prompt Strike Guide Vehicles, and long-range hypersonic weapons) Vincent Boulanin (ed.) *Artificial Intelligence, Strategic Stability and Nuclear* (SIPRI Publications, Stockholm: June 2020), pp.42–43.

82 In the context of statistical decision theory, a Type I error (also known as a 'false positive') refers to where a particular action was not taken that should have been taken. A Type II error (also known as a 'false negative') refers to where an action was taken that was not warranted. See, Intriligator and Brito, 'Minimizing the Risks of Accidental Nuclear War: An Agenda for Action', pp.227–230.

83 Both Types I and II errors can cause misperceptions and mischaracterizations between adversaries in ways that may amplify the risk of accidental escalation. Lora Saalman, 'Fear of false negatives: AI and China's nuclear posture', *Bulletin of the Atomic Scientists*, 24 April 2018.

84 See Avery Goldstein, 'First Things First: The Pressing Danger of Crisis Instability in U.S.-China Relations', *International Security* 37, no. 4 (2013), pp.49–89; Caitlin Talmadge, 'Would China Go Nuclear? Assessing the Risk of Chinese Nuclear Escalation in a Conventional War with the United States', *International Security*, Vol. 41, No. 4 (Spring 2017), pp. 50–92; and James Acton, 'Escalation through Entanglement: How the Vulnerability of Command-and-Control Systems Raises the Risks of an Inadvertent Nuclear War', *International Security* 43, no. 1 (1 August 2018), pp.56–99.

85 The uncertainty and ambiguity of states' co-mingled nuclear and non-nuclear weapon systems can lead to mischaracterization that causes false positives or false negatives. See, James Acton, *Is this a Nuke? Pre-launch Ambiguity and Inadvertent Escalation*, p.3.

86 Unexpected accidental situations involving nuclear weapons (or 'broken arrow' incidents) have also included missile explosions, aircraft collisions, and even nuclear weapons release – without an actual nuclear detonation. See, Pavel Podvig, 'Blurring the line between nuclear and non-nuclear weapons: Increasing the risk of accidental nuclear war?' *Bulletin of the Atomic Scientists*, vol. 72, no. 3 (2016), pp.145–149; and Eric Schlosser, *Command and Control* (New York, NY: Penguin Group, 2014).

87 Lin, 'Escalation Dynamics and Conflict Termination in Cyberspace' p.59.

88 'Is Launch Under Attack Feasible?' *Nuclear Threat Initiative*, 4 August 2016.

89 Perrow, *Normal Accidents: Living with High-Risk Technologies*, p. 259.

90 Schelling, *Arms and Influence*, p.234.

91 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*, p.6.

92 Acton, 'Escalation through Entanglement: How the Vulnerability of Command-and-Control Systems Raises the Risks of an Inadvertent Nuclear War', pp.56–99; and Goldstein, 'First Things First: The Pressing Danger of Crisis Instability in U.S.-China Relations', pp.49–89.

93 For example, during the Cold War, the Soviets developed a computer program known as VRYAN (a Russian acronym for 'Surprise Nuclear Missile Attack') 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 feedback loop, which heightened the Kremlin's fear that the United States was pursuing first-strike superiority. President's Foreign Intelligence Advisory Board (PFIAB), "The Soviet 'War Scare,'" 15 February 1990, vi, 24 et seq.

94 Most nuclear-armed states possess modern NC3 structures with inbuilt redundancies, coupled with robust internal ballistics, mechanical arming, and other safety-related mechanisms and protocols. In some cases (i.e., North Korea and Pakistan), however, national NC3 systems may not be rigorously and continuously tested and not function coherently or reliably in functional performance under stress. These NC3 systems are also considered insufficiently redundant and lack clear lines of responsibility and accountability. Peter Hayes, Binoy Kampmark, Philip Reiner, and Deborah Gordon, 'Synthesis Report, NC3 Systems and Strategic Stability: A Global Overview', *NAPSNet Special Reports*, 5 May 2019, https://nautilus.org/?p=97769

95 On 'automation bias' 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), pp.991–1006; and Mary L. Cummings, 'Automation Bias in Intelligent Time-Critical Decision Support Systems', *AIAA 1st Intelligent Systems Technical Conference*, 2004, pp.557–562. On technology and 'organizational failure' see, Bent Natvig, "Accidental Nuclear War Considered from the Area of Reliability of Large Technological Systems, in *Inadvertent Nuclear War*, eds., Hakan Wilberg, Ib Damgaard Petersen, and Paul Smoker (New York, NY: Pergammon Press, 1993), pp.55–69.

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

97 Militaries today have already begun developing AI solutions to substitute human prediction to deal with information overload – or the so-called 'ISR revolution.' See, Keith Dear, 'Artificial Intelligence and Decision-Making', *The RUSI Journal*, 164:5–6, (2019), pp.18–25; and Yang Feilong and Li Shijiang 'Cognitive Warfare: Dominating the Era of Intelligence', *PLA Daily*, 19 March 2020.

98 Paul Bracken, *The Second Nuclear Age: Strategy, Danger, and the New Power Politics* (New York: Times Books, 2012); and Michael Krepon, 'Can Deterrence Ever Be Stable?' *Survival* 57, no. 3 (2015), pp.111–32.

99 Bracken, *The Second Nuclear Age: Strategy, Danger, and the New Power Politics.*

100 Kenneth N. Waltz, *Theory of international politics* (Reading, MA: Addison-Wesley, 1979), p.168.

101 James Johnson, "Artificial Intelligence in Nuclear Warfare: A Perfect Storm of Instability? *The Washington Quarterly* 43:2 (2020), pp.197–211.

102 Drone swarming, as a robotics field, considers large groups of robots typically operate autonomously and coordinate their behavior through decentralized command and control. Working as a collective, the swarm can, in theory, perform both simple and complex tasks in a way that a single robot would be unable to do, thus increasing the robustness and flexibility of the swarm group as a whole. See, Andrew Ilachinski, *AI, Robots, and Swarms, Issues, Questions, and Recommended Studies*, Center for Naval Analysis, January 2017, p.108; and Iñaki Navarro and Fernando Matía, 'An Introduction to Swarm Robotics', *International Scholarly Research Notices* (2013) DOI:10.5402/2013/608,164

103 The question of how other nuclear-armed states – especially allies and partners of the competitive dyad – might respond to a non-state against another member of that group – e.g., NATO states covered by Article 5. Ayson, 'After a Terrorist Nuclear Attack: Envisaging Catalytic Effects', pp.584–585.

104 Conversely, the discovery of an adversary's attempt to degrade a state's nuclear forces or attendant systems would heighten mistrust and tension in future nuclear crises, indirectly contributing to escalation risks. See, Gartzke and Lindsay, 'Thermonuclear Cyberwar', pp.37–48.

105 John J. Mearsheimer, *Conventional Deterrence* (Ithaca, NY: Cornell University Press, 1984), p.210.

106 Thomas J. Christensen, 'The Meaning of the Nuclear Evolution: China's Strategic Modernization and U.S.-China Security Relations', *Journal of Strategic Studies* 35, no. 4 (August 2012), pp.467–71.

