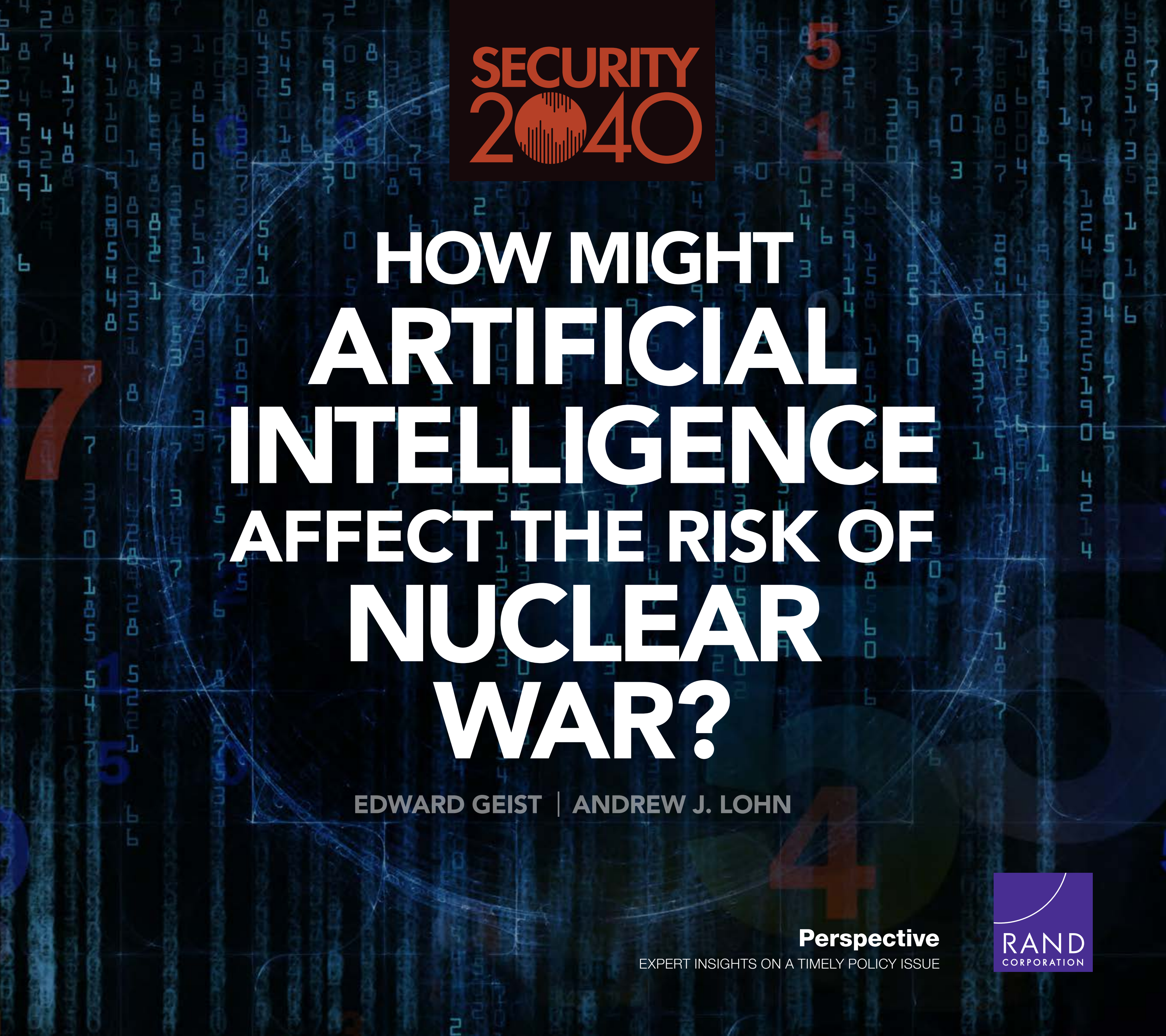


SECURITY 2040

HOW MIGHT ARTIFICIAL INTELLIGENCE AFFECT THE RISK OF NUCLEAR WAR?

EDWARD GEIST | ANDREW J. LOHN

Perspective
EXPERT INSIGHTS ON A TIMELY POLICY ISSUE

RAND CORPORATION

RAND's publications do not necessarily reflect the opinions of its research clients and sponsors. **RAND**® is a registered trademark.

For more information on this publication, visit www.rand.org/t/PE296.

Today's nuclear balance relies on several conditions that may not hold. Progress in computing and data availability are making it possible for machines to accomplish many tasks that once required human effort or were considered altogether impossible. This artificial intelligence (AI) might portend new capabilities that could spur arms races or increase the likelihood of states escalating to nuclear use—either intentionally or accidentally—during a crisis.[1] The RAND Corporation convened a series of workshops that brought together experts in AI and nuclear security to explore ways that AI might be a stabilizing—or destabilizing—force by the year 2040.

The effect of AI on nuclear strategy depends as much or more on adversaries' perceptions of its capabilities as on what it can actually do. For instance, it is extremely technically challenging for a state to develop the ability to locate and target all enemy nuclear-weapon launchers, but such an ability also yields an immense strategic advantage. States therefore covet this capability and might pursue it irrespective of technical difficulties and the potential to alarm rivals and increase the likelihood of conflict. The case could be made on technical grounds that advanced AI would still struggle to overcome obstacles originating from data limitations and information-theoretic arguments, but the tracking and targeting system needs only to be *perceived* as capable to be destabilizing. A capability that is nearly effective might be even more dangerous than one that already works.

The trajectory of AI development, together with that of complementary information technology and other advancements, will have a large effect on nuclear-security issues in the next quarter century. AI technology could continue to evolve rapidly, as it has in recent years, or it could plateau once current techniques mature. Some theorists postulate that machines might develop the ability to improve their own intelligence at some point, resulting in "superintelligences" with abilities that humans could neither comprehend nor control, but there is little consensus about how AI will advance, including the plausibility of superintelligences. Some envision an initial breakthrough followed by setbacks; others suspect that progress will remain incremental.

The two extreme cases have only limited relevance for the future of nuclear warfare. Stalling development (also referred to as *AI winter*) would result in only minor changes from the current nuclear-security environment. With superintelligence, AI would render the world unrecognizable and either save or destroy humanity in the process. The other two cases, in which AI progresses substantially and enables many new capabilities while still remaining fallible and inferior to humans in at least some respects, seem to have more support from the expert community, although experts disagree about the national security implications of such capabilities. Some fall in the category of "Complacents": These tend to believe that producing an AI capable of performing the types of tasks that would destabilize the nuclear balance is sufficiently difficult that it is unlikely to be achieved. "Alarmists" hold the opposite view, that an AI could be capable of certain tasks but should not be included in any aspect of nuclear war. A third group, "Subversionists," focus on an adversary's ability to alter, mislead, divert, or otherwise trick the AI, which could prove either stabilizing or destabilizing.

One example discussed in the workshops was an AI that acts as a decision support system. Without being directly connected to the nuclear launchers, an AI could still provide advice to humans on matters of escalation. It seems reasonable that such a capability, at least for some aspects of the decisionmaking process, could be achieved by 2040 given the progress AI is making in increasingly complex and poorly specified tasks. Alarmists might be concerned that such a capability could be incorporated before it is sufficiently robust or without fully understanding its limitations. If an AI adviser were proven effective, however, it could increase stability by reducing the likelihood of human error and by providing radical transparency, which could reduce the risk of miscalculation. But many experts were concerned by the potential for an adversary to subvert even a very capable AI by hacking, poisoning its training data, or manipulating its inputs.

Maintaining strategic stability in the coming decades will require revisiting the foundations of deterrence theory in a multipolar world. Effective deterrence will require us to contend with the rapidly changing set of capabilities being driven by progress in AI. Key considerations include the impact of the actual capabilities, the perceived potential of those capabilities (whether they exist or not), and the premature use or fallibility of those capabilities, especially as a result of adversarial actions. With care and some forward-thinking, these risks can potentially be identified and mitigated.

> *Without being directly connected to the nuclear launchers, an AI could still provide advice to humans on matters of escalation.*

Hints of Major Changes Ahead for the Nuclear Balance

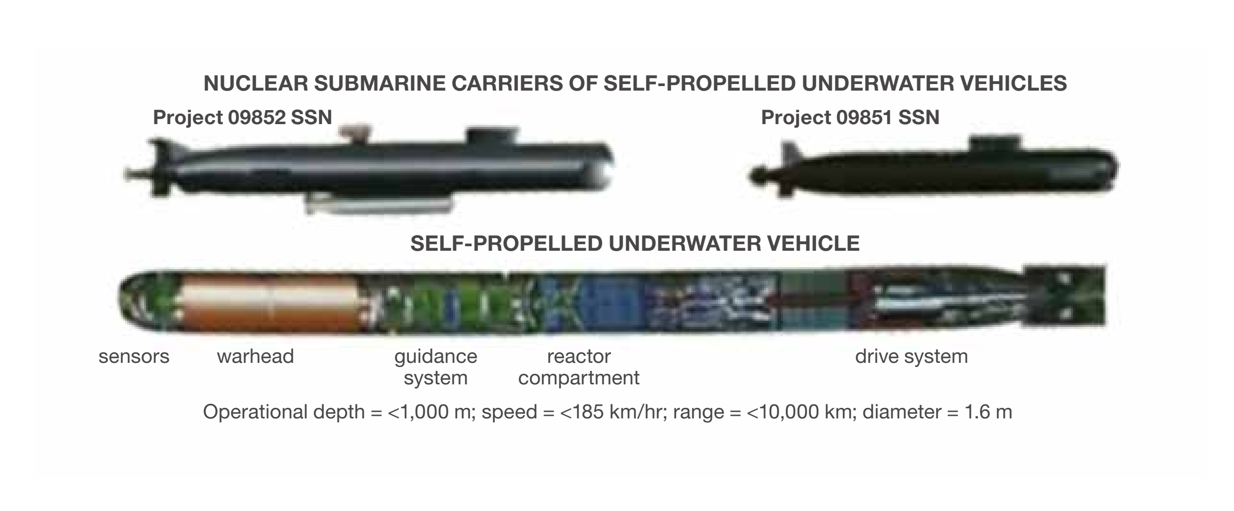
November 2015, Russia revealed that it was developing the ultimate "killer robot": a nuclear powered undersea drone designed to carry an enormous thermonuclear warhead. Russian television revealed the existence of this nightmarish weapon in an "accidental" leak that most Western observers concluded was intentional. Television cameras lingered momentarily on an ostensibly classified briefing slide for President Vladimir Putin describing the "Oceanic Multipurpose System Status-6." Shaped like an enormous torpedo and powered by a compact nuclear reactor (see the figure on page 3), Status-6 would overcome enemy defenses through a combination of speed and range that would enable it to outrun almost anything in the ocean

