AI and International Stability

Risks and Confidence-Building Measures

Michael C. Horowitz
Paul Scharre

Center for a
New American
Security

Center for a New American Security
1152 15th Street NW, Suite 950, Washington, DC 20005
T: 202.457.9400 | F: 202.457.9401 | CNAS.org | @CNASdc

Acknowledgements

The authors thank Lora Saalman, Helen Toner, and Luke Muehlhauser for their thoughtful reviews of this report. Thank you to Maura McCarthy, Melody Cook, Emma Swislow, Chris Estep, Megan Lamberth, and Lauren Kahn for their work in the production and design of this report.

About the Authors

Michael C. Horowitz is an Adjunct Senior Fellow in the Technology and National Security Program at the Center for a New American Security. He is Director of Perry World House and Richard Perry Professor at the University of Pennsylvania, the author of *The Diffusion of Military Power: Causes and Consequences for International Politics*, and co-author of *Why Leaders Fight*. Dr. Horowitz won the 2017 Karl Deutsch Award given by the International Studies Association for early career contributions to the fields of international relations and peace research. He has published in a wide array of peer-reviewed journals and popular outlets. His research interests include the intersection of emerging technologies such as artificial intelligence (AI) and robotics with global politics, military innovation, the role of leaders in international politics, and geopolitical forecasting methodology. Dr. Horowitz previously worked for the Office of the Undersecretary of Defense for Policy in the Department of Defense and is a member of the Council on Foreign Relations. Dr. Horowitz received his PhD in government from Harvard University and his BA in political science from Emory University.

Paul Scharre is a Senior Fellow and Director of the Technology and National Security Program at the Center for a New American Security (CNAS). He is the award-winning author of *Army of None: Autonomous Weapons and the Future of War*, which won the 2019 Colby Award and was one of Bill Gates's top five books of 2018. Dr. Scharre worked in the Office of the Secretary of Defense in the Bush and Obama administrations where he played a leading role in establishing policies on unmanned and autonomous systems and emerging weapons technologies. He led the Department of Defense working group that drafted DoD Directive 3000.09, establishing the Department's policies on autonomy in weapon systems. He holds a PhD in war studies from King's College London and an MA in political economy and public policy and a BS in physics, cum laude, from Washington University in St. Louis. Prior to working in the Office of the Secretary of Defense, Scharre served as an infantryman, sniper, and reconnaissance team leader in the Army's 3rd Ranger Battalion and completed multiple tours to Iraq and Afghanistan. He is a graduate of the Army's Airborne, Ranger, and Sniper Schools and Honor Graduate of the 75th Ranger Regiment's Ranger Indoctrination Program.

About This Report

This report is part of the Artificial Intelligence and International Stability Project at CNAS and draws on insights from workshops conducted in Washington, at Oxford University, at Stanford University, and virtually in 2019 and 2020. The project was made possible by a grant from Carnegie Corporation of New York. For additional CNAS work on AI and global security, see cnas.org/AI.

Table of Contents

Center for a New American Security
1152 15th Street NW, Suite 950, Washington, DC 20005
T: 202.457.9400 | F: 202.457.9401 | CNAS.org | @CNASdc

Executive Summary

Militaries around the world believe that the integration of machine learning methods throughout their forces could improve their effectiveness. From algorithms to aid in recruiting and promotion, to those designed for surveillance and early warning, to those used directly on the battlefield, applications of artificial intelligence (AI) could shape the future character of warfare. These uses could also generate significant risks for international stability. These risks relate to broad facets of AI that could shape warfare, limits to machine learning methods that could increase the risks of inadvertent conflict, and specific mission areas, such as nuclear operations, where the use of AI could be dangerous. To reduce these risks and promote international stability, we explore the potential use of confidence-building measures (CBMs), constructed around the shared interests that all countries have in preventing inadvertent war. Though not a panacea, CBMs could create standards for information-sharing and notifications about AI-enabled systems that make inadvertent conflict less likely.

Introduction

In recent years, the machine learning revolution has sparked a wave of interest in artificial intelligence (AI) applications across a range of industries. Nations are also mobilizing to use AI for national security and military purposes.[1] It is therefore vital to assess how the militarization of AI could affect international stability and how to encourage militaries to adopt AI in a responsible manner. Doing so requires understanding the features of AI, the ways it could shape warfare, and the risks to international stability resulting from the militarization of artificial intelligence.

AI is a general-purpose technology akin to computers or the internal combustion engine, not a discrete technology like missiles or aircraft. Thus, while concerns of an "AI arms race" are overblown, real risks exist.[2] Additionally, despite the rhetoric of many national leaders, military spending on AI is relatively modest to date. Rather than a fervent arms race, militaries' pursuit of AI looks more like routine adoption of new technologies and a continuation of the multi-decade trend of adoption of computers, networking, and other information technologies. Nevertheless, the incorporation of AI into national security applications and warfare poses genuine risks. Recognizing the risks is not enough, however. Addressing them requires laying out suggestions for practical steps states can take to minimize risks stemming from military AI competition.[3] One approach states could take is adopting confidence-building measures (CBMs): unilateral, bilateral, and/or multilateral actions that states can take to build trust and prevent inadvertent military conflict. CBMs generally involve using transparency, notification, and monitoring to attempt to mitigate the risk of conflict.[4] There are challenges involved in CBM adoption due to differences in the character of international competition today versus during the Cold War, when CBMs became prominent as a concept. However, considering possibilities for CBMs and exploring ways to shape the dialogue about AI could make the adoption of stability-promoting CBMs more likely.

This paper briefly outlines some of the potential risks to international stability arising from military applications of AI, including ways AI could influence the character of warfare, risks based on the current limits of AI technology, and risks relating to some specific mission areas, such as nuclear operations, in which introducing AI could present challenges to stability. The paper then describes possible CBMs to address these risks, moving from broad measures applicable to many military applications of AI to targeted measures designed to address specific risks. In each discussion of CBMs, the paper lays out both the opportunities and potential downsides of states adopting the CBM.

Center for a New American Security
1152 15th Street NW, Suite 950, Washington, DC 20005
T: 202.457.9400 | F: 202.457.9401 | CNAS.org | @CNASdc

Military Uses of AI: A Risk to International Stability?

Militaries have an inherent interest in staying ahead of their competitors, or at least not falling behind. National militaries want to avoid fielding inferior military capabilities and so will generally pursue emerging technologies that could improve their ability to fight. While the pursuit of new technologies is normal, some technologies raise concerns because of their impact on stability or their potential to shift warfare in a direction that causes net increased harm for all (combatants and/or civilians). For example, around the turn of the 20th century, great powers debated, with mixed results, arms control against a host of industrial era technologies that they feared could alter warfare in profound ways. These included submarines, air-delivered weapons, exploding bullets, and poison gas.

After the invention of nuclear weapons, concerns surrounding their potential use dominated the attention of policymakers given the weapons' sheer destructive potential. Especially after the Cuban Missile Crisis illustrated the very real risk of escalation, the United States and the Soviet Union engaged in arms control on a range of weapons technologies, including strategic missile defense, intermediate-range missiles, space-based weapons of mass destruction (WMDs), biological weapons, and apparent tacit restraint in neutron bombs and anti-satellite weapons. The United States and the Soviet Union also, at times, cooperated to avoid miscalculation and improve stability through measures such as the Open Skies Treaty and the 1972 Incidents at Sea Agreement.

It is reasonable and, in fact, vital to examine whether the integration of AI into warfare might also pose risks that policymakers should attend. Some AI researchers themselves have raised alarm at militaries' adoption of AI and the way it could increase the risk of war and international instability.[5] Understanding risks stemming from military use of AI is complicated, however, by the fact that AI is not a discrete technology like missiles or submarines. As a general-purpose technology, AI has many applications, any of which could, individually, improve or undermine stability in various ways.