107 James Johnson, 'Delegating Strategic Decision-Making to Machines: Dr. Strangelove Redux?' *Journal of Strategic Studies* (2020) DOI: 10.1080/01402390.2020.1759038

108 John J. Mearsheimer, 'The Gathering Storm: China's Challenge to US Power in Asia', *The Chinese Journal of International Politics*, Volume 3, Issue 4 (Winter 2010), pp.381–396.

109 For example, see, Yang Yaohui, 'A Vision of a New Kind of Combat Systems', *PLA Daily*, 20 June 2020, http://www.81.cn/theory/2019-06/20/content_9534942.htm; and Yang Feilong and Li Shijiang 'Cognitive Warfare: Dominating the Era of Intelligence', *PLA Daily*, 19 March 2020.

110 Keren Yarhi-Milo, *Knowing the Adversary* (Princeton NY: Princeton University Press, 2014), p.250.

111 David E. Apter, *The Politics of Modernization* (Chicago, IL: Chicago University Press, 1967).

112 See, Michael J. Mazarr, Ryan Michael Bauer, Abigail Casey, Sarah Heintz, and Luke J. Matthews, *The Emerging Risk of Virtual Societal Warfare: Social Manipulation in a Changing Information Environment* (Santa Monica, CA: RAND Corporation, 2019).

113 Adam P. Liff, 'Cyberwar: a new "absolute weapon"? The proliferation of cyberwarfare capabilities and interstate war', *Journal of Strategic Studies*, 35:3 (2012), pp.401–428.

114 The use of false information, false signals, and elaborate spoofing between adversaries are not entirely new. Roberta Wohlstetter, *Pearl Harbor: Warning and Decision* (Stanford, CA: Stanford University Press, 1962), pp.393–394.

115 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.

116 Allan Dafoe, Jonathan Renshon, Paul Huth, 'Reputation and Status as Motives for War', *Annual Review of Political Science*, 17:1 (2014), pp.371–393.

117 Leavitt and Bracken, 'Nuclear Proliferation: Neither Safe nor Stable', pp.215–216.

118 Robert Jervis, *How Statesmen Think: The Psychology of International Politics*, p.196–197.

119 For example, see Yuan Yi, 'The Development of Military Intelligentization Calls for Related International Rules', *PLA Daily*, 16 October 2019, http://military.workercn.cn/32824/201910/16/191016085645085.shtml; and Lin Juanjuan, Zhang Yuantao, and Wang Wei, 'Military intelligence is profoundly affecting future operations', *Ministry of National Defense of the People's Republic of China*, 10 September 2019.

120 Brian W. Everstine, 'DOD AI Leader Wants Closer Collaboration With NATO', *Airforce Magazine*, 15 January 2020.

121 Bahar Gholipour, 'New AI Tech Can Mimic Any Voice', *Scientific American*, 2 May 2017; Will Knight, 'These Incredibly Realistic Fake Faces Show How Algorithms Can Now Mess with Us', *MIT Technology Review*, 14 December 2018.

122 Generative adversarial networks (GANs) operate by integrating a competitive function into software, with one network seeking to generate an item (i.e., an image or video clip), while the other network judges the item to determine whether it looks real. 'Fake News: You Ain't Seen Nothing Yet', *The Economist*, 1 July 2017.

123 Researchers at the University of Washington are developing AI algorithms that convert audio clips into a realistic video by training software to generate matching mouth shapes on a speaker. Jennifer Langston, 'Lip-Syncing Obama: New Tools Turn Audio Clips into Realistic Video', *UW News*, 11 July 2017.

124 For example, a machine-learning artist at Google successfully generated images of imagined faces with a resolution of 768 pixels aside, more than twice as big as anything previously achieved. Also, AI experts anticipate that amateurs will produce compelling YouTube fake videos within the next few years. 'Fake News: You Ain't Seen Nothing Yet.'

125 Mazarr, Michael J., Ryan Michael Bauer, Abigail Casey, Sarah Heintz, and Luke J. Matthews, *The Emerging Risk of Virtual Societal Warfare: Social Manipulation in a Changing Information Environment* (Santa Monica, CA: RAND Corporation, 2019), p.85.

126 Johnson, 'Delegating Strategic Decision-Making to Machines: Dr. Strangelove Redux?'

127 Adrian Chen, 'The Agency', *New York Times Magazine*, 2 June 2015.

128 See, Martin C. Libicki, *Cyber deterrence and Cyberwar* (Santa Monica, CA: RAND, 2009), p. 44.

129 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.

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

131 Richard Fontaine and Kara Frederick, 'The Autocrat's New Tool Kit', *The Wall Street Journal*, 15 March 2019.

132 Hazel K. Kwon, Chris C. Bang, Michael J. Egnoto and Raghav Rao, 'Social media rumors as an improvised public opinion: Semantic network analyses of Twitter discourses during Korean sabre-rattling 2013', *Asian Journal of Communication* (2016), pp.1–22.

133 Between August 2017 and January 2018, six social media instances 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', p.1.

134 DFRLab, '#FakeNews: American bomber did not drop a bomb on a house in Lithuania', *Medium.com*, 16 June 2017, https://medium.com/dfrlab/fakenews-american-bomber-didnot-drop-a-bomb-on-a-house-in-lithuania-6ae64241fe9e

135 Drew Harwell, 'An Artificial-Intelligence First: Voice-Mimicking Software Reportedly Used in a Major Theft', *Washington Post*, 4 September 2019.

136 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*, 5 June 2004.

137 Kahneman, *Thinking, Fast and Slow.*

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

139 Daniel Kahneman and Amos Tversky, 'Availability: A heuristic for judging frequency and probability', *Cognitive Psychology* 5, no. 2 (September 1973), pp.207–232.

140 Robert Jervis, *How Statesmen Think: The Psychology of International Politics*, p.222.

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

142 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 of the information being conveyed questioned. This information can then be used by AI systems in real-time to determine the messages' influence; thus, learning to become more effective in its task.

143 Alina Polyakova, 'Weapons of the Weak: Russia and AI-Driven Asymmetric Warfare', *Brookings*, 15 November 2018.

144 A non-state third party could relatively easily insert a false narrative into social media or a crowdsourcing platform, with a low probability of increasing the risks of catalytic escalation. For example, non-credible sources or counter-factual information, counter-detection tools, inauspicious timing, or simply that the machine-learning algorithms failed to detect the fakery in time.

145 These fictional scenarios are premised on three assumptions: (1) the involvement of a nuke-armed adversarial dyad adversary during a crisis or period of tension; (2) the existence of a degree of information asymmetry; and (3) the technical feasibility of the operations described – either currently in existence or considered technically viable shortly.

146 Assumptions are needed to explain what an actor might hope to achieve by sparking nuclear confrontation – which would be a sub-rational act. Gartzke and Lindsay, 'Thermonuclear Cyberwar', pp.37–48.

147 While a cyber false flag operation would not require particularly sophisticated technical expertise – at least compared to the level of scientific and military infrastructure required to develop nuclear weapons – it would need the organizational know-how to collect and parse intelligence and conduct clandestine activities.

148 Lin, 'Escalation Dynamics and Conflict Termination in Cyberspace', p.59.

149 From an intelligence standpoint, nuclear solid-fuel missiles and tracked TELs reduce ISR systems' ability to detect signs of launch preparation. Solid fuel also increases the speed that missiles can be launched and reduces the number of support vehicles to support an operation.