(Sutyagin, 2016). The drone would be launched from submarines in the Russian arctic, traverse the ocean at perhaps 100 km/hr while autonomously circumventing antisubmarine defenses, and deliver its deadly payload to the U.S. coastline, presumably after a villainous American first-strike attack destroying the Kremlin. The difficulty of communicating underwater would require a degree of autonomous capability on the part of the drone that has become possible only recently as a result of progress in AI.[2]

Status-6 is not just a concrete application of AI; it is a reflection of AI's potential looming impact on nuclear deterrence—the use of retaliatory threats to dissuade an adversary from attacking a state or its allies.[3] The nuclear drone is the latest manifestation of Russian leaders' concerns about the credibility of their retaliatory forces in the face of U.S. counterforce targeting capability and missile defenses. Unable to match these capabilities in kind, contemporary Russia hopes to exploit AI to ensure the credibility of its deterrent. It might succeed by 2040 because the Kremlin continues to explore novel ways of employing AI for military purposes. This effort is in keeping with its decades-old strategy of developing "asymmetric responses" to superior U.S. capabilities. Russia's undersea "doomsday drone" is merely the most extreme example of this phenomenon so far.[4]

Will nuclear deterrence be recognizable in 2040? Status-6 is a stark warning that if technological progress undermines nuclear

Components of Status-6


NUCLEAR SUBMARINE CARRIERS OF SELF-PROPELLED UNDERWATER VEHICLES
Project 09852 SSN
Project 09851 SSN
SELF-PROPELLED UNDERWATER VEHICLE
sensors
warhead
guidance system
reactor compartment
drive system
Operational depth = <1,000 m; speed = <185 km/hr; range = <10,000 km; diameter = 1.6 m

The "Oceanic Multipurpose System Status-6," shaped like an enormous torpedo and powered by a compact nuclear reactor, would overcome enemy defenses through a combination of speed and range that would enable it to outrun almost anything in the ocean.

powers' sense of security, those powers could attempt to salvage their nuclear deterrents by embracing unprecedented new weapon systems and force postures. These unfamiliar strategic arrangements could prove less stable than those that kept an uneasy peace between the United States and the Soviet Union, and instability increases the probability of nuclear war. The extent to which risk increases depends in considerable part on the rate and extent of progress in AI, which could enable new ways of both delivering nuclear weapons and defending against nuclear attack. In May and June of 2017, the RAND Corporation convened three workshops with nuclear-security professionals and AI researchers to discuss the impact of AI on nuclear security. Participants appeared to agree that advanced AI could severely compromise nuclear strategic stability and thereby increase the risk of nuclear war. However, there was not agreement about how and why AI would have this effect, even within respective constituencies.

Methodology and Description of Workshops

To investigate the potential influence of advanced AI on nuclear security in the next quarter century, RAND conducted a series of workshops in May and June of 2017. These workshops brought together a variety of expert groups, including both nuclear-security professionals and AI researchers, as well as participants from government and industry, resulting in a variety of diverse perspectives.

Workshop 1

The first workshop was held at RAND's Santa Monica office on May 1, 2017, and many of the 16 participants were RAND researchers working in nuclear or AI-related fields. The aim of the workshop was to envision strategic environments with which AI might interact, building on the premise that the future geostrategic order is more predictable than the development of AI technology. Discussion was seeded with several specific scenarios in which conflicts between the nuclear powers became more acute. These included:

1. a "resurgent Russia" scenario, in which the New START treaty collapses and Russia achieves a significant advantage in strategic nuclear arms over the United States by the early 2030s
2. a "rising China" scenario, in which China gradually expands its strategic nuclear arsenal and achieves parity with the United States and Russia
3. a "successful limited use" scenario, in which Pakistan successfully uses tactical nuclear weapons to persuade India to withdraw an invading force, breaking the "nuclear taboo"
4. a "regional nuclear war" scenario, in which a North Korean regime undergoing collapse lashes out against South Korea, Japan, and China, resulting in the devastation of the region.

These unfamiliar strategic arrangements could prove less stable than those that kept an uneasy peace between the United States and the Soviet Union.

Workshop participants were asked to flesh out these scenarios with technical details of respective powers' nuclear forces—including number and capabilities of delivery systems and C4ISR (command, control, communications, computers, intelligence, surveillance, and reconnaissance)—with the aim of identifying AI applications that might be of interest to nuclear states. Future combat systems were presumed to be similar to those in development today because military acquisition time lines are slow, but AI progress can be made much faster than defense system acquisitions. Participants seemed to agree that applying advanced AI to these systems would likely be a destabilizing influence in a future standoff. However, participants also postulated that for every destabilizing technology, there is a counter, stabilizing technology. This theme was developed further in the second workshop.

Workshop 2

A group of 19 participants contributed to the second workshop, which was held in San Francisco on May 25, 2017. Seven of these participants described themselves as being in the AI field, five described themselves as being in national security, three said they were in both, and four in neither. The AI-focused contributors included prominent figures from commercial, academic, and nonprofit AI research organizations, as well as AI policy communities, while the national security participants included nuclear-weapon experts from the national laboratories. Participants divided into subgroups to discuss three issues.

The first issue was whether AI could enable states to track and target adversary retaliatory forces and thereby undermine the premise of assured retaliation that forms the basis of much of nuclear strategic theory (discussed in depth in a later section). One group, dominated by nuclear-security experts, concluded that AI could accomplish this but met with disagreement from a second group that included a prominent expert on generative adversarial networks (a technique in which a generator neural network interacts with a classifier neural network to learn to create increasingly realistic fake examples). In this group's view, vulnerability to adversarial manipulation attacks is intrinsic to most of the learning techniques of most machines, so states will be able to employ these approaches as a means to prevent an adversary from tracking its launchers.

The second issue at this workshop addressed was the use of AI in decision support systems to advise decisionmakers on strategic nuclear issues in crisis or conflict. The groups disagreed considerably about the use of AI for these tasks, with some saying they should be kept under strictly human control, while others declared that to be unrealistic. This topic is discussed in detail later on.

Finally, the workshop characterized possible lessons from nuclear arms control for future AI applications. Participants seemed to agree that it is not possible to replicate the kind of legal structures and norms that have been used to forestall nuclear proliferation to head off military AI applications because nuclear technology and AI are too different. In the specific case of AI for nuclear war–related tasks, participants pointed out that controlling AI might be difficult, but other components essential for those applications (i.e., sensor platforms) could be subjected to monitoring and control. Participants discussed whether it might be possible to control AI by controlling data, human talent, or processing resources. Several of the AI researcher participants argued that the current shortage of human talent is temporary and training data would become less important as simulations improve, but that hardware might then become the limiting factor. In their view, the limited number of factories making such components as graphics processing units might make it possible to construct some sort of control regime, but many other participants were skeptical of this.

One subgroup suggested provocatively that a future AI system could essentially be the arms control regime, monitoring compliance and adjudicating violations without human input.

Workshop 3

The third and final workshop took place at RAND's office in Arlington, Virginia, on June 9, 2017. This workshop had 15 participants, including eight who described themselves as focusing on nuclear issues and five as focused on AI, although most of the latter work in the policy space rather than as AI research practitioners. The remaining two participants contributed valuable expertise with respect to acquisition policy. The group included both RAND researchers and representatives from the U.S. Army, the Office of the Secretary of Defense, and the State Department Bureau for Arms Control, Verification, and Compliance. This workshop built on the findings of the previous two, asking the attending policy practitioners how they would address the challenges identified in the previous two workshops.

The first discussion focused on the tracking and targeting problem and asked participants to consider how they would try to thwart an adversary seeking to render strategic forces vulnerable using AI. Participants suggested trying to neutralize this capability by attacking the associated sensors and communications network rather than the AI itself. In the subsequent discussion, participants considered the challenges posed by the generative adversarial techniques emphasized by the AI researchers in the second workshop, although no consensus emerged (possibly due to the unfamiliarity of most of the attendees with technical details of these methods).

The second discussion addressed whether the United States needs to reconsider the trajectory of its nuclear force-modernization programs in light of the possibility that AI might significantly reshape the strategic landscape. Participants noted that the current program has many vulnerabilities but that apparent alternatives are not obviously better even if they probably merit at least some analysis. Domestic and international institutional pressures discourage the United States from departing significantly from the current "triad" of intercontinental ballistic missiles (ICBMs), submarines, and manned bombers.