Militaries are only beginning the process of adopting AI, and in the near term, military AI use is likely to be limited and incremental. Over time, the cognization of warfare through the introduction of artificial intelligence could change warfare in profound ways, just as industrial revolutions in the past shaped warfare.[6] Even if militaries successfully manage safety and security concerns and field AI systems that are robust and secure, properly functioning AI systems could create challenges for international stability.

For example, both Chinese and American scholars have hypothesized that the introduction of AI and autonomous systems in combat operations could accelerate the tempo of warfare beyond the pace of human control. Chinese scholars have referred to this concept as a battlefield "singularity,"[7] while some Americans have coined the term "hyperwar" to refer to a similar idea.[8] If warfare evolves to a point where the pace of combat outpaces humans' ability to keep up, and therefore control over military operations must be handed to machines, it would pose significant risks for international stability, even if the delegation decision seems necessary due to competitive pressure. Humans might lose control over managing escalation, and war termination could be significantly complicated if machines fight at a pace that is faster than humans can respond. In addition, delegation of escalation control to machines could mean that minor tactical missteps or accidents that are part and parcel of military operations in the chaos and fog of war, including fratricide, civilian casualties, and poor military judgment, could spiral out of control and reach catastrophic proportions before humans have time to intervene.

The logic of a battlefield singularity, or hyperwar, is troubling precisely because competitive pressures could drive militaries to accelerate the tempo of operations and remove humans "from the loop," even if they would rather not, in order to keep pace with adversaries. Then-Deputy Secretary of Defense Robert Work succinctly captured this dilemma when he posed the question, "If our competitors go to Terminators

... and it turns out the Terminators are able to make decisions faster, even if they're bad, how would we respond?"[9] While this "arms race in speed" is often characterized tactically in the context of lethal autonomous weapon systems, the same dynamic could emerge operationally involving algorithms designed as decision aids. The perception by policymakers that war is evolving to an era of machine-dominated conflict in which humans must cede control to machines to remain competitive could also hasten such a development, particularly if decision makers lack appropriate education about the limits of AI. In extremis, the shift toward the use of algorithms for military decision-making, combined with a more roboticized battlefield, could potentially change the nature of war. War would still be the continuation of politics by other means in the broadest sense, but in the most extreme case it might feature so little human engagement that it is no longer a fundamentally human endeavor.[10]

The widespread adoption of AI could have a net effect on international stability in other ways. AI systems could change strategy in war, including by substituting machines for human decision-making in some mission areas, and therefore removing certain aspects of human psychology from parts of war.[11] Warfare today is waged by humans through physical machinery (rockets, missiles, machine guns, etc.), but decision-making is almost universally human. As algorithms creep closer to the battlefield, some decisions will be made by machines even if warfare remains a human-directed activity that is fought for human political purposes. The widespread integration of machine decision-making across tactical, operational, and strategic levels of warfare could have far-reaching implications. Already, AI agents playing real-time computer strategy games such as StarCraft and Dota 2 have demonstrated superhuman aggressiveness, precision, and coordination. In other strategy games such as poker and Go, AI agents have demonstrated an ability to radically adjust playing styles and risk-taking in ways that would be, at best, challenging for humans to mimic for psychological reasons. AI dogfighting agents have similarly demonstrated superhuman precision and employed different tactics because of the ability to take greater risk to themselves.[12]

In many ways, AI systems have the ability to be the perfect strategic agents, unencumbered by fear, loss aversion, commitment bias, or other human emotional or cognitive biases and limitations.[13] While the specific algorithms and models used for computer games are unlikely to transfer well to combat applications, the general characteristics and advantages of AI agents relative to humans could have applications in the military domain. As in the case of speed, the net consequence of machine decision-making on the psychology of combat could change the character of warfare in profound ways.[14]

AI could have other cumulative effects on warfare. Policymakers generally assess adversaries' behavior based on an understanding of their capabilities and intentions.[15] Shifts toward AI could undermine policymaker knowledge in both of those arenas. The transition of military capabilities to software, already underway but arguably accelerated by the adoption of AI and autonomous systems, could make it harder for policymakers to accurately judge relative military capabilities. Incomplete information about adversary capabilities would therefore increase, conceivably increasing the risks of miscalculation. Alternatively, the opposite could be true—AI and autonomous systems used for intelligence collection and analysis could radically increase transparency about military power, making it easier for policymakers to judge military capabilities and anticipate the outcome of a conflict in advance. This added transparency could decrease the risks of miscalculation and defuse some potential conflicts before they begin.

The integration of AI into military systems, in combination with a shift toward a more roboticized force structure, could also change policymakers' threshold for risk-taking, either because they believe that fewer human lives are at risk or that AI systems enable greater precision, or perhaps because they see AI systems as uniquely dangerous. The perceived availability of AI systems could change policymakers' beliefs about their ability to foresee the outcome of conflicts or to win.

Center for a New American Security
1152 15th Street NW, Suite 950, Washington, DC 20005
T: 202.457.9400 | F: 202.457.9401 | CNAS.org | @CNASdc

It is, no doubt, challenging to stand at the beginning of the AI age and imagine the cumulative consequence of AI adoption across varied aspects of military operations, including effects that hinge as much on human perception of the technology as the technical characteristics themselves. The history of attempts to regulate the effects of industrial age weapons in the late 19th and early 20th centuries suggests that even when policymakers accurately anticipated risks from certain technologies, such as air-delivered weapons or poison gas, they frequently crafted regulations that turned out to be ill-suited to the specific forms these technologies took as they matured. Furthermore, even when both sides desired restraint, it frequently (although not always) collapsed under the exigencies of war.[16] There is no reason to think that our prescience in predicting the path of future technologies or ability to restrain warfare is any better today. There is merit, however, in beginning the process of thinking about the many ways in which AI could influence warfare, big and small.

Even beyond the scenarios described above, it is possible to frame how military applications of AI could impact international stability into two broad categories: (1) risks related to the character of algorithms and their use by militaries, and (2) risks related to militaries using AI for particular missions.

RISKS DUE TO THE LIMITATIONS OF AI

A challenge for military adoption of AI is that two key risks associated with new technology adoption are in tension. First, militaries could fail to adopt—or adopt quickly enough or employ in the right manner—a new technology that yields significant battlefield advantage. As a recent example, despite the overall growth in the military uninhabited, or unmanned, aircraft market, the adoption of uninhabited vehicles has, at times, been a source of contention within the U.S. defense establishment, principally based on debates over the merits of this new technology relative to existing alternatives.[17]

Alternatively, militaries could adopt an immature technology too quickly, betting heavily and incorrectly on new and untested propositions about how a technology may change warfare. Given the natural incentive militaries have in ensuring their capabilities work on the battlefield, it may be reasonable to assume that militaries would manage these risks reasonably well, although not without some mishaps. But when balancing the risk of accidents versus falling behind adversaries in technological innovation, militaries arguably place safety as a secondary consideration.[18] Militaries may be relatively accepting of the risk of accidents in the pursuit of technological advantage, since accidents are a routine element of military operations, even in training.[19] Nevertheless, there are strong bureaucratic interests in ultimately ensuring that fielded capabilities are robust and secure, and existing institutional processes may be able to manage AI safety and security risks with some adaptation.