150 Alternative outcomes from this fictional scenario are, of course, possible. For example, counter-AI systems might uncover the leak's source or false nature 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 users' ability 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.

151 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.

152 See, Lin, 'Escalation Dynamics and Conflict Termination in Cyberspace', pp.46–70.

153 Micro-drone UAVs range from small flying insects to that of palm-sized devices equivalent to small birds and can carry payloads, including small guns and explosives and unconventional munitions such as poisons and nerve agents. See, Berenice Baker, 'Dogfighting Drones – Swarms of Unmanned Battle-Bots Take to the Skies', Airforce-technology.com, 23 July 2013, www.airforcetechnology.com/features/featuredogfight-drones-unmanned-battle-bot-swarms/; and Juanjuan Lin, Yuantao Zhang, and Wei Wang, 'Military intelligence is profoundly affecting future operations', *Ministry of National Defense of the People's Republic of China*, 10 September 2019, http://www.mod.gov.cn/jmsd/2019-09/10/content_4850148.htm.

154 Because of the rapid diffusion of technologies such as virtual reality goggles, 3D printing, modular payloads, and AI, experts anticipate that within two decades terrorist, and insurgent forces may be able to field on their autonomous UAVs without state sponsorship. See Zachary Kallenborn and Philipp C. Bleek, 'Swarming destruction: drone swarms and chemical, biological, radiological, and nuclear weapons', *The Nonproliferation Review*, 25:5–6 (2019), pp.523–543.

155 Herman Kahn defined forty-four 'rungs' on a complex metaphorical ladder of escalating conflict, ranging from 'sub-crisis maneuvering' to civilization-destroying nuclear exchanges. Herman Kahn, On Escalation: Metaphors and Scenarios (New York, NY: Praeger, 1965). Recently, scholars have questioned the usefulness of Kahn's 'escalation ladder' in the context of the rapid diffusion of asymmetric technologies to signal intent and attack adversaries. For example, see Rebecca Hersman, 'Wormhole Escalation in the New Nuclear Age', *Texas National Security Review*, Vol. 3, Issue 3 (Autumn 2020), pp.99–110.

156 James Johnson, 'China's "Guam Express" and "Carrier Killers": The anti-ship asymmetric challenge to the U.S. in the Western Pacific', *Comparative Strategy*, 36:4, (2017), pp.319–332.

157 Scholars remain divided, however, about the effectiveness of drone strikes in counterterrorism and counterinsurgency activities. Michael J. Boyle, 'The costs and consequences of drone warfare', *International Affairs*, 89:1 (2013), pp.1–29; and Megan Smith and James I. Walsh, 'Do drone strikes degrade Al Qaeda? Evidence from propaganda output', *Terrorism Political Violence* 25:2 (2012), pp.311–27.

158 The U.S. Office of Naval Research (ONR), for example, envisions currently experimenting with swarms of unmanned surface vehicles to form a defensive perimeter around larger ships and surround enemy ships as part of the counterterrorism effort. Thomas Claburn, 'Navy Tests Swarming Autonomous Boats', *InformationWeek*, 7 November 2014.

159 Generative adversarial networks (GANs) is a new approach that involves two artificial neural network systems that spar with each other to create an original image, audio, or video content, which machines have never been able to do properly before. Karen Hao, 'Inside the World of AI that Forges Beautiful Art and Terrifying Deepfakes', *MIT Technology Review*, 1 December 2018.

160 Additional assumptions are needed to explain why the non-state actor expects to benefit from starting a nuclear war, which may not be entirely rational or appear tactically and strategically optimal.

161 Similar to the previous fictional scenario, alternatives to this worse scenario outcome can also be conceived. At the most basic level, at rung of the escalation ladder that led to catastrophe (i.e., the source of the GANs), subsequent subversive tactics could have been detected, and thus, crisis averted.

The case study also assumes the victims of this offensive operation viewed each other as adversaries, making them more suspicious of the others' intentions in the event of a crisis. Were *either* side able to demonstrate empathy or display restraint, this escalatory outcome may be avoided.

162 Scott Rosenberg, 'Firewalls Don't Stop Hackers, AI Might', *Wired*, 27 August 2017.

163 Andrew Futter, *Hacking the Bomb: Cyber Threats and Nuclear Weapons*, (Washington DC: Georgetown University Press, 2018), p.10.

164 Traditional deterrence theory maintains that for deterrence to succeed, all parties in a conflict or crisis must be interested in limiting their costs and credibly threaten their opponents with punishment should they attempt to escalate – a condition that seldom exists in situations between states and non-state adversaries. Morgan, Mueller, Medeiros, Pollpeter, and Cliff, *Dangerous Thresholds: Managing Escalation in the 21st Century*, pp.173–174.

165 For example, states can deter terrorists from attacking specific targets such as NC3 systems by protecting those targets more aggressively, thus *denying* adversaries the ability to escalate along certain lines. Ibid.

166 The approach used to control or mitigate escalation risk will vary according to the class of escalatory mechanisms involved (i.e., intention vs. unintentional); these approaches might also conflict with each other. Podvig, 'Risks of Nuclear Command and Control Accidents;' Wilfred Wan, *Nuclear Risk Reduction: A Framework for Analysis* (Geneva, Switzerland: UNIDIR, 2019).

167 Budget constraints, design modifications, and bureaucratic inertia have hampered previous PAL devices and other safety components to nuclear management systems. Eric Schlosser, *Command and Control* (New York, NY: Penguin Group, 2014).

168 For example, in 2017, NATO established a Strategic Communications Center of Excellence, which supports the development of best practices designed to raise awareness of the risks of disinformation posed by the nefarious dissemination of misinformation. NATO Takes Aim at Disinformation Campaigns," *NPR Morning Edition*, 10 May 2017.

169 Peter Maurer, *Nuclear Weapons: Averting a Global Catastrophe, ICRC Statement*, 23 April 2018.

170 For example, the implementation of best practices to reduce the risk of human fallibilities and measures to extend the decision-making, allowing for clarification of early warnings readings that turn out to be erroneous. U.S. Office of the Director of National Intelligence, *A Guide to Cyber Attribution*, 14 September 2018; Matthew Bunn, Nickolas Roth, and William H. Tobey, 'Revitalizing Nuclear Security in an Era of Uncertainty', *Project on Managing the Atom*, January 2019.

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

172 Perrow, *Normal Accidents: Living with High-Risk Technologies*; and Unal and Lewis, *Cybersecurity of Nuclear Weapons Systems: Threats, Vulnerabilities, and Consequences*.

173 Sico van der Meer, 'Reducing Nuclear Weapons Risks: A Menu of 11 Policy Options', *Policy Brief, Clingendael: Netherlands Institute of International Relations*, June 2018.

174 For example, the 'Memorandum of Agreement Between the United States of America and The Russian Federation on the Establishment of a Joint Center for the Exchange of Data from Early Warning Systems and Notifications of Missile Launches', *The U.S. Department of State*, 4 June 2000.

175 National Security Commission on Artificial Intelligence (NSCAI) Interim Report to Congress, November 2019, p.46.

176 For a recent multilateral (non-binding) effort, see the 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, 24 June 2013.