The third discussion addressed how AI might contribute to nuclear arms control. AI might be used for such tasks as treaty verification, by enabling increased transparency and trust. One subgroup suggested provocatively that a future AI system could essentially *be* the arms control regime, monitoring compliance and adjudicating violations without human input. Finally, the participants considered whether it is possible or desirable to apply arms control to AI itself. Most participants were skeptical of the feasibility and desirability of this goal; many regarded it as either a practical impossibility or something that would require extreme and unacceptable interventions, such as interning AI researchers.

Theoretical and Historical Background for Assessing AI's Potential Influence

During the Cold War, both the United States and the Soviet Union begrudgingly accepted the condition of *mutual assured destruction* (MAD)—the premise that any all-out attack would be met with an apocalyptic retaliatory strike ensuring that both societies would be destroyed. MAD was a condition, rather than a strategy—one that both superpowers hoped to escape if possible (Buchan et al., 2003). Even if mutual vulnerability made a general nuclear exchange less likely, the omnipresent possibility that war might still occur by accident or miscalculation weighed heavily on the minds of superpower leaders. Ronald Reagan, for instance, called on scientists to create a missile defense that would render nuclear weapons "impotent and obsolete," while the Soviet Union developed an elaborate civil defense program (Garthoff, 1987; Geist, 2012). Nor was MAD a sufficient basis for U.S. or Soviet nuclear strategy. While MAD credibly deterred a Soviet preemptive strike on the United States, it also undermined the plausibility of U.S. promises to defend its European allies in the North Atlantic Treaty Organization, even at the risk of nuclear war. If Washington relied solely on MAD, the Soviet Union could exploit its conventional superiority to invade western Europe and the United States would face a stark choice between capitulation or an all-out nuclear war. As a consequence, American strategists and government officials developed the more comprehensive doctrine of assured retaliation—the prospect that any enemy provocation would be met by an appropriate and effective response (Long, 2008). By threatening a retaliation scaled to likely enemy provocations, "assured retaliation" sought to credibly deter minor and all-out attacks. In the later decades of the Cold War, a variant of this approach called the countervailing strategy sought to deter all manner of attacks, including preemptive counterforce attacks, by assuring that any such attacks would fail to accomplish their objectives because of U.S. retaliation (Slocombe, 1981).

Nuclear strategy is about more than just deterrence (see the table opposite). *Deterrence* is the use of retaliatory threats to dissuade an adversary from attacking oneself or one's allies. Deterrence can be categorized into *central deterrence* (deterrence of an attack on one's homeland) and *extended deterrence* (deterrence of an attack on one's strategic partners) (Cimbala, 2002). Nuclear weapons can also be used for *compellence*—coercing the enemy into doing something that it does not want to do (Long, 2008, p. 9). In addition to coercive deterrence and compellence threats, nuclear weapons can be employed for warfighting, the way they were at the end of the Second World War. The practical complexities of nuclear strategy stem from the challenges of *assurance*—making extended deterrence credible. During the Cold War, the United States accumulated its massive stockpile of strategic and tactical nuclear weapons to convince its allies that that it would be willing to retaliate to conventional Soviet attacks in Europe with nuclear responses. As United Kingdom Defence Minister Denis Healey observed, it took "only five per cent credibility of American retaliation to deter the Russians, but ninety-

Categories of Nuclear Strategy Goals

Aspect	Definition
Coercion	
Deterrence	Dissuade adversaries from doing something they want to do
Compellence	Force adversaries to do something they do not wish to do
Assurance	Convince allies that security guarantees are credible
Reassurance	Convince adversaries that they will not be attacked so long as they refrain from provocative behavior

five per cent credibility to reassure the Europeans" (Healey, 1989). The scale of the U.S. nuclear arsenal, however, alarmed Soviet leaders, who believed that the Americans might be attempting to develop a first-strike capability against them. This distrust underscored the need for *reassurance*—convincing adversaries that they will not be attacked as long as they refrain from the behavior that is being deterred (Schelling, 1966).

Strategic stability exists when adversaries lack a significant incentive to engage in provocative behavior.[5] There are several kinds of strategic stability that are distinguished by their varying temporal scales. *First-strike stability* exists when no state can carry out an attack out of the blue against its opponent without significant fear of a devastating retaliation. Such a possibility is best deterred by the threat of overwhelming and automatic retaliation from secure second-strike forces (Cimbala, 2002, p. 66). *Crisis stability*, by contrast, aims to prevent or manage escalation during crises, as occurred in Berlin and Cuba in the early 1960s (Cimbala, 2002, p. 98). In these circumstances, national leaders are under immense pressure not to show weakness by backing down, but the chance of inadvertent escalation increases significantly as states attempt to maneuver the nuclear forces for signaling purposes. In this context, the kind of large automatic retaliation that is ideal for maximizing first-strike stability is a recipe for disaster.

Finally, arms race stability is achieved when there are no exploitable inequalities in adversaries' military capabilities (Cimbala, 2002, p. 110). States avoid these inequalities to manage the risks and costs of long-term competition and to avoid compromising first-strike stability and crisis stability in the future. Nuclear strategy is difficult because these objectives are in tension with each other.

In an extreme case, AI could undermine the condition of MAD and make nuclear war winnable, but it takes much less to undermine strategic stability. AI advancements merely need to cast doubt on the credibility of retaliation at some level of conflict. Major nuclear powers, such as the United States, Russia, and China, have a shared interest in maintaining the credibility of central deterrence, but they seek regional advantages in pursuit of what they regard as their core strategic interests. Areas where credibility is already strained, such as certain extended deterrence guarantees, are particularly vulnerable to destabilization. The increasingly multipolar strategic environment is also encouraging forms of competition that threaten stability. For instance, the United States is interested in developing the capability to track and target a minor nuclear power's mobile missile launchers, but Russia and China fear that the same technology could mature into a threat to their more sophisticated retaliatory forces. In a crisis situation, the employment or availability of AI-enabled intelligence, surveillance, and reconnaissance (ISR) or weapon systems could stoke tensions and increase the chances of inadvertent escalation. Finally, the pursuit of advanced military capabilities is liable to cause arms race instability even if those technologies are nonviable, as in the historical case of missile defense.

The challenge AI poses to strategic stability is not unique to this particular technology, but it is more acute because of rapid technical progress in AI and its many potential intersections with nuclear strategy. Most of the specific applications AI are likely to be used for, such

Strategic stability exists when adversaries lack a significant incentive to engage in provocative behavior.

as analysis of ISR data, controlling autonomous sensor platforms, and automated target recognition (ATR) have been eagerly sought for decades but were beyond the capability of available technology. Even without further breakthroughs, incremental progress using existing AI techniques may make these long-sought goals practical realities in the foreseeable future.

Both Russia and China appear to believe that the United States is attempting to leverage AI to threaten the survivability of their strategic nuclear forces, stoking mutual distrust that could prove catastrophic in a crisis. As Paul Bracken observes, ongoing improvements in technology such as AI threaten to "undermine minimum deterrence strategies" and "blur the line between conventional and nuclear war" (Bracken, 2017).

AI in the Cold War

AI pioneer Marvin Minsky defined AI as "the science of making machines do things that would require intelligence if done by men" (Minsky, 1968, p. v). Since AI research began in the 1950s, the boundaries of the field have shifted as computers have reshaped how humans comprehend "intelligence." AI has also evolved as theoretical paradigms have shifted in and out of vogue. From the 1950s until the 1980s, a "symbolic" paradigm that aimed to replicate high-level human reasoning predominated, only to be supplanted by a "connectionist" paradigm that sought to emulate the biological basis of human cognition using artificial neural networks. In the 20th century, neither paradigm worked particularly well outside laboratory demonstrations. This triggered occasional periods (sometimes characterized as AI winters) during which funding for AI research was scarce. Thanks to decades of progress in computer science, advances in computing and communications hardware and software, and the rise of cloud computing and big data, AI has advanced rapidly in the past few years, most prominently in the field of "deep neural networks" (DNNs), or neural networks with many layers (Goodfellow, Bengio, and Courville, 2016). The increase in the performance of DNNs has been so spectacular that they have become almost synonymous with AI, but in actuality the older paradigms are also continuing to progress and are in widespread commercial and military use. Some impressive recent AI systems, such as Alphabet DeepMind's AlphaGo program, which beat the world Go champion, employ DNNs in combination with such older techniques as searching the tree of possible moves. One thing that has remained constant over AI's 60-year history is its proponents' high hopes. With enough intelligence, might it be possible to conquer such seemingly impossible problems as poverty and illness—or even win a nuclear war?

The intersection between AI and nuclear warfare became a science fiction cliché more than 50 years ago, but their real-world connections are even older. The earliest AI researchers were deeply involved in national security work and secured government support by suggesting that their theoretical studies would soon translate into practical military applications. Claude Shannon asserted in his foundational 1950 article "Programming a Computer for Playing Chess" that making computers play that venerable game would impart theoretical insights that would make "machines for making strategic decisions in simplified military operations" possible "in the near-term future" (Shannon, 1950, p. 256). In the mid-1950s, researchers created the earliest working AI programs with support from the U.S. Air Force (Simon and Newell, 1958; Newell, Shaw, and Simon, 1959). Potential applications of such machines soon began appearing in the writings of strategic theorists. In the late 1950s, Herman Kahn postulated the notion of "doomsday machines" that would employ computers programmed to recognize unacceptable enemy provocations and retaliate (Kahn, 1960, pp. 145–154). While Kahn intended these as thought experiments illustrating how not to conduct nuclear strategy, science fiction authors latched onto the idea of intelligent computers controlling nuclear weapons, inspiring numerous novels and such films as *Colossus* (1970), *WarGames* (1983), and *Terminator* (1984).