For militaries, balancing between the risks of going too slow versus going too fast with AI adoption is complicated by the fact that AI, and deep learning in particular, is a relatively immature technology with significant vulnerabilities and reliability concerns.[20] These concerns are heightened in situations where there may not be ample data on which to train machine learning systems. Machine learning systems generally rely on very large data sets, which may not exist in some military settings, particularly when it comes to early warning of rare events (such as a nuclear attack) or tracking adversary behavior in a multidimensional battlefield. When trained with inadequate data sets or employed outside the narrow context of their design, AI systems are often unreliable and brittle. AI systems can often seem deceptively capable, performing well (sometimes better than humans) in some laboratory settings, then failing dramatically under changing environmental conditions in the real world. Self-driving cars, for example, may be safer than human drivers in some settings, then inexplicably turn deadly in situations where a human operator would not have trouble. Additionally, deep learning methods may, at present, be insufficiently reliable for safety-critical applications even when operating within the bounds of their design specifications.[21]

Center for a New American Security
1152 15th Street NW, Suite 950, Washington, DC 20005
T: 202.457.9400 | F: 202.457.9401 | CNAS.org | @CNASdc

For example, concerns about limits to the reliability of algorithms across demographic groups have hindered the deployment of facial recognition technology in the United States, particularly in high-consequence applications such as law enforcement. Militaries, too, should be concerned about technical limitations and vulnerabilities in their AI systems. Militaries want technologies that work, especially on the battlefield. Accordingly, the AI strategy of the Department of Defense (DoD) calls for AI systems that are "resilient, robust, reliable, and secure."[22] This is undoubtedly the correct approach but a challenge, at least in the near term, given the reliability issues facing many uses of algorithms today and the highly dynamic conditions of battlefield use.

An additional challenge stems from security dilemma dynamics. Competitive pressures could lead nations to shortcut test and evaluation (T&E) in a desire to field new AI capabilities ahead of adversaries. Similar competitive pressures to beat others to market appear to have played an exacerbating role in accident risk relating to AI systems in self-driving cars and commercial airplane autopilots.[23] Militaries evaluating an AI system of uncertain reliability could, not unjustifiably, feel pressure to hasten deployment if they believe others are taking similar measures. Historically, these pressures are highest immediately before and during wars, where the risk/reward equation surrounding new technologies can shift due to the very real lives on the line. For example, competitive pressures may have spurred the faster introduction of poison gas in World War I.[24] Similarly, in World War II, Germany diverted funds from proven technologies into jet engines, ballistic missiles, and helicopters, even though none of the technologies proved mature until after the war.[25] This dynamic risk might spark a self-fulfilling prophecy in which countries accelerate deployment of insufficiently tested AI systems out of the fear that others will deploy first.[26] The net effect is not an arms race but a "race to the bottom" on safety, leading to the deployment of unsafe AI systems and heightening the risk of accidents and instability.

Even if military AI systems are adequately tested, the use of AI to enable more autonomous machine behavior in military systems raises an additional set of risks. In delegating decision-making from humans to machines, policymakers may de facto be fielding forces with less flexibility and ability to understand context, which would then have deleterious effects on crisis stability and managing escalation. While machines have many advantages in speed, precision, and repeatable actions, machines today cannot come close to human intelligence in understanding context and flexibly adapting to novel situations. This brittleness of machine decision-making may particularly be a challenge in pre-conflict crisis situations, where tensions among nations run high. Military forces from competing nations regularly interact in militarized disputes below the threshold of war in a variety of contested regions (e.g., the India-Pakistan border, China-India border, South China Sea, Black Sea, Syria, etc.). These interactions among deployed forces sometimes run the risk of escalation due to incidents or skirmishes that can inflame tensions on all sides. This poses a challenge for national leaders, who have imperfect command-and-control over their own military forces. Today, however, deployed military forces rely on human decision-making. Humans can understand broad guidance from their national leadership and commander's intent, such as "defend our territorial claims, but don't start a war." Relative to humans, even the most advanced AI systems today have no ability to understand broad guidance, nor do they exhibit the kinds of contextual understanding that humans frequently label "common sense."[27] Militaries already employ uninhabited vehicles (drones) in contested areas, which have been involved in a number of escalatory incidents in the East China Sea, South China Sea, Syria, and Strait of Hormuz.[28] Over time, as militaries incorporate more autonomous functionality into uninhabited vehicles, that functionality could complicate interactions in these and other contested areas.

Autonomous systems may take actions based on programming that, while not a malfunction, are other than what a commander would have wanted a similarly situated human to do in the same situation. While the degree of flexibility afforded subordinates varies considerably by military culture and doctrine, humans have a greater ability to flexibly respond to complex and potentially ambiguous escalatory incidents in ways that may balance competing demands of ensuring national resolve while managing escalation.[29] Autonomous systems will simply follow their programming, whatever that may be, even if those rules no

longer make sense or are inconsistent with a commander's intent in the given situation. This challenge is compounded by the fact that human commanders cannot anticipate all of the possible situations that forward-deployed military forces in contested regions may face. Employing autonomous systems in a crisis effectively forces human decision makers to tie their own hands with certain pre-specified actions, even if they would rather not.

Unintended actions by autonomous systems in militarized disputes or contested areas are a challenge for militaries as they adopt more autonomous systems into their forces. The complexity of many autonomous systems used today, even ones that rely on rule-based decision-making, may mean that the humans employing autonomous systems lack sufficient understanding of what actions the system may take in certain situations.[30] Humans' ability to flexibly interpret guidance from higher commanders, even to the point of disregarding guidance if it no longer seems applicable, is by contrast a boon to managing escalation risks by retaining human decision-making at the point of interaction among military forces in contested regions.[31]

Unintended escalation is not merely confined to lethal actions, such as firing on enemy forces. Nonlethal actions, such as crossing into another state's territory, can be perceived as escalatory. Even if such actions do not lead directly to war, they could heighten tensions, increase suspicion about an adversary's intentions, or inflame public sentiment. While in most cases, humans would still retain agency over how to respond to an incident, competing autonomous systems could create unexpected interactions or escalatory spirals. Complex, interactive dynamics between algorithms have been seen in other settings, including financial markets,[32] and even in situations where the algorithms are relatively simple.[33] Another problem stems from the potential inability of humans to call off autonomous systems once deployed. One reason for employing autonomous functionality is so that uninhabited vehicles can continue their missions even if they are operating without reliable communication links to human controllers. When there is no communication link between human operators and an autonomous system, human operators would have no ability to recall the autonomous system if political circumstances changed such that the system's behavior was no longer appropriate. This could be a challenge in de-escalating a conflict, if political leaders decide to terminate hostilities but have no ability to recall autonomous systems, at least for some period of time. The result could be a continuation of hostilities even after political leaders desire a cease-fire. Alternatively, the inability to fully cease hostilities could undermine truce negotiations, leading to the continuation of conflict. These problems are not unique to autonomous systems. Political leaders have imperfect command-and-control over human military forces, which has, at times, led to similar incidents with human-commanded deployed forces. For example, the Battle of New Orleans in the War of 1812 was fought after a peace treaty ended the war because of the slowness of communications to deployed forces.

RISKS DUE TO THE USE OF AI FOR PARTICULAR MILITARY MISSIONS

The introduction of AI into military operations could also pose risks in certain circumstances due to the nature of the military mission, even if the AI system performs correctly and consistent with human intentions. Some existing research already focuses on the intersection of AI with specific military mission areas, most notably nuclear stability.[34] Nuclear stability is an obvious area of concern given the potential consequences of an intentional or unintentional nuclear detonation.[35] Lethal autonomous weapon systems (LAWS), a particular use of AI in which lethal decision-making is delegated from humans to machines, also represents a focus area of existing research. Other areas may deserve special attention from scholars concerned about AI risks. The intersections of AI with cybersecurity and biosecurity are areas worthy of exploration where there has been relatively less work at present.[36]

Center for a New American Security
1152 15th Street NW, Suite 950, Washington, DC 20005
T: 202.457.9400 | F: 202.457.9401 | CNAS.org | @CNASdc

Potentially risky applications of AI extend beyond the battlefield to the use of AI to aid in decision-making in areas such as early warning and forecasting adversary behavior. For example, AI tools to monitor, track, and analyze vast amounts of data on adversary behavior for early indications and warning of potential aggression have clear value. However, algorithms also have known limitations and potentially problematic characteristics, such as a lack of transparency or explainability, brittleness in the face of distributional shifts in data, and automation bias. AI systems frequently perform poorly under conditions of novelty, suggesting a continued role for human judgment. The human tendency toward automation bias, coupled with the history of false alarms generated by non-AI early warning and forecasting systems, suggests policymakers should approach the adoption of AI in early warning and forecasting with caution, despite the potential value of using AI in intelligent decision aids.[37] Education and training to ensure the responsible use of AI in early warning and forecasting scenarios will be critical.[38]

Finally, autonomous systems raise novel challenges of signaling in contested areas because of ambiguity about whether their behavior was intended by human commanders. Even if the system performs as intended, adversaries may not know whether an autonomous system's behavior was consistent with human intent because of the aforementioned command-and-control issues. This can create ambiguity in a crisis situation about how to interpret an autonomous system's behavior. For example, if an autonomous system fired on a country's forces, should that be interpreted as an intentional signal by the commanding nation's political leaders, or an accident? This, again, is not a novel problem; a similar challenge exists with human-commanded military forces. Nations may not know whether the actions of an adversary's deployed forces are fully in line with their political leadership's guidance. Autonomous systems could complicate this dynamic due to uncertainty about whether the actions of an autonomous system are consistent with any human's intended action.