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

178 Maas, 'How viable is international arms control for artificial military intelligence? Three lessons from nuclear weapons', pp.285–311.

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

180 The 'AI control problem' refers to the issue that, under certain conditions, AI systems can learn in unexpected and counterintuitive ways that engineers and operators may behave expected or align with their goals. See Stuart Russell, *Human Compatible* (New York, NY: Penguin Random House, 2019), p.251.

181 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.

Bibliography

- Acton, James , 'Escalation through Entanglement: How the Vulnerability of Command-and-Control Systems Raises the Risks of an Inadvertent Nuclear War', *International Security* 43/1 (August 1 2018), 56–99. doi:10.1162/isec_a_00320 [Crossref], [Web of Science ®], [Google Scholar]
- Acton, James , *Is This a Nuke? Pre-launch Ambiguity and Inadvertent Escalation* (New York, NY: Carnegie Endowment for International Peace 2020). [Google Scholar]
- Acton, James M. , 'Is It a Nuke? Pre-Launch Ambiguity and Inadvertent Escalation', *Carnegie Endowment for International Peace* , 9 April 2020, https://carnegieendowment.org/2020/04/09/is-it-nuke-pre-launch-ambiguity-and-inadvertent-escalation-pub-81446 [Google Scholar]
- Allison, Graham , *Nuclear Terrorism: The Ultimate Preventable Catastrophe* (New York: Owl Books 2004). [Google Scholar]
- Apter, David E. , *The Politics of Modernization* (Chicago, IL: Chicago University Press 1967). [Google Scholar]
- Ayson, Robert , 'After a Terrorist Nuclear Attack: Envisaging Catalytic Effects'. *Studies in Conflict & Terrorism* 33/7 (2010), 571–93. doi:10.1080/1057610X.2010.483756 [Taylor & Francis Online], [Web of Science ®], [Google Scholar]
- Baker, Berenice , 'Dogfighting Drones - Swarms of Unmanned Battle-Bots Take to the Skies', *Airforce-*

technology.com , 23 July 2013, www.airforcetechnology.com/features/featuredogfight-drones-unmanned-battle-bot-swarms/ [Google Scholar]

- Bawden, David and RobinsonLyn , 'The Dark Side of Information: Overload, Anxiety and Other Paradoxes and Pathologies'. *Journal of Information Science* 35/2 (April 2009), 180–91. doi:10.1177/0165551508095781 [Crossref], [Web of Science ®], [Google Scholar]
- Biddle, Stephen , *Military Power: Explaining Victory & Defeat in Modern Battle* (Princeton, New Jersey: Princeton University Press 2004). [Crossref], [Google Scholar]
- Boulanin, Vincent , ed., *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk Vol. I Euro-Atlantic Perspectives* (Stockholm: SIPRI Publications May 2019). [Google Scholar]
- Boulanin, Vincent , ed., *Artificial Intelligence, Strategic Stability and Nuclear* (Stockholm: SIPRI Publications June 2020). [Google Scholar]
- Boyle, Michael J. , 'The Costs and Consequences of Drone Warfare'. *International Affairs* 89/1 (2013), 1–29. doi:10.1111/1468-2346.12002 [Crossref], [Web of Science ®], [Google Scholar]
- Bracken, Paul , *The Command and Control of Nuclear Forces* (New Haven: Yale University Press 1983). [Google Scholar]
- Bracken, Paul , *The Second Nuclear Age: Strategy, Danger, and the New Power Politics* (New York: Times Books 2012). [Google Scholar]
- Brody, Richard A. , 'Deterrence Strategies: An Annotated Bibliography', *The Journal of Conflict Resolution* 4 (1960), 443–57. doi:10.1177/002200276000400406 [Crossref], [Google Scholar]
- Buchanan, Ben , *The Cybersecurity Dilemma: Hacking, Trust and Fear between Nations* (Oxford: Oxford University Press 2017). [Crossref], [Google Scholar]
- Bunn, Matthew , NickolasRoth, and William H.Tobey , 'Revitalizing Nuclear Security in an Era of Uncertainty', *Project on Managing the Atom* , January 2019. [Google Scholar]
- Bunn, Matthew and AnthonyWier , 'Terrorist Nuclear Weapon Construction: How Difficult?'. *The Annals of the American Academy* 6/1 (2006), 133–49. doi:10.1177/0002716206290260 [Crossref], [Google Scholar]
- Burns, Arthur L. , *The Rationale of Catalytic War, Research Monograph No. 3* (Princeton, NJ: Center of International Studies, Princeton University 1959). [Google Scholar]
- Charles, Fred , 'Nth Countries and Disarmament', *Bulletin of the Atomic Scientists* 16 (1960), 391–94. doi:10.1080/00963402.1960.11454156 [Taylor & Francis Online], [Web of Science ®], [Google Scholar]
- Chen, Adrian , 'The Agency', *New York Times Magazine* , 2 June 2015, https://www.nytimes.com/2015/06/07/magazine/the-agency.html [Google Scholar]

- Christensen, Thomas J. , 'The Meaning of the Nuclear Evolution: China's Strategic Modernization and U.S.-China Security Relations'. *Journal of Strategic Studies* 35/4 (August 2012), 467–71. doi:10.1080/01402390.2012.714710 [Taylor & Francis Online], [Web of Science ®], [Google Scholar]
- Cimbala, Stephen J. , *The Dead Volcano: The Background and Effects of Nuclear War Complacency* (Westport, Connecticut: Praeger 2002). [Google Scholar]
- Citron, Danielle K. and RobertChesney , *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security* (107 California Law Review 1753 2019). [Google Scholar]
- Cummings, Mary L. , 'Automation Bias in Intelligent Time-Critical Decision Support Systems', AIAA 1st Intelligent Systems Technical Conference, Chicago, IL, 2004, pp.557–62. [Crossref], [Google Scholar]
- Dafoe, Allan , JonathanRenshon, and PaulHuth , 'Reputation and Status as Motives for War'. *Annual Review of Political Science* 17/1 (2014), 371–93. doi:10.1146/annurev-polisci-071112-213421 [Crossref], [Google Scholar]
- Dalton, Toby and GauravKalwani , 'Might India Start the Next South Asian Crisis?" *War on the Rocks* , 1 November 2019. [Google Scholar]
- Dear, Keith , 'Artificial Intelligence and Decision-Making'. *The RUSI Journal* 164/5–6 (2019), 18–25. doi:10.1080/03071847.2019.1693801 [Taylor & Francis Online], [Web of Science ®], [Google Scholar]
- DFRLab , '#fakenews: American Bomber Did Not Drop a Bomb on a House in Lithuania', *Medium.com* , 16 June 2017, https://medium.com/dfrlab/fakenews-american-bomber-didnot-drop-a-bomb-on-a-house-in-lithuania-6ae64241fe9e [Google Scholar]
- Dumbacher, Erin D. and PageSoutland , 'NTI Nuclear Security Index: Building a Framework for Assurance, Accountability, and Action: Fourth Edition', *Nuclear Threat Initiative* , September 2018. [Google Scholar]
- Erik, Gartzke and Jon R.Lindsay , 'Thermonuclear Cyberwar'. *Journal of Cybersecurity* 3/1 (February 2017), 37–48. [Web of Science ®], [Google Scholar]
- Everstine, Brian W. "DOD AI Leader Wants Closer Collaboration With NATO', *Airforce Magazine* , 15 January 2020. [Google Scholar]
- 'Fake News: You Ain't Seen Nothing Yet', *The Economist* , 1 July 2017, https://www.economist.com/science-and-technology/2017/07/01/fake-news-you-aint-seen-nothing-yet [Google Scholar]
- Fearon, James D. , 'Rationalist Explanations for War'. *International Organization* 49/3 (1995), 379–414. doi:10.1017/S0020818300033324 [Crossref], [Web of Science ®], [Google Scholar]
- Fearon, James D. , 'Cooperation, Conflict, and the Costs of Anarchy'. *International Organization* 72/3 (2018), 523–59. doi:10.1017/S0020818318000115 [Crossref], [Web of Science ®], [Google Scholar]