While fictional thrillers spin tales of nuclear armed computers run amok, real-world attempts to apply AI to nuclear strategic problems tended to be much more mundane. Neither U.S. nor Soviet officials were inclined to entrust launch decisions to computers, both because they jealously reserved this prerogative for themselves and because automating retaliation was not a logical response to difficult strategic problems, such as compellence or crisis stability. The sole notable exception came from the Soviet Union at the end of the Cold War. Perceiving that the United States aspired to a first-strike capability and anxious that they might be the objects of a decapitation strike, Soviet leaders sought measures to ensure that capitalist aggressors would never go unpunished.[6] The Soviet Union reportedly considered developing a system that would have automatically launched surviving ICBMs at the United States following a first strike if it could not contact the Soviet political leadership. It seems that the fully automated version, nicknamed the "Dead Hand," was rejected in favor of a version, dubbed "Perimetr," that would automatically delegate launch authority to field commanders but would always require a human in the loop (Hoffman, 2009). According to Russian media accounts, the Perimetr system still exists and uses some kind of AI.[7] The United States, meanwhile, explored the possibility that AI could be used to bolster its counterforce capability. One late 1980s research project, the Survivable Adaptive Planning Experiment (SAPE), sought to use the AI technology of the time to enable the United States to target the Soviet Union's mobile ICBM launchers. The SAPE would not control nuclear weapons directly; rather, it would employ expert systems to translate reconnaissance data into nuclear targeting plans that would then be carried out by manned B-2 bombers. The SAPE was just one part of an envisioned suite of systems and capabilities that, if actualized, would have severely challenged the survivability of the Soviet Union's nuclear arsenal (Roland and Shiman, 2002, p. 305; Long and Green, 2012).

AI and the Emerging Geopolitical Order

Although 20th-century AI struggled to actualize these applications, more-recent advances in computing could release their potential. Such contemporary techniques as deep learning are dramatically advancing machine vision and other signal processing applications, which can enhance autonomy and sensor fusion. Autonomy and sensor fusion may be of paramount strategic relevance because they could greatly improve ISR, ATR, and terminal guidance capabilities. All of these might severely erode the means by which nuclear powers assure the survivability of their nuclear forces. Because increased weapon accuracy has long since undermined the survivability of silo-based ICBMs, the United States, Russia, and China put nuclear weapons on submarines and mobile ICBMs that were deemed more likely to survive a first strike. Technologies that make it more likely that survivable forces (such as submarine and mobile missiles) could be targeted and destroyed make it more plausible that one country might threaten a first strike. This undermines strategic stability, because even if the state possessing these capabilities has no intention of actually using them, the adversary cannot be sure of that. Thus, the capabilities can still be used to pressure potential adversaries and perhaps extract concessions during a crisis. Such a capability does not have to be exploited during a crisis to be politically useful. As Alfred T. Mahan observed, "force is never more operative than when it is known to exist but is not brandished" (Mahan, 1912, p. 105). As long as adversaries fear that the capability may exist, they can be cowed into submission without explicit confrontation—the more powerful state can in effect preemptively "win" the crisis. As a consequence, counterforce targeting capability is an enticing prospect for many despite its potential to compromise strategic stability.

AI technologies could help enable new breakthroughs in tracking and targeting and in antisubmarine warfare or make it easier for high-precision conventional munitions to destroy hardened ICBM silos (Holmes, 2016). Such capabilities would be especially destabilizing because decisionmakers could threaten to employ conventional weapons much more plausibly than any kind of nuclear attack. A conventional threat would place the adversary under enormous pressure during a crisis, which could force it to capitulate—but could also spiral into nuclear war. Such escalation could happen because the adversary felt the need to use its weapons before being disarmed, in retaliation for an unsuccessful disarming strike, or simply because the crisis triggered accidental use.

Potential U.S. adversaries, such as Russia, take seriously the possibility that the United States might leverage its advantage in such technologies as AI to radically improve its counterforce capabilities. For the past several years, Russian military analysts have been engaged in a vociferous debate in the military press about the extent of their country's strategic vulnerabilities.[8] Their tendency to assume that current and future U.S. capabilities pose a dire threat to Russia's security stokes these anxieties.

A major challenge of nuclear strategy is that adversaries may interpret one nation's secure retaliatory forces as a first-strike threat or a doomsday machine and react accordingly. For instance, the Russians probably conceived of Status-6 as a last-ditch second-strike option exploiting AI to autonomously circumvent U.S. defenses, but Western observers interpreted it as a Strangelovean "cobalt bomb." AI progress is also contributing to Russia's doubling down on older types of systems with undesirable strategic properties. For example, with its RS-28 "Sarmat" missile, Russia is reinvesting in large, silo-based ICBMs with multiple independently targetable reentry vehicle (MIRV) warheads, a category of weapon it once planned to abandon under the now-defunct Strategic Arms Reduction Talks II treaty. Western strategic theory generally considers large MIRVed ICBMs to be destabilizing because they are ideal for preemptive strikes and are vulnerable to preemption.

At the dawn of the millennium, Moscow believed that it could ensure the survivability of its forces by emphasizing mobile ICBMs and scrapping large silo-based missiles inherited from the Soviet Union. However, Russian leaders' anxieties about potential U.S. threats to the survivability of the mobile ICBMs seem to have changed this calculus and led them to try to ensure retaliation by launching during a U.S. attack instead of riding it out.

A major challenge of nuclear strategy is that adversaries may interpret one nation's secure retaliatory forces as a first-strike threat or a doomsday machine and react accordingly.

The increasingly multipolar nuclear environment also aggravates the potential strategic impact of AI.

This is tantamount to the adoption of a launch-under-attack posture that could place great pressure on Russian leaders to launch first in a crisis, increasing the chances of accidental escalation. The Russians recognize that the Sarmat silo would be unlikely to survive a preemptive attack on its own, so its survivability hinges on an associated active defense system, code-named "Mozyr'," that would attempt to force enemy warheads to detonate at a slight distance from a silo, allowing it to survive the nuclear explosions.

The increasingly multipolar nuclear environment also aggravates the potential strategic impact of AI. While six states had the bomb during the Cold War, five of them considered the Soviet Union their primary enemy, making the strategic order essentially bipolar. This bipolarity encouraged both crisis and arms race stability. Today, there are nine nuclear-weapon states and multiple strategic rivalries that indirectly affect one another. The United States worries about Russia and China; Russia plans for confrontations with both the United States and China; China regards the United States, Russia, and India as potential adversaries; India is embroiled in strategic competition with China and Pakistan; and North Korea is a headache for almost everyone.

Much work remains in developing theories of strategic stability applicable to these complex multipolar conflicts. As analysts develop approaches to this challenging problem, they need to consider the various means by which AI could raise or lower the risk of intentional or accidental thermonuclear war. Even with our current imperfect understanding, it is possible to start considering the impact of some emerging capabilities and their interactions.

Commonly Held Expert Opinions on Possible AI Futures

As discussed earlier, several perspectives dominated discussions at the workshops.

Anticipating Progress in AI

There are four main schools of thought regarding progress in AI. It is difficult to provide rigorous justification in favor of any one school over any other. Nonetheless, their proponents often argue emphatically on their behalf. For the most part, the experts who attended in our workshops were familiar with the various schools of thought and were able to make arguments from all perspectives.

For each of the possible future states of AI, the common views from each of the categories of expert opinion with respect to nuclear security are summarized in the table on page 13. These views and the arguments supporting or refuting them are discussed further in the following sections.

Superintelligence

Superintelligence is anticipated by some to be an inevitable state where machines come to hopelessly outmatch humans intellectually. Such theorists as Oxford philosopher Nick Bostrom (2014) argue that once a superintelligence exists, two outcomes are possible: The superintelligence is benevolent and solves all humanity's problems, or the superintelligence destroys humanity, either maliciously or incidentally. Bostrom believes that recursively self-improving AIs might evolve to superhuman intelligence extremely quickly while committing few or no mistakes. This "intelligence explosion" might take hours or minutes.

In this case, the role of nuclear security is made trivial: if benevolent, superintelligence would save humanity from nuclear war; if malevolent, nuclear strikes would be just one of many possible methods for extinction.

Superintelligence does not seem to be viewed as imminent or inevitable by the majority of experts in AI, but many supporters believe it merits attention because of the extreme nature of its costs and benefits, even if the likelihood of its occurrence is low.

Limited Breakout

Short of creation of a true superintelligence, large and discontinuous jumps in intelligence might also be possible, leading to greatly increased intellects that would still be subhuman in at least some respects. This could happen, for example, if a recursively reprogrammable software system were to rapidly increase intelligence until reaching the peak of what its hardware is capable of and being unable to advance further.

In this case, depending on the exact capabilities that emerged relative to humans, the AI would most likely be used to exploit its comparative advantages and humans would be used to maximize theirs. The AI remains fallible, as do humans, and the range and impact of possible outcomes rely heavily on the nature of that fallibility.