The Role of Confidence-Building Measures

AI potentially generates risks for international security due to ways AI could change the character of warfare, the limitations of AI technology today, and the use of AI for specific military missions such as nuclear operations. Especially given the uncertain technological trajectory of advances in AI, what are options to reduce the risks that military applications of AI can pose to international stability?

To advance the conversation about ensuring that military AI adoption happens in the safest and most responsible way possible, this paper outlines a series of potential confidence-building measures aimed at mitigating risks from military uses of AI.[39] We introduce these ideas as preliminary concepts for future research, discussion, and examination, rather than to specifically advocate for any of these options. But progress in mitigating the risks from military AI competition requires moving beyond the recognition that risk mitigation is important to the hard work of suggesting, evaluating, and examining the benefits and drawbacks of specific mechanisms.[40]

This paper focuses on confidence-building measures, a broad category of actions that states can take to reduce instability risks. CBMs include actions such as transparency, notification, and monitoring designed to reduce various risks arising from military competition between states. They generally encompass four areas, as Marie-France Desjardins describes:[41]

- Information-sharing and communication
- Measures to allow for inspections and observers
- "Rules of the road" to govern military operations
- Limits on military readiness and operations

Center for a New American Security
1152 15th Street NW, Suite 950, Washington, DC 20005
T: 202.457.9400 | F: 202.457.9401 | CNAS.org | @CNASdc

Confidence-building measures are related to, but distinct from, arms control agreements. Arms control encompasses agreements states make to forgo researching, developing, producing, fielding, or employing certain weapons, features of weapons, or applications of weapons. The set of possible actions states could take is broad, and this paper will focus on the potential benefits and drawbacks of specific AI-related confidence-building measures. Arms control for military AI applications is a valuable topic worthy of exploration, but beyond the scope of this paper.[42]

HISTORICAL APPLICATIONS OF CBMS

Confidence-building measures as a concept rose to prominence during the Cold War as a tool to reduce the risk of inadvertent war. In the wake of the Cuban Missile Crisis, the United States and the Soviet Union began exploring ways to improve their communication. While both sides recognized that war might occur, they had a shared interest, due to the potentially world-ending consequences of a global nuclear war, in ensuring that any such outbreak would be due to a deliberate decision, rather than an accident or a misunderstanding.

The desire to build confidence led to a series of bilateral measures. Less than a year after the Cuban Missile Crisis, in June 1963, the United States and the Soviet Union signed a memorandum of understanding to create a hotline between the senior leadership of the two nations.[43] The idea was that this line of communication would provide a mechanism for U.S. and Soviet leaders to reach out to their counterparts and discuss crises in a way that made inadvertent escalation less likely. In 1972, as part of the Strategic Arms Limitation Talks (SALT I) arms control agreement, the United States and the Soviet Union went further, signing the Incidents at Sea Agreement, which they had been negotiating since 1967. The Incidents at Sea Agreement, not initially considered a prominent part of the 1972 SALT I accord, created a mechanism for communication and information surrounding the movement of U.S. and Soviet naval vessels. The agreement regulated dangerous maneuvers and harassment of vessels, established means for communicating the presence of submarines and surface naval movements, and generated a mechanism for regular consultation.[44] These successes helped lead to the formalization of the CBM concept in 1975 in Helsinki at the Conference on Security and Cooperation in Europe.[45]

As the Cold War drew to a close, confidence-building measures expanded beyond the U.S.-Soviet context and the European theater. For example, India and China have a series of CBMs intended to prevent escalation in their disputed border area, while India and Pakistan have a hotline designed to make accidental escalation in South Asia less likely. In Southeast Asia, through the Regional Forum of the Association of Southeast Asian Nations (ASEAN), member nations have pursued CBMs designed to reduce the risk of conflict among themselves, and between any ASEAN member and China, due to territorial disputes in the South China Sea.[46] These CBMs used outside of the Cold War have had mixed effects.

In the China-India case, for example, border-related CBMs did not prevent the ongoing conflict in 2020 between those two nations along the Line of Actual Control in the Himalayan region. However, norms surrounding the types of “legitimate” military activities promoted by CBMs have likely reduced the death toll of the clashes, with both sides generally avoiding the use of firearms, consistent with agreements from 1996 and 2005.[47]

In Southeast Asia, while the ASEAN Regional Forum is a principal forum for dialogue, the consensus-based character of ASEAN makes it challenging for that dialogue to translate into policies on contested issues. Recent multilateral dialogues about emerging technologies such as cyber systems have also featured efforts to create CBMs that could be building blocks for cooperation. Unfortunately, a lack of international agreement on basic definitions and some countries’ interest in dodging limitations on behavior in cyberspace have limited the development of effective norms.[48] CBMs rely on shared interests

to succeed, and major powers such as the United States, China, and Russia do not have clearly shared interests concerning behavior in cyberspace, making it difficult to use CBMs to build trust or successfully design "rules of the road" agreements likely to generate widespread adherence.

CBMs may be a useful tool for managing risks relating to military AI applications. There are a number of possible CBMs that states could adopt that may help mitigate the various AI-related risks previously outlined. These include broad CBMs applicable to AI as a category, CBMs designed to address some of the limitations of AI, and CBMs focused on specific missions for which militaries might use AI.[49]

BROAD CBMS

These CBMs focus broadly on mechanisms for dialogue and agreement surrounding military uses of AI, rather than the specific content of agreements. Given that a key goal of CBMs is to enhance trust, mechanisms that serve as a building block for more substantive dialogue and agreement can, in some cases, be an end in themselves and not just a means to an end.[50] These could include promoting international norms for how nations develop and use military AI systems, Track II academic-to-academic exchanges, direct military-to-military dialogues, and agreements between states regarding military AI, such as a code of conduct or mutual statement of principles.

Promoting Norms

In 2019, the U.S. Defense Innovation Board proposed a set of AI principles for the U.S. Defense Department, which DoD subsequently adopted in early 2020. While these principles no doubt have domestic audiences in the U.S. defense community and tech sector, they also serve as an early example of a state promulgating norms about appropriate use of AI in military applications. The DoD AI principles included a requirement that DoD AI systems be responsible, equitable, traceable, reliable, and governable.[51] (The full set of DoD AI principles is included in the Appendix). Similarly, the DoD's unclassified summary of its AI strategy, released in 2019, called for building AI systems that were "resilient, robust, reliable, and secure."[52] A focus of the strategy was "leading in military ethics and AI safety."[53]

There is value in states promoting norms for responsible use of AI, including adopting and employing technology in a way that reflects an understanding of the technical risks associated with AI systems. While stating such principles is not the same as putting in place effective bureaucratic processes to ensure their compliance, there is nevertheless value in states publicly signaling to others (and to their own bureaucracies) the importance of using AI responsibly in military applications. While these norms are at a high level, they nevertheless signal some degree of attention by senior military and civilian defense officials to some of the risks of AI systems, including issues surrounding safety, security, responsibility, and controllability. These signals may aid internal bureaucratic efforts to mitigate various AI-related risks, as bureaucratic actors can point to these official documents for support. Additionally, to the extent that other nations find these statements credible, they may help signal to other nations at least some degree of awareness and attention to these risks, helping to incentivize others to do the same.