- Feilong, Yang and LiShijiang , 'Cognitive Warfare: Dominating the Era of Intelligence', *PLA Daily* , 19 March 2020. [Google Scholar]
- Ferguson, Charles C. and William C.Potter , *The Four Faces of Nuclear Terrorism* (Monterey, CA: Center for Nonproliferation Studies and Nuclear Threat Initiative 2004). [Google Scholar]
- Fine, Cordelia , *A Mind of Its Own: How Your Brain Distorts and Deceives* (New York, NY: W. W. Norton & Company 2008). [Google Scholar]
- Fitzpatrick, Mark , *The World After: Proliferation, Deterrence and Disarmament if the Nuclear Taboo Is Broken* (Paris: Ifri Security Studies Centre 2009). [Google Scholar]
- Fontaine, Richard and KaraFrederick , 'The Autocrat's New Tool Kit', *The Wall Street Journal* , 15 March 2019, https://www.wsj.com/articles/the-autocrats-new-tool-kit-11552662637 [Google Scholar]
- Forest, James , 'Framework for Analyzing the Future Threat of WMD Terrorism'. *Journal of Strategic Security* 5/4 (2012), 51–68. doi:10.5038/1944-0472.5.4.4 [Crossref], [Google Scholar]
- Fox, William T. R. , 'Political and Diplomatic Prerequisites of Arms Control', *Daedalus* 89 (1960), 1000–14. [Web of Science ®], [Google Scholar]
- Fritz, Jason , 'Hacking Nuclear Command and Control, International Commission on Nuclear Nonproliferation and Disarmament', July 2009, www.icnnd.org/Documents/Jason_Fritz_Hacking_NC2.doc [Google Scholar]
- Futter, Andrew , 'Is Trident Safe from Cyber-attack?" *European Leadership Network* , 5 February 2016, https://www.europeanleadershipnetwork.org/report/is-trident-safe-from-cyber-attack/ [Google Scholar]
- Futter, Andrew , *Hacking the Bomb: Cyber Threats and Nuclear Weapons* (Washington DC: Georgetown University Press 2018). [Google Scholar]
- Garfinkel, Ben and AllanDafoe , 'How Does the Offense-defense Balance Scale?'. *Journal of Strategic Studies* 42/6 (2019), 736–63. doi:10.1080/01402390.2019.1631810 [Taylor & Francis Online], [Web of Science ®], [Google Scholar]
- Gartzke, Erik , 'War Is in the Error Term'. *International Organization* 53/3 (1999), 567–87. doi:10.1162/002081899550995 [Crossref], [Web of Science ®], [Google Scholar]
- Geist, Edward and AndrewLohn , *How Might Artificial Intelligence Affect the Risk of Nuclear War?* (Santa Monica, CA: RAND Corporation 2018). [Crossref], [Google Scholar]
- Gholipour, Bahar , 'New AI Tech Can Mimic Any Voice', *Scientific American* , 2 May 2017. [Google Scholar]
- Goldgeier, James M. and Philip E.Tetlock , 'Psychology and International Relations Theory'. *Annual Review of Political Science* 4/1 (2001), 67–92. doi:10.1146/annurev.polisci.4.1.67 [Crossref], [Web of Science ®], [Google Scholar]

- Goldman, Russell , ‘Reading Fake News, Pakistani Minister Directs Nuclear Threat at Israel’, *New York. Times* , 24 December 2016, https://www.nytimes.com/2016/12/24/world/asia/pakistan-israel-khawaja-asif-fake-news-nuclear.html [Google Scholar]
- Goldstein, Avery , ‘First Things First: The Pressing Danger of Crisis Instability in U.S.-China Relations’. *International Security* 37/4 (2013), 49–89. doi:10.1162/ISEC_a_00114 [Crossref], [Web of Science ®], [Google Scholar]
- Hadfield-Menell, Dylan , AncaDragan, PieterAbbeel, and StuartRussell , ‘Cooperative Inverse Reinforcement Learning’, 30th Conference on Neural Information Processing Systems, Barcelona, Spain, 2016 NIPS. [Google Scholar]
- Haegel, Hannah E. and ReemaVerma , ‘The Terrifying Geography of Nuclear and Radiological Insecurity in South Asia’, *Bulletin of Atomic Scientists* , 27 January 2017. [PubMed], [Google Scholar]
- Hamilton, Shane P. and Michael P.Kreuzer , ‘The Big Data Imperative: Air Force Intelligence for the Information Age’, *Air and Space Power Journal* 32/1 (Spring 2018), 4–20. [Google Scholar]
- Hao, Karen , ‘Inside the World of AI that Forges Beautiful Art and Terrifying Deepfakes’, *MIT Technology Review* , 1 December, 2018, https://www.technologyreview.com/2018/12/01/138847/inside-the-world-of-ai-that-forges-beautiful-art-and-terrifying-deepfakes/ [Google Scholar]
- Harwell, Drew , ‘An Artificial-Intelligence First: Voice-Mimicking Software Reportedly Used in a Major Theft’, *Washington Post* , 4 September 2019. [Google Scholar]
- Hayes, Peter , ‘Non-State Terrorism and Inadvertent Nuclear War’, *Nautilus Institute for Security and Sustainability Special Reports* , 18 January 2018. [Google Scholar]
- Hayes, Peter , BinoyKampmark, PhilipReiner, and DeborahGordon , ‘Synthesis Report, NC3 Systems and Strategic Stability: A Global Overview’, *NAPSNet Special Reports* , 5 May 2019, https://nautilus.org/?p=97769 [Google Scholar]
- Herman, Kahn , ‘The Arms Race and Some of Its Hazards’, *Dxdalus* 89 (1960), 744–80. [Web of Science ®], [Google Scholar]
- Hersman, Rebecca , ‘Wormhole Escalation in the New Nuclear Age’, *Texas National Security Review* 3/3 (Autumn 2020), 99–110. [Google Scholar]
- Horowitz, Michael C. , *The Diffusion of Military Power: Causes and Consequences for International Politics* (Princeton, N.J: Princeton University Press 2010). [Crossref], [Google Scholar]
- Horowitz, Michael C. , PaulScharre, and AlexanderVelez-Green , ‘A Stable Nuclear Future? the Impact of Autonomous Systems and Artificial Intelligence’, *arXiv* (December 2019). https://arxiv.org/abs/1912.05291 [Google Scholar]
- Ikle, C. Fred , HansSpeier, BernardBrodie, Alexander L.George, L. AliceHsieh, and KramishArnold , *The Diffusion of Nuclear Weapons to Additional Countries: The “Nth*