Continuous Incremental Progress

A third possibility is that a state similar to the one outlined above is reached not through discontinuous jumps in the progress of AI but as a result of continuous incremental advancement. This progress

Alternative AI Futures from Expert Opinions

	Possible Future States of AI		
Categories of Expert Opinion	AI Winter	Limited Breakout or Continued Incremental Progress	Superintelligence
Complacents	Likely	Data insufficient and problems too complex for even advanced AI	Unlikely to exist but would probably be safer than humans
Alarmists	Unlikely	Algorithms that barely work would alarm adversaries and could fail if used	Inevitable eventually and likely to destroy humanity either intentionally or inadvertently
Subversionists	Neutral	AI could be made to fail catastrophically, or ability to cause AI to fail could provide stabilizing assurances	Superintelligence resistant to both subversion and human control

would rely on increases in computing speeds, hardware architectures, algorithmic development, data availability, and decreasing costs. This is arguably the most plausible interpretation of recent trends in AI, where the increases in AI capability are mundane when viewed at any one moment in time but compelling when viewed over a few years—when projected over more than two decades to the year 2040, progress could be astonishing in the same way that the internet of today would be barely recognizable to nontechnologists in the mid-1990s.

While perhaps distinct from a philosophical or preventive policy perspective, continuous incremental progress over more than two decades is largely indistinguishable from the limited breakout school in terms of outcomes and impact on the world of nuclear security. Just as for limited breakout, there would likely be aspects of superiority for both humans and machines while both remain fallible, and that fallibility would drive the risks.

AI Plateau

A final perspective expressed by a few workshop participants posited that AI progress might plateau once current techniques reach technological maturity. Such an outcome would be distinct from historical AI winters, during which AI research continued to advance even though funding and popular interest dropped markedly. For instance, computer hardware might stagnate and deprive AI of the computational resources required to reach its theoretical potential.

Currently active AI researchers are frequently skeptical of this line of thought, although they acknowledge it is conceivable. Current levels of direct investment in AI development around the world are unprecedented and may not be sustainable, but the expressed level of commitment from private firms and such governments as China suggest that funding will remain robust for the foreseeable future. With such high levels of investment, projections of progress can become self-fulfilling, much as Moore's Law drove semiconductor development for decades (Mack, 2011). Moreover, an "AI plateau" might occur after considerable progress from present-day capabilities, creating many of the same challenges for nuclear security as the previous two scenarios.

Anticipated Impacts on Nuclear Security

The focus of this Perspective is mainly on the limited breakout and continuous incremental progress cases because the other two are of limited relevance from a nuclear perspective. The limited breakout and continuous incremental progress projections are characterized by AI that far exceeds human capacity in increasingly complex and data-limited tasks. Experts disagree about what such capabilities imply for nuclear security.

Complacents

One common view is that AI will not profoundly change the status quo aside from improving efficiency and transparency. Subscribers to this "complacent" view tend to be focused more on technological issues than on strategy or policy. It should be noted that, although including some of the world's most capable AI engineers, participants were probably more interested in AI safety than is typical in their fields, so this view may have been underrepresented. Complacents would be more likely to believe that the complexity of nuclear war is too challenging for AI to contribute significantly and therefore AI's impact on the existing balance would be negligible. Complacents would assert, for example, that challenges in data collection and in distinguishing between real systems or actions and decoys would be impossible for an AI to overcome, even by 2040. They would also

Moreover, an "AI plateau" might occur after considerable progress from present-day capabilities, creating many of the same challenges for nuclear security as the previous two scenarios.

be more likely to view the problem of identifying and interpreting inputs for decisions regarding nuclear escalation as being sufficiently broad to be AI-complete. That is to say, any computer that could outperform humans at this task would necessarily have made the jump to being able to outperform humans generally.

Alarmists

At the opposite extreme are Alarmists, who tend to believe that AI will render existing systems vulnerable or will upset the present strategic balance enough to be of grave concern. This camp includes those who would never entrust any aspect of nuclear decisionmaking to an algorithm. Some of the participants had personal experience with historical attempts to create algorithms to achieve related objectives. In some cases, the ineptitude of those algorithms and their inability to consider the emotional and ethical aspects of a decision had made the participants uncomfortable with the intersection of AI and nuclear issues. Alarmists also argue that an AI needs only to be *perceived* as highly effective to be destabilizing—for example, in the tracking and targeting of adversary launchers. Threatened with potential loss of its second-strike capability, an adversary would be pressured into a preemptive first strike or into expanding its arsenal, both undesirable outcomes.

Subversionists

A third perspective that can lead to positions that fall between the Complacent and Alarmist camps is rooted in concerns over AI's susceptibility to adversarial actions. This stems from theoretical considerations and demonstrations that such adversarial attacks are likely to be highly effective. This view does not always lead to the same conclusions as the cases in which AI shows limited progress or is perceived to be effective but is not, although there is overlap with both.

In numerous convincing demonstrations, small amounts of adversarial effort toward subverting machine learning algorithms have shown outsized effect. Some researchers argue that this is a pervasive trait of machine learning and that they expect that it will persist for years to come. Where an effective AI for tracking and targeting might be destabilizing and lead to proliferation or worse, an adversary may regain trust in the survivability of its second-strike forces if it is confident in its ability to forestall detection using these adversarial methods, thereby reestablishing strategic stability. On the other hand, an actor may believe that it can subvert an AI's ability to identify a preemptive first strike, making such a strike a viable option and therefore destabilizing.

Illustrative Case: Tracking Mobile Missile Launchers

AI may be strategically destabilizing not because it works too well but because it works just well enough to feed uncertainty. To illustrate this point, in this section we describe the results of earlier RAND research on the problem of targeting mobile missile launchers.

Impact on Secure Second Strike

Most nuclear powers favor mobile missile launchers because they are difficult to track and target and therefore are considered survivable. These missiles move regularly via road or rail, and unless the enemy can keep apprised of their locations at all times, the only way to threaten them (other than a first strike destroying the weapons before they are deployed to the field) is by attempting to target their sizable patrol areas with nuclear weapons. Even such bombardment strategies are really practical only if the possible locations of the missiles can be narrowed down at least somewhat. Cold War–era schemes to target Soviet mobile ICBM launchers combined bombardment strategies with intelligence about patterns in the way the Soviet Union moved its missiles.

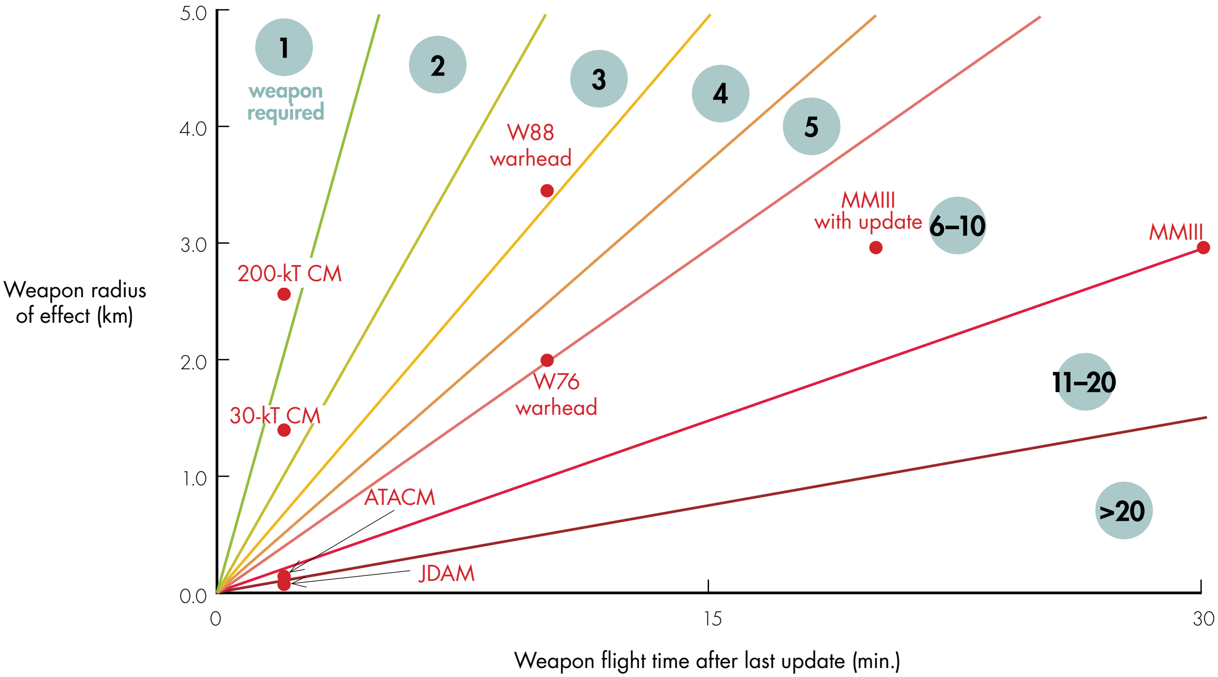
AI could make critical contributions to ISR and analysis systems, upending these assumptions and making mobile missile launchers vulnerable to preemption. This possibility seriously alarms Russian and Chinese defense planners because those states rely heavily on mobile ICBMs for deterrence. Even if AI only modestly improves the ability to integrate data about the disposition of enemy missiles, it might substantially undermine a state's sense of security and undermine crisis stability. At present, the requisite capabilities for ATR, sensor integration, and signal processing remain forbiddingly difficult. But it appears plausible that these challenges could fall in the uncomfortable median between working well enough to render these weapons entirely obsolete and utterly lacking credibility.