One risk to such statements is that if they appear manifestly at odds with a state's actions, they can ring hollow, undermine a state's credibility, or undermine the norm itself. For example, loudly proclaiming the importance of AI ethics while using AI systems in a clearly unethical manner, such as for internal repression or without regard for civilian casualties, could not only undermine a state's credibility but also undermine the value of the norm overall, especially if other states fail to highlight the disconnect. Following through with meaningful actions to show how a state puts these norms into practice is essential for them to have real value.

Track II Academic Dialogues

One confidence-building measure is already underway: Track II dialogues between academic experts from different countries with expertise surrounding military uses of AI.[54] Because these dialogues occur among experts who are not government officials, they are low risk because they do not commit countries to actually doing anything. This also places a cap on their potential benefits. Track II dialogues can nevertheless be useful building blocks for more substantive cooperation among countries and an avenue to explore various potential modes of cooperation without fear of commitment by states. Track II dialogues can help facilitate mutual understanding among expert communities in different states and build shared trust between experts.[55] Additionally, if some of those experts transition into government positions in the future, the lessons from these dialogues can improve the prospects for cooperation in more formal venues.

While there are risks to misleading statements in the context of formal government dialogues, as discussed below, the consequences of such activities in a Track II context are minimal. The nature of the dialogue is that participants are not government officials and it is to be expected that some of their statements may not be entirely in line with their government's policies. Thus, Track II dialogues can build trust and be an end in themselves, even as they serve as the means to broader cooperation and understanding.

Military-to-Military Dialogues

Direct military-to-military engagement on deconfliction measures for AI and autonomous systems could be valuable, both as a precursor to potentially more fulsome specific measures, but also a valuable communication mechanism in their own right. For example, if militaries deploy an autonomous vehicle into a contested area where other military forces will be operating, a direct military-to-military channel would give the other side an opportunity to ask questions about its behavior and the deploying side an opportunity to communicate expectations, to avoid unintended escalation or incidents. Similarly, such a venue would give militaries an opportunity to ask questions and communicate information about other capabilities or investments that may threaten mutual stability, such as investments in AI, autonomy, or automation in nuclear operations. There are many advantages of direct, private communication over more indirect, public communication. Nations can send targeted messages just to the intended audience, rather than dealing with multiple audiences, including domestic ones. There may be reduced political pressure to save face or show strength publicly, although of course some of these pressures may still exist in private channels. And direct discussions afford more high-bandwidth information exchange with greater back-and-forth between sides than may be possible via public messages broadcast to a wider audience.

One challenge, of course, is that these dialogues are most challenging precisely when they are needed the most, when there is a lack of transparency and trust on both sides. However, history shows that such dialogues are possible and indeed can be valuable measures in increasing transparency and reducing mutual risks.

Code of Conduct

Nations could agree to a written set of rules or principles for how they adopt AI into military systems. These rules and principles, even if not legally binding, could nevertheless serve a valuable signaling and coordination function to avoid some of the risks in AI adoption. A code of conduct, statement of principles, or other agreement could include a wide range of both general and specific statements, including potentially on any or all of the confidence-building measures listed above.

Even if countries cannot agree on specific details beyond promoting safe and responsible military use of AI, more general statements could nevertheless be valuable in signaling to other nations some degree of mutual understanding about responsible use of military AI and help create positive norms of behavior. Ideally, a code of conduct would have support from a wide range of countries and major military powers. However, if this were not possible, then a multilateral statement of principles from like-minded countries could still have some value in increasing transparency and promulgating norms of responsible state behavior.

There are a few potential drawbacks to a broad code of conduct. First, a broader code of conduct, lacking the specificity of some of the measures discussed above, might undercut momentum toward broader cooperation, rather than serve as a building block. Second, there would be risk in negotiating a code of conduct that disagreements over some of the specifics could derail the entire endeavor or lead to forum shopping, whereby countries then spin off to create their own dialogues about a code of conduct. This is arguably what has happened in the cyber realm, where several different ongoing dialogue processes about codes of conduct have not led to substantive success. Third, a more formal code of conduct might start to raise the prospects of triggering some of the costs associated with CBMs. Specifically, if a country reduced its investments in military applications of AI or did not pursue capability areas because it believed adversaries were following a code of conduct, it could expose itself in the event of cheating. This might be of particular concern for democracies, given that, in many cases, democracies are more likely to comply with the agreements they sign, in part because democracies often have rigorous internal bureaucratic processes to ensure compliance.[56] Thus, one might imagine that the incentives might lead to a less formal code of conduct designed as a building block, rather than something that might cause countries to restrain capabilities.

THE LIMITATIONS OF AI

Accident risk is a significant concern for military applications of AI. Competitive pressures could increase accident risk by creating pressures for militaries to shortcut testing and rapidly deploy new AI-enabled systems. States could take a variety of options to mitigate the risks of creating unnecessary incentives to shortcut test and evaluation,[57] including publicly signaling the importance of T&E, increasing transparency about T&E processes, promoting international T&E standards, and sharing civilian research on AI safety.

Additionally, AI will enable more capable autonomous systems, and their increased use may pose stability risks, particularly when deployed into contested areas. To mitigate these risks, states could adopt CBMs such as “rules of the road” for the behavior of autonomous systems, marking systems to signal their degree of autonomy, and adhering to off-limits geographic areas for autonomous systems.

Public Signaling

To reduce AI accident risk, national security leaders could publicly emphasize the importance of strong T&E requirements for military AI applications. This potentially could be linked to a formal multilateral statement or something more informal. Publicly promoting AI T&E could be valuable in signaling that nations agree, at least in principle, about the importance of T&E to avoid unnecessary accidents and mishaps. Public statements would be more powerful when used in combination with major investments in T&E institutions and processes. Promoting AI T&E as a CBM would be designed to create positive spillover effects. As major countries investing in AI come together to promote AI safety, it demonstrates the importance of the issue. It could encourage other governments to sign on and signal that AI experts within the bureaucracy can advocate for AI T&E measures.

The downsides of publicly signaling the prioritization of AI T&E are relatively limited. A critic might argue that, to the extent that accidents are a necessary part of the innovation and capabilities development process, an overemphasis on T&E might discourage experimentation. However, promoting experimentation and innovation does not have to come at the expense of building robust and assured systems, especially since it is through experimentation and testing that accident risks are likely to be revealed, leading to the deployment of more capable systems. Ensuring that AI systems function as intended is part of fielding effective military capabilities, and effective T&E processes are aligned with the goal of fielding superior military capabilities. Rigorous T&E processes would, by definition, add time to the development process in order to ensure that systems are robust and secure before deployment, but the result would be more effective systems once deployed. In peacetime, taking additional measures to ensure that military systems will perform properly in wartime has little downside, so long as accident risk does not become a bureaucratic excuse for inaction. In wartime, the tradeoffs in delaying fielding may become more acute, and militaries may balance these risks differently. There are potential transparency downsides if countries say they emphasize AI T&E in public, but do not do so in private,[58] but that would not impose costs on countries whose actions match their rhetoric.