Country" Problem (Santa Monica, CA: RAND Corporation 1960). [Crossref], [Google Scholar]

- Ilachinski, Andrew , *AI, Robots, and Swarms, Issues, Questions, and Recommended Studies* (Washington, DC: Center for Naval Analysis January 2017). [Google Scholar]
- Intriligator Michael, D. and Dagobert L.Brito , 'Minimizing the Risks of Accidental Nuclear War: An Agenda for Action', in WilbergHakan, Ib DamgaardPetersen, and PaulSmoker (eds.), *Inadvertent Nuclear War* (New York, NY: Pergammon Press 1993), 228. [Crossref], [Google Scholar]
- James, Johnson , 'Artificial Intelligence & Future Warfare: Implications for International Security'. *Defense & Security Analysis* 35/2 (2019), 147–69. doi:10.1080/14751798.2019.1600800 [Taylor & Francis Online], [Web of Science ®], [Google Scholar]
- Jervis, Robert , *Perception and Misperception in International Politics* (Princeton, NJ: Princeton University Press 1976). [Google Scholar]
- Jervis, Robert , 'Cooperation under the Security Dilemma'. *World Politics* 30/2 (1978), 167–214. doi:10.2307/2009958 [Crossref], [Web of Science ®], [Google Scholar]
- Jervis, Robert , 'War and Misperception'. *Journal of Interdisciplinary History* 18/4 (1988), 675–700. doi:10.2307/204820 [Crossref], [Web of Science ®], [Google Scholar]
- Jervis, Robert , *How Statesmen Think: The Psychology of International Politics* (New Jersey, Princeton: Princeton University Press 2017). [Google Scholar]
- Jervis, Robert and MiraRapp-Hooper , 'Perception and Misperception on the Korean Peninsula', *Foreign Affairs* , 5 April 2018. [Google Scholar]
- Johnson, James , 'China's "Guam Express" and "Carrier Killers": The Anti-ship Asymmetric Challenge to the U.S. In the Western Pacific'. *Comparative Strategy* 36/4 (2017), 319–32. doi:10.1080/01495933.2017.1361204 [Taylor & Francis Online], [Google Scholar]
- Johnson, James , 'Artificial Intelligence in Nuclear Warfare: A Perfect Storm of Instability?'. *The Washington Quarterly* 43/2 (2020), 197–211. doi:10.1080/0163660X.2020.1770968 [Taylor & Francis Online], [Web of Science ®], [Google Scholar]
- Johnson, James , 'Delegating Strategic Decision-Making to Machines: Dr. Strangelove Redux?'. *Journal of Strategic Studies* (2020), 1–39. doi:10.1080/01402390.2020.1759038 [Taylor & Francis Online], [Web of Science ®], [Google Scholar]
- Johnson, James , 'The End of Military-techno Pax Americana? Washington's Strategic Responses to Chinese AI-enabled Military Technology', *The Pacific Review* (2019). doi:10.1080/09512748.2019.1676299 [Taylor & Francis Online], [Web of Science ®], [Google Scholar]

- Juanjuan, Lin , ZhangYuantao, and WangWei , 'Military Intelligence Is Profoundly Affecting Future Operations', *Ministry of National Defense of the People's Republic of China* , 10 September 2019. [Google Scholar]
- Kahneman, Daniel , *Thinking, Fast and Slow* (New York: Farrar, Straus and Giroux 2011). [Google Scholar]
- Kahneman, Daniel and AmosTversky , 'Availability: A Heuristic for Judging Frequency and Probability'. *Cognitive Psychology* 5/2 (September 1973), 207–32. doi:10.1016/0010-0285(73)90033-9 [Crossref], [Web of Science ®], [Google Scholar]
- Kallenborn, Zachary and Philipp C.Bleek , 'Swarming Destruction: Drone Swarms and Chemical, Biological, Radiological, and Nuclear Weapons'. *The Nonproliferation Review* 25/5–6 (2019), 523–43. doi:10.1080/10736700.2018.1546902 [Taylor & Francis Online], [Google Scholar]
- Kissinger, Henry A. , *The Necessity for Choice* (New York: Harper & Bros. 1960). [Google Scholar]
- Knight, Will , 'The US Military Is Funding an Effort to Catch Deepfakes and Other AI Trickery', *MIT Technology Review* , 23 May 2018, https://www.technologyreview.com/2018/05/23/142770/the-us-military-is-funding-an-effort-to-catch-deepfakes-and-other-ai-trickery/. [Google Scholar]
- Knight, Will , 'These Incredibly Realistic Fake Faces Show How Algorithms Can Now Mess with Us', *MIT Technology Review* , 14 December 2018, https://www.technologyreview.com/2018/12/14/103659/these-incredibly-real-fake-faces-show-how-algorithms-can-now-mess-with-us/ [Google Scholar]
- Knox, MacGregor and WilliamsonMurray , *The Dynamics of Military Revolution, 1300-2050* (New York: Cambridge University Press 2001). [Crossref], [Google Scholar]
- Kobe, Donald H , 'A Theory of Catalytic War'. *Journal of Conflict Resolution* 6/2 (1962), 125–42. doi:10.1177/002200276200600203 [Crossref], [Google Scholar]
- Krepon, Michael , 'Can Deterrence Ever Be Stable?'. *Survival* 57/3 (2015), 111–32. doi:10.1080/00396338.2015.1046228 [Taylor & Francis Online], [Web of Science ®], [Google Scholar]
- Kwon, Hazel K. , Chris C.Bang, Michael J.Egnoto, and RaghavRao , 'Social Media Rumors as an Improvised Public Opinion: Semantic Network Analyses of Twitter Discourses during Korean Saber-rattling 2013'. *Asian Journal of Communication* 26/3 (2016), 1–22. [Web of Science ®], [Google Scholar]
- Lamothe, Dan , 'U.S. Families Got Fake Orders to Leave South Korea. Now Counterintelligence Is Involved', *The Washington Post* , 22 September 2017, https://www.washingtonpost.com/gdpr-consent/?next_url=https%3a%2f%2fwww.washingtonpost.com%2fnews%2fcheckpoint%2fwp%2f2017%2f09%2f22%2fu-s-families-got-fake-orders-to-leave-south-korea-now-counterintelligence-is-involved%2f [Google Scholar]