A Difficult Technical Challenge

RAND has developed a model for tracking and targeting adversary forces that incorporates limitations in sensing, image processing, and the weapon velocities and kill radii.[9] While some of these limitations could be overcome by advances in AI over relatively short time lines, others are less likely by the year 2040. For example, even with perfect knowledge of the target location, mobile targets can move between the time a weapon is launched and the time it arrives. Weapons for targeting mobile systems might be able to fly faster and adjust course better, but weapons would still need extremely sophisticated terminal guidance capabilities to substantially reduce the amount of ordnance required. As a result, even with advances in image processing and target recognition, many large weapons would be needed, or smaller ones would need to be launched from close range. The figure on page 17 shows the number of warheads of various types that would be required to destroy a mobile target with a weapon radius of effect between 0 and 5 kilometers. Despite their huge "kill radius" measuring kilometers in diameter, multiple thermonuclear warheads delivered by ballistic missiles would be required to have a high assurance of destroying a missile launcher. For instance, three 475-kT W88 warheads delivered by Trident II missiles with a ten-minute flight time would be required to cover one target, while five 100-kT W76 warheads would be necessary to cover it. The analysis finds, however, that accurate cruise missiles (CMs) launched from a position close to the targets (30-kT CM and 200-kT CM in the figure) could cover the mobile missile launchers with only one or two

> *Fired from very close distances . . . , even conventional munitions could become viable options, thereby significantly increasing the credibility of preemptive counterforce strikes.*

Minimum Number of Weapons Required to Cover Target



This figure shows the number of warheads of various types that would be required to destroy a mobile target with a weapon radius of effect between 0 and 5 kilometers. Despite their huge "kill radius" measuring kilometers in diameter, multiple thermonuclear warheads delivered by ballistic missiles would be required to have a high assurance of destroying a missile launcher. ATACM = Army Tactical Missile System; JDAM = Joint Direct Attack Munition; kT = kiloton; MMIII = Minuteman III.

warheads. Fired from very close distances (i.e., flight times of a few minutes), even conventional munitions could become viable options, thereby significantly increasing the credibility of preemptive counterforce strikes.

These finding suggest that AI in conjunction with mobile, possibly autonomous, sensor platforms could enable the development of strategically destabilizing threats to the survivability of mobile ICBM launchers but also offer some hope that arms control could help forestall threats. To pose a credible threat to the mobile ICBM launchers, the attacking forces need to be based very close to them, even if the AI system that is tracking and targeting the launchers is relatively sophisticated. Even under such conditions, "windows of vulnerability" during which the attack might be carried out would last a period of only minutes, pressuring the would-be attacker to seize the opportunity if it arose. States worried about a disarming strike would be extremely alarmed by the appearance of such forces around their periphery, perhaps leading them to believe that they were in a "use it or lose it" situation. Therefore, such moves could create a vicious cycle of distrust and actually incite a conflict that neither side intended. Such undesirable outcomes might be avoided, however, by verifiable agreements not to base or deploy weapons that might be used for such a disarming strike within a certain distance of the mobile missile launchers.

Illustrative Case—AI as a Trusted Adviser

In addition to potentially reducing confidence in second-strike forces, AI could inadvertently compromise a state's ability to navigate road-to-war, escalation, and launch decisions. Autonomous control is unlikely to be implemented directly in any of the domestic launchers or command-and-control platforms, but doing so is not necessary for AI to exert influence. This already happens to the degree that computer programs, simulations, or data analysis procedures are used to inform human decisions. AI is expected to become more widely used in aids to decisionmaking (commonly termed *decision support systems*).

Probable Roles for AI in Decisionmaking

AI is making rapid progress, exhibiting superhuman performance at increasingly complex tasks. Alphabet DeepMind's AlphaGo defeating the world champion at Go astonished even AI and strategy experts (Etherington, 2017). To be sure, the decisionmaking in Go is far simpler to address than in nuclear war; the moves are sequential and clearly defined. But DeepMind's developers have been working toward AI that can play the computer game Starcraft (Woyke and Kim, 2017), which mirrors a military engagement complete with logistics, infrastructure, and a range of moves and strategies that are difficult to specify. Starcraft, too, is far simpler than nuclear war, but by the year 2040, it does not seem unreasonable to expect that an AI system might be able to play aspects or stages of military wargames or exercises at superhuman levels. Once that capability has been demonstrated, it is likely that humans making command decisions will treat the AI system's suggestions as on par with or better than those of human advisers. This potentially unjustified trust presents new risks that must be considered.

> *AI is making rapid progress, exhibiting superhuman performance at increasingly complex tasks.*

Some workshops participants were convinced that humans would be unwilling to let the computer influence decisions about nuclear war, while others could easily envision growing comfortable with the idea. Anecdotally, the difference in perspective was generational, suggesting that those who will have inherited the reins by 2040 will be more comfortable with abdicating some degree of control, especially as AI continues to prove itself in increasingly complex and day-to-day tasks over the coming decades. It is already common for Americans to rely on AI to make routing decisions when driving, facilitate scheduling tasks, and respond to simple e-mails. Perilously, these successes may build confidence that is unwarranted considering the chasm between routine decisions and nuclear war.

Limits to Effectiveness Because of Adversarial Actions

There are two main challenges in developing AI for tasks relating to nuclear war. First, nuclear weapons have not been used since the U.S. bombings of Japan in 1945 triggered unconditional surrender, and there has never been a nuclear exchange. Therefore, there is a complete lack of real training data. However, simulations, wargames, and exercises might help alleviate that problem, and it should not be forgotten that the same lack of real data limits human learning and decisionmaking as well.

The second distinguishing trait is that all parties involved in such common pursuits as navigation or scheduling tend to have the same incentive to complete the task successfully, whereas nuclear war is inherently adversarial. There is a range of different approaches to subverting AI systems, and it appears that subversion is likely to be an effective option for a long time to come. We will briefly discuss hacking, training data attacks, and input manipulation as illustrations of the types of concerns that exist.

Hacking

Hacking is not specific to AI, but as long as AI involves computers, it must be considered vulnerable to hacking. The intelligence itself can be hacked, as will be described later, but data might also be altered at the inputs, outputs, or en route from the output to the display, for example. Of course, any AI that played a role in the nuclear enterprise would be carefully protected, but it would also be a high-value target.

Training Data Attacks

Another way to subvert AI is to tamper with the training data. That can be achieved in several ways: insiders replacing data, hacking to switch out data, including erroneous samples in openly available data, or an adversary carefully selecting its behaviors in ways that set false precedent.

A range of studies have started to outline strategies for, and effects of, poisoning training data for various machine learning algorithms (Anderson et al., 2017; Biggio, Nelson, and Laskov, 2012; Kearns and Li, 1992), but much work is left to be done and many more discoveries should be anticipated. Much of this work is being led by the antivirus community, which is among the few other application spaces that are adversarial by nature—in recent years, this community has turned to machine learning as opposed to more-traditional, signature-based methods. Some have sought ways to ensure that machine learning remains effective despite data manipulation attacks, but those efforts remain nascent (Kegelmeyer et al., 2015). Data tampering is expected to be a threat for a long time to come.

Input Manipulation

A third opportunity to subvert AI comes after it is fully trained. Manipulating the inputs in subtle ways can lead even a high-performing AI system to come to any of its possible conclusions that the attacker prefers. This has been demonstrated for image recognition, where changes so small that they are undetectable to humans have been made to an image and caused the AI to classify the altered image as a category of the attacker's choosing (Karpathy, 2015). This may be more difficult in nuclear matters, where a human may not have precise knowledge of all the inputs or possible classifications. In the image-recognition case, the adversary range of inputs was simple, restricted to pixels. For other tasks, adversaries may need to mobilize forces in a specific pattern or release statements with a specific message in a specific sequence, but it is still possible—at least in principle—to “trick” fully trained AI systems. Importantly, input manipulation attacks do not require the adversary to have access to the trained system, so even a well-protected AI can still be vulnerable (Papernot, McDaniel, and Goodfellow, 2016). More research will be needed to understand the extent of vulnerabilities and to understand which parts of the inputs, outputs, and data would need to be kept secure.

Impact of Limited Effectiveness on Nuclear Security

In the previous case of tracking and targeting, AI threats relate to undermining strategic stability because of the adversary’s exaggerated faith in its effectiveness, but the opposite could also be the case. In the case of decision support systems, it is more pernicious for the force employing the AI to believe that it is effective when it is not. It is also possible that an adversary could become convinced that it is able to subvert an AI and avoid retaliation, leading it to pursue paths that would otherwise be escalatory in nature, up to and including preemptive first strike. For example, the adversary might be convinced that it has discovered a pattern of launches and trajectories that would lead the AI to view the data and conclude that such a pattern is safe even as missiles are en route to targets.

AI presents an array of new vulnerabilities that are difficult to detect in real time. Yet it will almost certainly—eventually or gradually—be given more prominence in road-to-war, escalation, and even launch decisions. Any system with those responsibilities should have to go through rigorous testing that would include adversarial approaches. The simulation of adversaries in testing is fully effective only if the tester can envision the full range of attacks an adversary might create. This impossibly tall order is nonetheless faced for all military systems that are deployed.

Progress in AI appears inexorable, with firms and governments rushing to employ it for an ever-widening range of applications, including both offensive and defensive uses.

Some Possible Stability-Enhancing Effects of AI

Given the many decades that have passed without nuclear attack, it is easy to take strategic stability for granted. While the previous sections have outlined ways in which AI progress could undermine strategic stability, this need not be the case. Progress in AI appears inexorable, with firms and governments rushing to employ it for an ever-widening range of applications, including both offensive and defensive uses. The effects of AI on these strategic applications will become apparent only with time. AI has the potential to exacerbate the tensions among different aspects of nuclear strategy, but it might, under favorable circumstances, alleviate these tensions and enhance strategic stability instead. Despite their mutual distrust, nuclear states may be motivated by self-interest to coordinate toward this end.