Increased Transparency about T&E Processes

A related unilateral or multilateral CBM could involve countries publicly releasing details about the T&E processes used for military applications of AI without revealing details about specific technical capabilities. This is similar to existing U.S. policy regarding legal weapons reviews. Currently, the U.S. military promotes norms in favor of stringent legal weapons reviews but does not share the actual reviews of specific weapons.[59]

Since this CBM would build on existing norms that the United States already promotes, transparency about T&E processes for military AI systems might be more likely to receive American support than more intrusive measures. Moreover, increasing knowledge about T&E processes might bring other countries that want to learn from the American military on board. The potential drawbacks of transparency surrounding T&E processes stem from what happens if the CBM succeeds. If successful, all countries, including potential adversaries, would have greater knowledge of how to design effective T&E processes for their military AI applications. This could improve their ability to field more effective military AI systems. This downside may be somewhat mitigated if a country only shares high-level information about its T&E bureaucratic processes and refrains from sharing technical information that could actually help an adversary execute more effective T&E. Nevertheless, an overarching concern with any T&E-related CBM that aims to reduce the risk to international stability from states building unsafe AI systems is that actually succeeding in improving other states' T&E could also lead to adversaries deploying more effective AI systems. Whether an adversary's improved AI capabilities or the prospect of an adversary deploying unsafe military AI systems is more of a danger to a country's security would need to be considered.

International Military AI T&E Standards

Another CBM regarding AI safety could entail establishing and promoting specific international standards for what constitutes effective T&E practices for military AI applications. Such an effort could build on private-sector and public-private standard-setting actions for non-military uses of AI.[60]

While not enforceable or verifiable, promoting common standards for AI T&E could be a useful focal point for like-minded states to promote responsible norms concerning the safe deployment of military AI systems. The downsides of promoting common T&E standards are similar to the potential downsides of a public emphasis on AI safety. These kinds of CBMs are early building blocks: While the gains are likely to be relatively limited, the downsides are limited as well, because they do not expose key information or require national commitments that limit capabilities. As with increasing transparency about T&E

processes, the most significant downside to effective T&E standards would be that, if successful, this CBM could increase the reliability of military AI systems by adversary states. The relative balance of danger between more reliable, and therefore more effective, adversary AI systems versus unreliable and more accident-prone AI systems would again need to be carefully weighed.

Shared Civilian Research on AI Safety

International efforts to promote shared civilian research on AI safety could be a low-level CBM that would not explicitly involve military action. Shared civilian research would build scientific cooperation between nations, which could serve as a building block for overall cooperation. Focusing cooperation on AI safety, an area of shared interest, might also make more nations willing to sign on to participate. An analogy to this in the U.S.-Soviet context is the Apollo-Soyuz mission in 1975, whose intent was to promote cooperation between civilian scientists on a shared agenda. Similarly, nations could work to foster increased cooperation and collaboration between civilian scientists on AI safety.

The potential drawbacks of cooperation stem from the general-purpose character of AI knowledge. If increasing cooperation on AI safety led to adversary breakthroughs in AI safety that made them better able to field effective military uses of AI, there could be negative consequences for other states' security. It may be possible to mitigate this downside by carefully scoping the shared civilian research, depending on the specific type of cooperation and degree of information-sharing required by participants.

International Autonomous Incidents Agreement

There are inherent risks when autonomous systems with any level of decision-making interact with adversary forces in contested areas. Given the brittleness of algorithms, the deployment of autonomous systems in a crisis situation could increase the risk of accidents and miscalculation. AI-related CBMs could build on Cold War agreements to reduce the risk of accidental escalation, with some modification to account for the new challenges AI-enabled autonomous systems present.

States have long used established "rules of the road" to govern the interaction of military forces operating with a high degree of autonomy, such as at naval vessels at sea, and there may be similar value in such a CBM for interactions with AI-enabled autonomous systems. The 1972 Incidents at Sea Agreement and older "rules of the road" such as maritime prize law provide useful historical examples for how nations have managed analogous challenges in the past. Building on these historical examples, states could adopt a modern-day "international autonomous incidents agreement" that focuses on military applications of autonomous systems, especially in the air and maritime environments. Such an agreement could help reduce risks from accidental escalation by autonomous systems, as well as reduce ambiguity about the extent of human intention behind the behavior of autonomous systems.

In addition to the Incidents at Sea Agreement, maritime prize law is another useful historical analogy for how states might craft a rule set for autonomous systems' interactions. Prize law, which first began in the 12th century and evolved more fully among European states in the 15th to 19th centuries, regulated how ships interacted during wartime. Because both warships and privateers, as a practical matter, operated with a high degree of autonomy while at sea, prize law consisted of a set of rules governing acceptable wartime behavior. Rules covered which ships could be attacked, ships' markings for identification, the use of force, seizure of cargo, and providing for the safety of ships' crews.[61]

Nations face an analogous challenge with autonomous systems as they become increasingly integrated into military forces. Autonomous systems will be operating on their own for some period of time, potentially interacting with assets from other nations, including competitors, and there could be value in establishing internationally agreed upon "rules of the road" for how such systems should interact. The

goal of such an agreement, which would not have to be as formal as the Incidents at Sea Agreement, would be to increase predictability and reduce ambiguity about the behavior of autonomous systems. Such an agreement could be legally binding but would not necessarily need to be in order to be useful. It would likely need to be codified in an agreement (or set of agreements), however, so that expectations are clear by all parties.

An ideal set of rules would be self-enforcing, such that it is against one's own interests to violate them. Examples of rules of this kind in warfare include prohibitions against perfidy[62] and giving "no quarter,"[63] violating either of which incentivizes the enemy to engage in counterproductive behavior, such as refusing to recognize surrender or fighting to the bitter end rather than surrendering.

An autonomous incidents agreement could also include provisions for information-sharing about potential deployments of autonomous systems in disputed areas and mechanisms for consultation at the military-to-military level to resolve questions that arise (including potentially a hotline to respond to incidents in real time).

One challenge with autonomous systems is that their autonomous programming is not immediately observable and inspectable from the outside, a major hurdle for verifying compliance with arms control. One benefit to an international rule set that governs the behavior of autonomous systems, particularly in peacetime or pre-conflict settings, is that the outward behavior of the system would be observable, even if its code is not. Other nations could see how another country's autonomous air, ground, or maritime drone behaves and whether it is complying with the rules, depending on how the rules are written.

Given the perceived success of the Incidents at Sea Agreement in decreasing the risk of accidental and inadvertent escalation between the United States and the Soviet Union, an equivalent agreement in the AI space might have potential to do the same for a new generation. The efficacy of any agreement would depend on the details, both in the agreement itself and in states' execution. For example, the United States and China have signed multiple CBM agreements involving air and maritime deconfliction of military forces, including the 1998 U.S.-China Military Maritime Consultative Agreement and the 2014 Memorandum of Understanding Regarding the Rules of Behavior for Safety of Air and Maritime Encounters, yet U.S.-China air and naval incidents have continued.[64]

However, the existence of these prior agreements themselves may be a positive sign about the potential for U.S.-China cooperation on preventing accidents and could be a building block for further collaboration. Moreover, in a February 2020 article, Senior Colonel Zhou Bo in China's People's Liberation Army (PLA) advocated for CBMs between the United States and China, including on military AI, drawing on the example of the 1972 Incidents at Sea Agreement.[65] Interest in at least some quarters in the Chinese military suggests that cooperation may be possible even in the midst of competition, especially if the PLA is willing to reciprocate American transparency.[66]

In the absence of an internationally agreed upon common rule set, a country could unilaterally make declaratory statements about the behavior of its autonomous systems. For example, a country could say, "If you fire at our autonomous ship/aircraft/vehicle, it will fire back defensively."[67] In principle, such a rule could incentivize the desired behavior by other nations (i.e., not shooting at the autonomous ship, unless you want to start a conflict). If every nation adopted this rule, coupled with a "shoot-second posture" for autonomous systems—they would not fire unless fired upon first—the result could be a mutually stable situation. A unilateral declaration of a set of rules for avoiding incidents would be analogous to declaring, "I will drive on the right side of the road. I suggest you do the same or we both will crash." This could work if countries' aim is to coordinate their behavior to avoid conflict, meaning they have some shared interests in avoiding accidental escalation.