- Langston, Langston , ‘Lip-Syncing Obama: New Tools Turn Audio Clips into Realistic Video’, *UW News* , 11 July 2017, https://www.washington.edu/news/2017/07/11/lip-syncing-obama-new-tools-turn-audio-clips-into-realistic-video/ [Google Scholar]
- Leavitt, Lloyd R. and PaulBracken , ‘Nuclear Proliferation: Neither Safe nor Stable’, in HakanWilberg, Ib DamgaardPetersen, and PaulSmoker (eds.), *Inadvertent Nuclear War* (New York, NY: Pergammon Press 1993), 203–17. [Crossref], [Google Scholar]
- Lebow, Richard N. , *Nuclear Crisis Management* (Ithaca, NY: Cornell University Press 1987). [Google Scholar]
- Leghorn, Richard S. , et al., ‘The Problem of Accidental War’. *Bulletin of the Atomic Scientists* 14 (1958), 221–58. doi:10.1080/00963402.1958.11453851 [Taylor & Francis Online], [Web of Science ®], [Google Scholar]
- Lehman, Joel et al., ‘The Surprising Creativity of Digital Evolution, A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities’, 14 August 2018, arXiv:1803.03453 https://arxiv.org/abs/1803.03453 [Crossref], [Google Scholar]
- Leonid, Ryabikhin , ‘Russia’s NC3 and Early Warning Systems’, *Tech4GS* , 11 June 2019 https://www.tech4gs.org/leonid-ryabikhin.html [Google Scholar]
- Libicki, Martin C. , *Cyber Deterrence and Cyberwar* (Santa Monica, CA: RAND 2009). [Google Scholar]
- Liff, Adam P. , ‘Cyberwar: A New ‘Absolute Weapon’? the Proliferation of Cyberwarfare Capabilities and Interstate War’. *Journal of Strategic Studies* 35/3 (2012), 401–28. doi:10.1080/01402390.2012.663252 [Taylor & Francis Online], [Web of Science ®], [Google Scholar]
- Lin, Herbert , ‘Escalation Dynamics and Conflict Termination in Cyberspace’, *Strategic Studies Quarterly* 6/3 (Fall 2012), 46–70. [Google Scholar]
- Lin, Zhang Yuantao and WangWei “Military Intelligence Is Profoundly Affecting Future Operations’, *Ministry of National Defense of the People’s Republic of China* , 10 September 2019, http://www.mod.gov.cn/jmsd/2019-09/10/content_4850148.htm. [Google Scholar]
- Maas, Matthijs , ‘How Viable Is International Arms Control for Artificial Military Intelligence? Three Lessons from Nuclear Weapons’. *Contemporary Security Policy* 40/3 (2019), 285–311. doi:10.1080/13523260.2019.1576464 [Taylor & Francis Online], [Web of Science ®], [Google Scholar]
- Macgregor, Knox and WilliamsonMurray , eds., *The Dynamics of Military Revolution, 1300–2050* (Cambridge, UK: Cambridge University Press 2001). [Google Scholar]
- Mazarr, Michael J. , Ryan M.Bauer, AbigailCasey, SarahHeintz, and Luke J.Matthews , *The Emerging Risk of Virtual Societal Warfare: Social Manipulation in a Changing Information Environment* (Santa Monica, CA: RAND Corporation 2019). [Crossref], [Google Scholar]

- McDermott, Rose , ‘Some Emotional Considerations in Cyber Conflict’. *Journal of Cyber Policy* 4/3 (2019), 309–25. doi:10.1080/23738871.2019.1701692 [Taylor & Francis Online], [Google Scholar]
- Mearsheimer, John J. , *Conventional Deterrence* (Ithaca, NY: Cornell University Press 1984). [Google Scholar]
- Mearsheimer, John J. , ‘The Gathering Storm: China’s Challenge to US Power in Asia’. *The Chinese Journal of International Politics* 3/4 (Winter 2010), 381–96. doi:10.1093/cjip/poq016 [Crossref], [Web of Science ®], [Google Scholar]
- “Memorandum of Agreement between the United States of America and the Russian Federation on the Establishment of a Joint Center for the Exchange of Data from Early Warning Systems and Notifications of Missile Launches’, *The U.S. Department of State* , 4 June 2000. [Google Scholar]
- Mercer, Jonathan , *Reputation and International Politics* (Ithaca, NJ: Cornell University Press 1996). [Google Scholar]
- Metz, Steven and JamesKievit , *Strategy and the Revolution in Military Affairs: From Theory to Policy* (Carlisle: Strategic Studies Institute 1995). [Crossref], [Google Scholar]
- Morgan, Forrest E. , Karl P.Mueller, Evan S.Medeiros, Kevin L.Pollpeter, and RogerCliff , *Dangerous Thresholds: Managing Escalation in the 21st Century* (Santa Monica, CA: RAND Corporation 2008). [Google Scholar]
- Mowatt-Larssen, Rolf , ‘Al Qaeda’s Pursuit of Weapons of Mass Destruction: The Authoritative Timeline’, *Foreign Policy* , 25 January 2010. [Google Scholar]
- National Security Commission on Artificial Intelligence (NSCAI) , ‘Interim Report to Congress’, November 2019. https://www.nationaldefensemagazine.org/-/media/sites/magazine/03_linkedfiles/nscai-interim-report-for-congress.ashx [Google Scholar]
- ‘NATO Takes Aim at Disinformation Campaigns’, *NPR Morning Edition* , 10 May 2017, https://choice.npr.org/index.html?origin=https://www.npr.org/2017/05/10/527720078/nato-takes-aim-at-disinformation-campaigns [Google Scholar]
- Natvig, Bent , ‘Accidental Nuclear War Considered from the Area of Reliability of Large Technological Systems’, in PetersenWilberg and Smoker (eds.), *Inadvertent Nuclear War* (New York, NY: Pergammon Press 1993), 55–69. [Crossref], [Google Scholar]
- 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* . [Google Scholar]
- Navarro, Iñaki and FernandoMatía , ‘An Introduction to Swarm Robotics’, *International Scholarly Research Notices* (2013). doi:10.5402/2013/608164 [Crossref], [Google Scholar]
- Nilsson, Nils J. , *The Quest for Artificial Intelligence: A History of Ideas and Achievements* (Cambridge: Cambridge University Press 2010). [Google Scholar]

- Nye, Joseph S. , ‘Deterrence and Dissuasion in Cyberspace’. *International Security* 41/3 (2017), 44–71. doi:10.1162/ISEC_a_00266 [Crossref], [Web of Science ®], [Google Scholar]
- Patricia, Lewis and BeyzaUnal , *Cybersecurity of Nuclear Weapons Systems: Threats, Vulnerabilities and Consequences* (London, United Kingdom: The Royal Institute of International Affairs, Chatham House 2019). [Google Scholar]
- Perrow, Charles , *Normal Accidents: Living with High-Risk Technologies* , 2nd ed. (Princeton, NJ: Princeton University Press 1999). [Google Scholar]
- Podvig, Pavel , ‘Blurring the Line between Nuclear and Non-nuclear Weapons: Increasing the Risk of Accidental Nuclear War?’. *Bulletin of the Atomic Scientists* 72/3 (2016), 145–49. doi:10.1080/00963402.2016.1170363 [Taylor & Francis Online], [Web of Science ®], [Google Scholar]
- Polyakov, Alina , ‘Weapons of the Weak: Russia and AI-Driven Asymmetric Warfare’, *Brookings* , 15 November, 2018, https://www.brookings.edu/research/weapons-of-the-weak-russia-and-ai-driven-asymmetric-warfare/ [Google Scholar]
- Porche, Isaac R., III , et al., *Data Flood: Helping the Navy Address the Rising Tide of Sensor Information* (Santa Monica, CA: RAND Corporation 2014). [Crossref], [Google Scholar]
- President’s Foreign Intelligence Advisory Board (PFIAB) , ‘The Soviet ‘War Scare’, 15 February 1990, vi, 24 et seq. [Google Scholar]
- Rosen, Stephen P. , *Winning the Next War: Innovation and the Modern Military* (Ithaca: Cornell University Press 1991). [Google Scholar]
- Rosenberg, Scott , ‘Firewalls Don’t Stop Hackers, AI Might’, *Wired* , 27 August 2017, https://www.wired.com/story/firewalls-dont-stop-hackers-ai-might/ [Google Scholar]
- Rubin, Jamie , ‘Deception: The Other ‘D’ in D-Day’, *NBC News* , 5 June 2004. https://www.nbcnews.com/id/wbna5139053 [Google Scholar]
- Russell, Stuart , *Human Compatible* (New York, NY: Penguin Random House 2019). [Google Scholar]
- Saalman, Lora , ‘Fear of False Negatives: AI and China’s Nuclear Posture’, *Bulletin of the Atomic Scientists* , 24 April 2018. [Google Scholar]
- Sagan, Scott , *The Limits of Safety; Diane Vaughan, the Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA* (Chicago: University of Chicago Press 1996). [Google Scholar]
- Schelling, Thomas C. , *Arms and Influence* (New Haven, CT, and London: Yale University Press 1966). [Google Scholar]
- Schlosser, Eric , *Command and Control* (New York, NY: Penguin Group 2014). [Google Scholar]