The Periods Beyond High Fallibility

Workshop participants agreed that the riskiest periods will occur immediately after AI enables a new capability, such as tracking and targeting or decision support about escalation. During this break-in period, errors and misunderstandings are relatively likely. With time and increased technological progress, those risks would be expected to diminish. If the main enabling capabilities are developed during peacetime, then it may be reasonable to expect progress to continue beyond the point at which they could be initially fielded, allowing time for them to increase in reliability or for their limitations to become well understood. Eventually, the AI system would develop capabilities that, while fallible, would be less error-prone than their human alternatives and therefore be stabilizing in the long term.

Potential Cooperation for Strategic Stability

One of the factors that perpetuates the risk of nuclear war is the contradiction between the first-strike stability requirement of assured retaliation, which encourages governments to adopt "launch under attack" postures, and the possibility of an accident or malfunction. For instance, a 1983 malfunction of the Soviet Union's early warning systems led to the "detection" of a nonexistent U.S. attack (Hoffman, 2009, pp. 6–11). Particularly during a crisis situation, such an incident might lead officials to order a retaliatory strike in response to a phantom assault. AI could help alleviate this contradiction by enabling the creation of more-reliable early warning systems. Greater first-strike stability should, in turn, help reduce the danger of accidental escalation in crises. Even so, this kind of confidence could be a mixed blessing. An aggressor state believing in its ability to predict escalation might feel emboldened to risk provocative actions from which uncertainty might otherwise dissuade it.

Improved accuracy in intelligence collection and analysis could also reinforce strategic stability by making deterrence, assurance, and reassurance more credible. If potential adversaries had less opportunity to prepare an attack in secret, threats to use force against oneself or one's allies would be less plausible. If strategic partners had access to more-comprehensive intelligence and analysis, they could be assured more easily. With smaller forces needed for assurance, a nuclear power such as the United States could reduce the size of its nuclear arsenal, which would enhance reassurance of the enemy. This process could develop into a virtuous cycle, ultimately greatly reducing the risk of war. Unfortunately, this outcome would require fortuitous conditions to materialize, irrespective of the state of AI technology. First, all actors would require equivalent access to intelligence and analysis capabilities. The weaker state in an emerging intelligence asymmetry would probably consider itself unacceptably vulnerable and deepen its suspicions of the adversary. Furthermore, the intentions of rival states would need to be genuinely benign. Finally, officials' confidence in the intelligence collection and analysis system (including non-AI components) needs to be well justified.

To actualize the potential of AI to bolster strategic stability, states need to begin coordinating as the technology matures to avoid these pitfalls. These discussions should include diplomatic and military officials, as well as technology experts.

Radical Transparency

In one highly optimistic possibility, an AI algorithm being used to provide support for decisions about escalation could be shared with the adversary. Such radical transparency would come with many risks. The adversary might then be able to pursue undesirable actions up to the very edge of the escalation threshold. It might also try to subvert the AI. At the same time, any AI that would be used as an aid in such decisionmaking should be required to undergo extensive testing, including of an adversarial nature. It is good practice in any case to attempt to design AI in such a way that it would remain secure even if an enemy were to get the algorithm; it is dangerous to presume that an adversary would be unable to obtain it (Kerckhoff, 1883). If the AI computer system must meet that high standard of robustness prior to fielding, disseminating it widely might alleviate fears and make miscalculations nearly impossible.

Conclusions

Overall workshop participants agreed that AI has significant potential to upset the foundations of nuclear stability and undermine deterrence by the year 2040, especially in the increasingly multipolar strategic environment. Dismissing the Hollywood nightmare of malevolent AIs trying to destroy humanity with nuclear weapons, experts were instead concerned with more-mundane issues arising from improving capabilities. AI applications discussed included the ability to track and target adversary launchers for counterforce targeting and the incorporation of AI into decision support systems informing choices about the use of nuclear weapons.

Some experts fear that an increased reliance on AI could lead to new types of catastrophic mistakes. There may be pressure to use it before it is technologically mature; it may be susceptible to adversarial subversion; or adversaries may believe that the AI is more capable than it is, leading them to make catastrophic mistakes.

On the other hand, if the nuclear powers manage to establish a form of strategic stability compatible with the emerging capabilities that AI might provide, the machines could reduce distrust and alleviate international tensions, thereby decreasing the risk of nuclear war.

At present, we cannot predict which—if any—of these scenarios will come to pass, but we need to begin considering the potential impact of AI on nuclear security before these challenges become acute. Maintaining strategic stability in the coming decades may prove extremely difficult, and all nuclear powers will have to participate in the cultivation of institutions to help limit nuclear risk. This goal will demand a fortuitous combination of technological, military, and diplomatic measures that will require rival states to cooperate. We hope that this Perspective will begin that discussion and open a path toward pragmatism and realism on these controversial and often polarizing topics.

Notes

[1] In this Perspective, we employ the term *artificial intelligence* in an informal sense that includes many computer science achievements in research programs broadly associated with AI, even though these accomplishments ultimately had little to do with emulating human intelligence per se. Such programs resulted in pattern-recognition algorithms, new programming languages, natural-language processing, and a host of other functions that were referred to as AI in previous decades but have long since entered the mainstream of computing.

[2] Russian press accounts attest that the Status-6 employs *iskusstvennyi intellekt* (AI) to achieve its autonomous capabilities. For instance, see Tuchkov (2016) and "Rossiiskii proekt 'Status-6' meniaet sootnosheniia iadernykh sil v mire" (2016). The latter article asserts that Status-6, "being equipped with artificial intelligence," could circumvent antisubmarine warfare measures by following "otherwise unreachable routes" to attack the enemy "where he least expects it."

[3] In keeping with common parlance, we are colloquially including both future and recent advances within the definition of AI, but not those with a long history of application, even if the tasks required humans at some point in the past.

[4] During the final years of the Cold War, the Soviet Union elected to counter prospective U.S. missile defenses by developing missile technologies designed to defeat them. For a Russian account of the Soviet "asymmetric response" to Ronald Reagan's Strategic Defense Initiative, see Oznobishev, Potapov, and Skokov (2008). Vladimir Putin continues to echo this language—for instance, in his 2012 declaration that "Russia's military-technical response to American global antimissile defense and its component in Europe will be effective and asymmetric" (Putin, 2012).

[5] The 2010 Nuclear Posture Review states that the goal of U.S. nuclear strategy is to "strengthen deterrence of regional adversaries," such as North Korea, while "reinforcing strategic stability" with Russia and China. The report does not provide a concise definition of "strategic stability," however (U.S. Department of Defense, 2010).

[6] The United States and Soviet Union each sought to develop a preemptive strike capability that could be used to disarm the other if an attack appeared to be imminent, but this should be distinguished from a first-strike capability designed to mount a "bolt from the blue" attack. The practical difficulty of distinguishing strategic forces intended for a preemptive attack from those intended for a first strike led officials in both superpowers to fear that the other side might be preparing to start a nuclear war.

[7] The assertion that Perimetr employs some kind of iskusstvennyi intellekt (AI) has appeared repeatedly in Russian state media. For example, see Timoshenko (2015) and Valagin (2014).

[8] For instance, see Akhmerov, Akhmerov, and Valeev (2016).

[9] Unpublished RAND research by Brien Alkire and Jim Powers.

References

Akhmerov, D. E., E. N. Akhmerov, and M. G. Valeev, "'Uiazvimost' kontseptsii neiadernogo razoruzheniia strategicheskikh iaderniyi sil Rossii" ["The Dubiousness of the Concept of a Non-Nuclear Disarming Strike Against Russia's Strategic Nuclear Forces"], *Vestnik akademii voennykh nauk*, Vol. 54, No. 1, 2016, pp. 37–41.

Anderson, H. S., A. Kharkar, B. Filar, and P. Roth, *Evading Machine Learning Malware Detection*, blackhat.com, July 2017. As of August 15, 2017: https://www.blackhat.com/docs/us-17/thursday/us-17-Anderson-Bot-Vs-Bot-Evading-Machine-Learning-Malware-Detection-wp.pdf

Biggio, B., B. Nelson, and P. Laskov, "Poisoning Attacks Against Support Vector Machines," *Proceedings of the 29th International Conference on Machine Learning*, July 2012, pp. 1467–1474. As of August 15, 2017: https://arxiv.org/pdf/1206.6389.pdf

Bostrom, Nick, *Superintelligence: Paths, Dangers, Strategies*, Oxford: Oxford University Press, 2014.

Bracken, P., "The Intersection of Cyber and Nuclear War," *The Strategy Bridge*, blog post, January 17, 2017. As of August 15, 2017: https://thestrategybridge.org/the-bridge/2017/1/17/the-intersection-of-cyber-and-nuclear-war

Buchan, G., D. Matonick, C. Shipbaugh, and R. Mesic, *Future Roles of U.S. Nuclear Forces: Implications for U.S. Strategy*, Santa Monica, Calif.: RAND Corporation, MR-1231-AF, 2003. As of March 8, 2018: https://www.rand.org/pubs/monograph_reports/MR1231.html

Cimbala, S. J., *The Dead Volcano: The Background and Effects of Nuclear War Complacency*, Westport, Conn.: Praeger, 2002.