One challenge to establishing rules of the road for autonomous systems' behavior would be if there were incentives to defect from the rules. For example, in World War I, technological developments enabled submarines, which were highly effective in attacking ships but unable to feasibly comply with existing prize law without putting themselves at risk of attack by surfacing. Despite attempts in the early 20th century to regulate submarines, the incentives for defecting from the existing rules were too great (and the rules failed to adapt), and the result was unrestricted submarine warfare.[68] Another challenge to a potential autonomous incidents agreement is fully exploring the incentives for trustworthiness, both in the signals that countries send about the behavior of their autonomous systems and adversaries' responses. Some declaratory policies would not be credible, such as the claim to have created a "dead hand" system such that if a country engaged in a particular type of action, an autonomous system would start a war and there would be nothing a leader could do to stop it.

Marking Autonomous Systems

One component of managing risks from interactions with autonomous systems might involve marking those systems to signal to adversaries their level of autonomy. This could be done through physical markings, such as paint, lights, flags, or other observable external characteristics, or through electronic means, such as radiofrequency broadcasts. One benefit of a marking system is that it builds on things militaries already do, even at the tactical level, to signal their intentions. For example, a fighter jet might tip its wing to show an adversary that it is carrying air-to-air missiles under the wing, communicating an unambiguous and credible signal about capability, and at least threatening some degree of intent. Because autonomous programming is not physically observable in the same way, militaries would have to intentionally design systems with observable markings reflecting their degree of autonomy. Another option could be that certain platforms are understood to have certain behavior (or not), the same way that conventional and nuclear capabilities may in some cases be segregated (e.g., some aircraft are nuclear-capable and some are not, which allows nations to send different kinds of signals).

Because potential markings for autonomous functionality are not forced by the capability itself but are rather an optional signal that militaries can choose to send, in order for such markings to be believable and useful, there would have to be strong incentives for sending truthful signals and few incentives for deception. This would be challenging, and nations would have to carefully think through what signals might be useful and believable in different circumstances, and how adversaries might interpret such signals. Additionally, because concepts such as "levels of autonomy" are often murky, especially for systems that have varying modes of operation, nations would have to think carefully about what kinds of signals could helpfully and clearly communicate autonomous functionality to other countries.[69] In the past, human operators of automated or autonomous systems have in some instances misunderstood the functionality of the system they themselves were operating, leading to accidents.[70] This problem would be significantly compounded for an external observer. Signals that were trusted but misunderstood could be equally or more dangerous than ambiguity, and states should strive for clear, unambiguous signals.

Off-limits Geographic Areas

Nations could agree to declare some geographic areas off-limits to autonomous systems because of their risk of unanticipated interactions. This could be to avoid unintended escalation in a contested region (e.g., a demilitarized zone) or because a region is near civilian objects (e.g., a commercial airliner flight path) and operating there poses a risk to civilians. Other examples of areas that nations could agree to make off-limits to autonomous military systems could be overlapping territorial claims or other countries' exclusive economic zones (EEZs) or airspace above their EEZs.

Center for a New American Security
1152 15th Street NW, Suite 950, Washington, DC 20005
T: 202.457.9400 | F: 202.457.9401 | CNAS.org | @CNASdc

Reaching agreement on specific regions could be challenging, however, since the areas most at risk of escalation are precisely the regions where nations disagree on territorial claims. Nations could perceive any agreement to refrain from deploying elements of military forces to a region as reflecting negatively on their territorial claims or freedom of navigation. Agreeing to declare some areas off-limits to autonomous systems is likely to be most constructive when there are already pre-established regions that countries agree are under dispute (even if they disagree on who has a claim to ownership) and where pre-existing military deconfliction measures already exist.

SPECIFIC MISSION-RELATED CBMS: NUCLEAR OPERATIONS

The integration of AI, autonomy, and/or automation into nuclear command-and-control, early warning, and delivery systems poses unique risks to international stability because of the extreme consequences of nuclear accidents or misuse.[71] One option for mitigating these risks could be for nations to set limits on the integration of AI, autonomy, or automation into their nuclear operations.

Some U.S. military leaders and official DoD documents have expressed skepticism about integrating uninhabited vehicles into plans surrounding nuclear weapons. The Air Force's 2013 *Remotely Piloted Aircraft (RPA) Vector* report proposed that nuclear strike "may not be technically feasible unless safeguards are developed and even then may not be considered for [unmanned aircraft systems] operations."[72] U.S. Air Force general officers have been publicly skeptical about having uninhabited vehicles armed with nuclear weapons. General Robin Rand stated in 2016, during his time as head of Air Force Global Strike Command, that: "We're planning on [the B-21] being manned. … I like the man in the loop … very much, particularly as we do the dual-capable mission with nuclear weapons."[73]

Other U.S. military leaders have publicly expressed support for limits on the integration of AI into nuclear command-and-control. In September 2019, Lieutenant General Jack Shanahan, head of the DoD Joint AI Center, said, "You will find no stronger proponent of the integration of AI capabilities writ large into the Department of Defense, but there is one area where I pause, and it has to do with nuclear command and control." In reaction to the concept of the United States adopting a "dead hand" system to automate nuclear retaliation if national leadership were wiped out, Shanahan said, "My immediate answer is '*No.* We do *not*.' … This is the ultimate human decision that needs to be made which is in the area of nuclear command and control."[74]

While the motivation for these statements about limits on the use of autonomy may or may not be strategic stability—bureaucratic factors could also be at play—they are examples of the kinds of limits that nuclear powers could agree to set, unilaterally or collectively, on the integration of AI, autonomy, and automation into their nuclear operations.

Nuclear states have a range of options for how to engage with these kinds of risks. On one end of the spectrum are arms control treaties with some degree of verification or transparency measures to ensure mutual trust in adherence to the agreements. On the other end of the spectrum are unilateral transparency measures, which could have varying degrees of concreteness ranging from informal statements from military or civilian leaders along the lines of Shanahan's and Rand's statements, all the way to formal declaratory policies. In between are options such as mutual transparency measures, statements of principles, or non-legally binding codes of conduct or other agreements between nuclear states to ensure human control over nuclear weapons and nuclear launch decisions. Even if states that desired these restraints found themselves in a position where others were unwilling to adopt more binding commitments, there may be value in unilateral transparency measures both to reduce the fears of other states and to promulgate norms of responsible state behavior. As with other areas, it is important to consider incentives for defection from an agreement and the extent to which one state's voluntary limitations depend on verifying others' compliance with an agreement. If some states, such as the United

States, desire strict positive human control over their nuclear weapons and nuclear launch authority for their own reasons, then verifying others' behavior, while desirable, may not be a necessary precondition to those states adopting their own limits on the use of AI, autonomy, or automation in nuclear operations.

Two possible CBMs for AI applications in the nuclear arena involve nuclear weapons states agreeing to strict human control over nuclear launch decisions and ensuring any recoverable delivery vehicles are human-inhabited, to ensure positive human control.

Strict Human Control Over Nuclear Launch Decisions

One CBM for uses of AI in the nuclear arena would involve an agreement by nuclear powers to ensure positive human control over all nuclear launch decisions. This type of agreement would preclude automated "dead hand" systems or any other automatic trigger for the use of nuclear weapons.

The benefit of such a CBM would be to reduce the risk of accidental nuclear war. It would preclude a machine malfunction leading directly to the use of nuclear weapons without a human involved in the process. Agreement on positive human control over nuclear launch decisions could also be a mechanism for dialogue with newer nuclear powers, helping generate more transparency over their nuclear launch decisions.

A drawback to this CBM would be forgoing any potential benefits of an automated "dead hand" or similar system. While not without controversy, automated nuclear response systems have a strategic logic under some circumstances. Some nuclear states could desire automated retaliatory systems to ensure a second strike in a decapitation scenario. To the extent that strategic stability depends on second strike capabilities, and a country believes it faces a real risk of decapitation if a conflict escalates, that country might prefer an automated option. (This was the intent behind the Soviet Perimeter system, which reportedly had a semiautomated "dead hand" functionality.)[75] The assurance of automated retaliation could be valuable as a deterrent and/or to reduce the incentives for a nation's leaders to launch a strike under ambiguous warning, if they felt confident that a second strike was assured. An agreement to rule out the use of automated "dead hand" systems might increase the risk of first strike instability, because nations could have a larger incentive to strike first—or perhaps launch in response to a false alarm—before being decapitated.