- Sechser, Todd S. , NeilNarang, and CaitlinTalmadge , 'Emerging Technologies and Strategic Stability in Peacetime, Crisis, and War'. *Journal of Strategic Studies* 42/6 (2019), 727–35. doi:10.1080/01402390.2019.1626725 [Taylor & Francis Online], [Web of Science ®], [Google Scholar]
- Skitka, Linda J , Kathleen LMosier, and MarkBurdick , 'Does Automation Bias Decision-Making?'. *International Journal of Human-computer Studies* 51/5 (1999), 991–1006. doi:10.1006/ijhc.1999.0252 [Crossref], [Web of Science ®], [Google Scholar]
- Smith, Megan and James I.Walsh , 'Do Drone Strikes Degrade Al Qaeda? Evidence from Propaganda Output'. *Terrorism Political Violence* 25/2 (2012), 311–27. doi:10.1080/09546553.2012.664011 [Taylor & Francis Online], [Web of Science ®], [Google Scholar]
- Snook, Scott A. , *Friendly Fire: The Accidental Shootdown of U.S. Black Hawks over Northern Iraq* (Princeton, NJ: Princeton University Press 2000). [Crossref], [Google Scholar]
- Stein, Janice G. , 'The Micro-Foundations of International Relations Theory: Psychology and Behavioral Economics'. *International Organization* 71/S1 (April 2017), 249–63. doi:10.1017/S0020818316000436 [Crossref], [Google Scholar]
- Stoutland, Page O. and SamanthaPitts-Kiefer , *Nuclear Weapons in the New Cyber Age: Report off the Cyber-Nuclear Weapons Study Group* (Washington, DC: Nuclear Threat Initiative September 2018). [Google Scholar]
- Taleb, Nassim N. , *Fooled by Randomness: The Hidden Role of Chance in Life and the Markets* , 2nd ed. (London, UK: Penguin 2004). [Google Scholar]
- Talmadge, Caitlin , '"Would China Go Nuclear? Assessing the Risk of Chinese Nuclear Escalation in a Conventional War with the United States'. *International Security* 41/4 (Spring 2017), 50–92. doi:10.1162/ISEC_a_00274 [Crossref], [Web of Science ®], [Google Scholar]
- Talmadge, Caitlin , 'Emerging Technology and Intra-war Escalation Risks: Evidence from the Cold War, Implications for Today'. *Journal of Strategic Studies* 42/6 (2019), 864–87. doi:10.1080/01402390.2019.1631811 [Taylor & Francis Online], [Web of Science ®], [Google Scholar]
- Tversky, Amos and DanielKahneman , 'Judgment under Uncertainty: Heuristics and Biases'. *Science* 185/4157 (September 1974), 1124–31. doi:10.1126/science.185.4157.1124 [Crossref], [PubMed], [Web of Science ®], [Google Scholar]
- 'U.S. Office of the Director of National Intelligence', *A Guide to Cyber Attribution* , 14 September 2018. [Google Scholar]
- United States Air Force , 'Artificial Intelligence Annex to DoD AI Strategy', Tech. Rep. (United States Air Force 2019). [Google Scholar]
- Van der Meer, Sico , 'Reducing Nuclear Weapons Risks: A Menu of 11 Policy Options', *Policy Brief* , Clingendael Netherlands Institute of International Relations, June 2018. [Google Scholar]

- Waltz, Kenneth N. , *Theory of International Politics* (Reading, MA: Addison-Wesley 1979). [Google Scholar]
- Wan, Wilfred , *Nuclear Risk Reduction: A Framework for Analysis* (Geneva, Switzerland: UNIDIR 2019). [Crossref], [Google Scholar]
- Weick, Karl E. and Kathleen M.Sutcliffe , *Managing the Unexpected: Sustained Performance in a Complex World* , 3rd ed. (Hoboken, New Jersey: John Wiley and Sons 2015). [Crossref], [Google Scholar]
- Wiberg, Haken , 'Accidental Nuclear War: The Problematique'. in HakanWilberg, Ib DamgaardPetersen, and PaulSmoker (eds.), *Inadvertent Nuclear War* (New York, NY: Pergammon Press 1993), 3–31. [Crossref], [Google Scholar]
- Williams, Andrew P. and Paul D.Scharre , eds., *Autonomous Systems - Issues for Defence Policy Makers* (Norfolk, VA: NATO Allied Command Transformation 2015). [Google Scholar]
- Williams, Heather , 'Asymmetric Arms Control and Strategic Stability: Scenarios for Limiting Hypersonic Glide Vehicles'. *Journal of Strategic Studies* 42/6 (2019), 789–813. doi:10.1080/01402390.2019.1627521 [Taylor & Francis Online], [Web of Science ®], [Google Scholar]
- Wohlstetter, Roberta , *Pearl Harbor: Warning and Decision* (Stanford, CA: Stanford University Press 1962). [Google Scholar]
- Wong, Yuna Huh , et al., *Deterrence in the Age of Thinking Machines* (Santa Monica, CA: RAND Corporation 2020). [Crossref], [Google Scholar]
- Woolf, Amy , *Defense Primer: Command and Control of Nuclear Forces* (Washington, DC: Congressional Research Service December 11 2018). [Google Scholar]
- Yaohui, Yang , 'A Vision of A New Kind of Combat Systems', *PLA Daily* , June 20, 2020, http://www.81.cn/theory/2019-06/20/content_9534942.htm [Google Scholar]
- Yarhi-Milo, Keren , *Knowing the Adversary* (Princeton NY: Princeton University Press 2014). [Google Scholar]
- Yi, Yuan , 'The Development of Military Intelligentization Calls for Related International Rules', *PLA Daily* , October 16, 2019, http://military.workercn.cn/32824/201910/16/191016085645085.shtml [Google Scholar]
- Zero, Global , 'Global Zero Commission on Nuclear Risk Reduction: De-Alerting and Stabilizing the World's Nuclear Force Postures', April 2015. [Google Scholar]