Etherington, D., "Google's AlphaGo AI Beats the World's Best Human Go Player," TechCrunch, May 23, 2017. As of August 15, 2017: https://techcrunch.com/2017/05/23/googles-alphago-ai-beats-the-worlds-best-human-go-player/

Garthoff, R. L., "Refocusing the SDI Debate," *Bulletin of the Atomic Scientists*, Vol. 43, No. 7, September 1987.

Geist, E., "Was There a Real 'Mineshaft Gap'? Bomb Shelters in the USSR, 1945–62," *Journal of Cold War Studies*, Vol. 14, No. 2, Spring 2012, pp. 3–28.

Goodfellow, I., Y. Bengio, and A. Courville, *Deep Learning*, Cambridge, Mass.: MIT Press, 2016.

Healey, D., *The Time of My Life*, London: Michael Joseph, 1989, p. 243.

Hoffman, D. E., *The Dead Hand*, New York: Doubleday, 2009.

Holmes, J. R., "Sea Changes: The Future of Nuclear Deterrence," *Bulletin of the Atomic Scientists*, Vol. 72, No. 4, 2016, pp. 228–233.

Kahn, H., *On Thermonuclear War*, Princeton, N.J.: Princeton University Press, 1960.

Karpathy, A., "Breaking Linear Classifiers on ImageNet," *Andrej Karpathy blog*, March 30, 2015. As of August 15, 2017: http://karpathy.github.io/2015/03/30/breaking-convnets/

Kearns, M., and M. Li, "Learning in the Presence of Malicious Errors," *SIAM Journal on Computing*, Vol. 22, No. 4, March 1992, pp. 807–837. As of August 15, 2017: https://doi.org/10.1137/0222052

Kegelmeyer, P., T. M. Shead, J. Crussell, K. Rodhouse, D. Robinson, C. Johnson, D. Zage, W. Davis, J. Wendt, J. Doak, T. Cayton, R. Colbaugh, K. Glass, B. Jones, and J. Shelburg, *Counter Adversarial Data Analytics*, Albuquerque, N.M.: Sandia National Laboratories, SAND2015-3711, May 2015. As of August 15, 2017: http://www.sandia.gov/~wpk/pubs/publications/cada-full-uur.pdf

Kerckhoff, A., "La Cryptographie Militaire," *Journal des Sciences Militaires*, January 1883.

Long, A., *Deterrence from Cold War to Long War: Lessons from Six Decades of RAND Research*, Santa Monica, Calif.: RAND Corporation, MG-636-OSD/AF, 2008. As of March 8, 2018: https://www.rand.org/pubs/monographs/MG636.html

Long, A., and B. R. Green, "Stalking the Secure Second Strike: Intelligence, Counterforce, and Nuclear Strategy," *Journal of Strategic Studies*, Vol. 38, Nos. 1–2, August 2012, pp. 38–76.

Mack, C. A., "Fifty Years of Moore's Law," *IEEE Transaction on Semiconductor Manufacturing*, Vol. 24, No. 2, May 2011, pp. 202–207. As of August 15, 2017: https://doi.org/10.1109/TSM.2010.2096437

Mahan, A. T., *Armaments and Arbitration: Or, The Place of Force in the International Relations of States*, New York: Harper & Brothers, 1912.

Minsky, M., *Semantic Information Processing*, Cambridge, Mass.: MIT Press, 1968.

Newell, A., J. C. Shaw, and H. A. Simon, *Report on a General Problem-Solving Program*, Santa Monica, Calif., RAND Corporation, Report P-1584, revised February 9, 1959.

Oznobishev, S. K., V. Ia. Potapov, and V. V. Skokov, *Kak gotovilisia "asymmetrichnyi otvet" na "Strategicheskuiu oboronnyiu initsiativu" R.Reigana. Velikhov, Kokoshin i drugie* [*How the "Asymmetric Response" to R. Reagan's "Strategic Defense Initiative" Was Prepared*], Moscow: Legand, 2008.

Papernot, N., P. McDaniel, and I. Goodfellow, "Transferability in Machine Learning: from Phenomena to Black-Box Attacks Using Adversarial Samples," arXiv, May 24, 2016. As of August 15, 2017:
https://arxiv.org/abs/1605.07277

Putin, V., "Byt' sil'nymi: garantii natsional'noi bezopasnosti dlia Rossii" ["Being Strong Is the Guarantee of National Security for Russia"], *Rossiiskaia gazeta*, February 20, 2012. As of December 5, 2017:
https://rg.ru/2012/02/20/putin-armiya.html/

Roland, A., and P. Shiman, *Strategic Computing: DARPA and the Quest for Machine Intelligence, 1983–1993*, Cambridge, Mass.: MIT Press, 2002.

"Rossiiskii Proekt 'Status-6' Meniaet Sootnosheniia Iadernykh Sil v Mire" ["The Russian Status-6 Project Is Changing the World's Nuclear Balance of Forces"], *Russkaia Politika*, November 14, 2016. As of December 4, 2016:
http://ruspolitica.ru/post/rossiyskiy-proekt-status-6-menyaet-sootnoshenie-yadernyih-sil-v-mire/

Schelling, T., *Arms and Influence*, New Haven, Conn.: Yale University Press, 1966.

Shannon, C. E. "Programming a Computer for Playing Chess," *Philosophical Magazine*, Vol. 41, No. 7, 1950, pp. 256–275.

Simon, H. A., and A. Newell, "Heuristic Problem Solving: The Next Advance in Operations Research," *Operations Research*, Vol. 6, No. 1, January–February 1958, pp. 1–10.

Slocombe, W., "The Countervailing Strategy," *International Security*, Vol. 5, No. 4, Spring 1981, pp. 18–27.

Sutyagin, I., "Russia's Underwater "Doomsday Drone": Science Fiction, but Real Danger," *Bulletin of the Atomic Scientists*, Vol. 72, No. 4, June 2016, pp. 243–246.

Timoshenko, M., 'Mertvaia ruka' na strazhe perimetra Rossii" ["The 'Dead Hand' Guarding Russia's Periphery"], *Telekanal "Zvezda,"* February 18, 2015. As of August 15, 2017:
http://tvzvezda.ru/news/krasnaya_zvezda/content/201502181414-gskc.htm

Tuchkov, V., *Status-6: Oruzhie Vosmezdia, Vognavshee Pentagon v Stupor* [*Status-6: The Retaliatory Weapon That Drove the Pentagon into a Stupor*], Svobodnaia Pressa, December 11, 2016. As of December 4, 2016:
http://svpressa.ru/war21/article/162378/

U.S. Department of Defense, *Nuclear Posture Review Report*, Washington, D.C., April 2010.

Valagin, A., "Garantirovannoe vozmezdie: Kak rabotaet rossiiskaia sistema 'Perimetr'" ["Assured Retaliation: How the Russian 'Perimetr' System Works"], *Rossiiskaia gazeta*, January 22, 2014. As of August 15, 2017:
https://rg.ru/2014/01/22/perimetr-site.html

Woyke, E., and Y. Kim, "Starcraft Pros Are Ready to Battle AI," *MIT Technology Review*, May 19, 2017. As of August 15, 2017:
https://www.technologyreview.com/s/607888/starcraft-pros-are-ready-to-battle-ai/

About This Perspective

We would like to thank RAND's Center for Global Risk and Security and its director, Andrew Parasiliti, for initiating this effort, and RAND Ventures for sponsoring it. We also thank Angela O'Mahoney and Bill Welser for their guidance throughout. We also indebted to Sonni Efron, Doug Irving, and Greg Baumann for their help in finding the stories as they emerged and Hosay Yaqub for making the events themselves a success. Finally, we would like to thank the workshop participants we leave unnamed in accordance with the Chatham House Rule.

Security 2040

This Perspective is part of a broader effort, an initiative of RAND Ventures, to envision critical security challenges in the world of 2040, considering the effects of political, technological, social, and demographic trends that will shape those security challenges in the coming decades. The research was conducted within the RAND Center for Global Risk and Security.

RAND Center for Global Risk and Security

The Center for Global Risk and Security (CGRS) works across the RAND Corporation to develop multidisciplinary research and policy analysis dealing with systemic risks to global security. The center draws on RAND's expertise to complement and expand RAND research in many fields, including security, economics, health, and technology. A board of distinguished business leaders, philanthropists, and former policymakers advises and supports the center activities, which are increasingly focused on global security trends and the impact of disruptive technologies on risk and security. For more information about the RAND Center for Global Risk and Security, visit www.rand.org/international/cgrs

RAND Ventures

RAND is a research organization that develops solutions to public policy challenges to help make communities throughout the world safer and more secure, healthier and more prosperous. RAND is nonprofit, nonpartisan, and committed to the public interest.

RAND Ventures is a vehicle for investing in policy solutions. Philanthropic contributions support our ability to take the long view, tackle tough and often-controversial topics, and share our findings in innovative and compelling ways. RAND's research findings and recommendations are based on data and evidence, and therefore do not necessarily reflect the policy preferences or interests of its clients, donors, or supporters.

Funding for this venture was provided by gifts from RAND supporters and income from operations.

About the Authors

Edward Geist is an associate policy researcher at RAND. Previously a MacArthur Nuclear Security fellow at Stanford University's Center for International Security and Cooperation (CISAC) and a Stanton Nuclear Security Fellow in RAND's Washington office, Edward received his doctorate in Russian history from the University of North Carolina in May 2013.

Andrew J. Lohn is an engineer at the RAND Corporation. He applies a wide range of mathematical and machine learning techniques to provide new insights into highly technical policy issues, such as cyberwarfare, artificial intelligence, or drone delivery. Lohn holds a doctorate in electrical engineering from the University of California, Santa Cruz.