Alternatively, countries that feel they need an automated nuclear response option might prefer to not sign a CBM or to sign and then cheat.[76] Fortunately, the "costs" of a counterpart cheating on this type of CBM are relatively minimal, since presumably most states would only sign such an agreement if they thought it was already consistent with their nuclear launch decision-making process.

Prohibitions on Uninhabited Nuclear Launch Platforms

An agreement to prohibit uninhabited nuclear launch platforms would involve nuclear weapon states agreeing to forgo a capability that, to our knowledge, no nuclear weapon state deploys today—an uninhabited ("unmanned") submarine, fighter, or bomber armed with nuclear weapons.[77] Such an agreement would not affect one-way nuclear delivery vehicles, such as missiles or bombs, instead only preventing a state from deploying two-way (recoverable) remotely piloted or autonomous platforms armed with a nuclear weapon. States have long employed uninhabited nuclear delivery vehicles (missiles, bombs, torpedoes) to carry a nuclear warhead to the target. At present, however, the recoverable launch platform (submarine, bomber, transporter erector launcher) is crewed. With crewed nuclear launch platforms, humans remain not only in control over the final decision to launch a nuclear weapon, but have direct physical access to the launch platform to maintain positive control over the nuclear launch decision.

A critical benefit of CBMs that sustain positive human control over nuclear weapons is a reduction in the risk of accidental nuclear war. Deploying nuclear weapons on an uninhabited launch platform, whether remotely piloted or autonomous, would by definition increase the risk that, in the case of an accident, whether mechanical or due to flawed software code, a machine, rather than a human, would make the decision about the use of nuclear weapons. Similarly, a crewed platform would have a redundant layer of direct onboard human physical control in the event that the system's software or communications links were hacked. As previously described, U.S. military leaders, often skeptical about capabilities of remotely piloted or autonomous systems, have expressed a degree of support for such a policy, even unilaterally. With American support, this type of CBM might have a better chance of succeeding and gathering support among other nuclear weapon states.

Critics might argue that, similar to the objection to a ban on automated nuclear launches, some types of nuclear states might view more autonomous platforms with nuclear weapons as critical to their second-strike capabilities because of their ability to stay in the air or concealed at sea for extended periods. Russian military officials have raised the idea of an uninhabited nuclear-armed bomber,[78] and Russia is reportedly developing a nuclear-armed uninhabited undersea vehicle, the Status-6.[79]
However, given that these platforms are not currently deployed, it may be easier to reach an agreement to prohibit these platforms compared with an agreement prohibiting a capability that already exists. Moreover, to the extent that this kind of CBM is more a commitment to avoid pursuing dangerous applications of AI, rather than a restriction on current capabilities, it would also be reversible if states decided such capabilities were both necessary and safe at a later time.[80]

Conclusion

Military use of AI poses several risks, including due to ways AI could change the character of warfare, the limitations of AI technology today, and the use of AI for specific military missions such as nuclear operations. Policymakers should be cognizant of these risks as nations begin to integrate AI into their military forces, and they should seek to mitigate these risks where possible. Because AI is a general-purpose technology, it is not reasonable to expect militaries to refrain from adopting AI overall, any more than militaries would refrain from adopting computers or electricity. *How* militaries adopt AI matters a great deal, however, and various approaches could mitigate risks stemming from military AI competition.

Confidence-building measures are one potential tool policymakers could use to help reduce the risks of military AI competition among states. There are a variety of potential confidence-building measures that could be used, all of which have different benefits and drawbacks. As scholars and policymakers move forward to better understand the risks of military AI competition, these and other confidence-building measures should be carefully considered, alongside other approaches such as traditional arms control.

Center for a New American Security
1152 15th Street NW, Suite 950, Washington, DC 20005
T: 202.457.9400 | F: 202.457.9401 | CNAS.org | @CNASdc

Appendix

DEPARTMENT OF DEFENSE (DOD) ARTIFICIAL INTELLIGENCE (AI) PRINCIPLES[81]

1. **Responsible.** DoD personnel will exercise appropriate levels of judgment and care, while remaining responsible for the development, deployment, and use of AI capabilities.

2. **Equitable.** The department will take deliberate steps to minimize unintended bias in AI capabilities.

3. **Traceable.** The department's AI capabilities will be developed and deployed such that relevant personnel possess an appropriate understanding of the technology, development processes, and operational methods applicable to AI capabilities, including with transparent and auditable methodologies, data sources, and design procedure and documentation.

4. **Reliable.** The department's AI capabilities will have explicit, well-defined uses, and the safety, security, and effectiveness of such capabilities will be subject to testing and assurance within those defined uses across their entire life cycles.

5. **Governable.** The department will design and engineer AI capabilities to fulfill their intended functions while possessing the ability to detect and avoid unintended consequences, and the ability to disengage or deactivate deployed systems that demonstrate unintended behavior.

Center for a New American Security
1152 15th Street NW, Suite 950, Washington, DC 20005
T: 202.457.9400 | F: 202.457.9401 | CNAS.org | @CNASdc

1. Tim Dutton, "An Overview of National AI Strategies," Politics + AI on Medium.com, June 28, 2018, https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd.

2. Andrew Imbrie, James Dunham, Rebecca Gelles, and Catherine Aiken, "Mainframes: A Provisional Analysis of Rhetorical Frames in AI," CSET Issue Brief (Center for Security and Emerging Technology, August 2020), https://cset.georgetown.edu/research/mainframes-a-provisional-analysis-of-rhetorical-frames-in-ai/; Heather M. Roff, "The frame problem: The AI 'arms race' isn't one," *Bulletin of the Atomic Scientists* (April 29, 2019), https://thebulletin.org/2019/04/the-frame-problem-the-ai-arms-race-isnt-one/; "Autonomous Weapons: An Open Letter from AI & Robotics Researchers," Future of Life Institute, July 28, 2015, https://futureoflife.org/open-letter-autonomous-weapons/?cn-reloaded=1; and Brandon Knapp, "DoD official: US not part of AI arms race," *C4ISRNET* (April 10, 2018), https://www.c4isrnet.com/it-networks/2018/04/10/dod-official-us-not-part-of-ai-arms-race/.

3. Giacomo P. Paoli, Kerstin Vignard, David Danks, and Paul Meyer, "Modernizing Arms Control: Exploring responses to the use of AI in military decision-making" (United Nations Institute for Disarmament Research, 2020), https://www.unidir.org/publication/modernizing-arms-control; Andrew Imbrie and Elsa B. Kania, "AI Safety, Security, and Stability Among Great Powers: Options, Challenges, and Lessons Learned for Pragmatic Engagement," CSET Policy Brief (Center for Security and Emerging Technology, December 2019), https://cset.georgetown.edu/research/ai-safety-security-and-stability-among-great-powers-options-challenges-and-lessons-learned-for-pragmatic-engagement/; and Michael C. Horowitz, Lauren Kahn, and Casey Mahoney, "The Future of Military Applications of Artificial Intelligence: A Role for Confidence-Building Measures?," *Orbis*, 64 no. 4 (Fall 2020), 528-543.

4. Exploring or even adopting CBMs does not preclude other approaches to managing the risks of military AI competition, such as arms control for military AI applications. Arms control approaches are beyond the scope of this paper, however.

5. "Autonomous Weapons: An Open Letter from AI & Robotics Researchers."

6. William H. McNeill, *The Pursuit of Power: Technology, Armed Force, and Society since A.D. 1000* (Chicago: University of Chicago Press, 1984).

7. Chen Hanghui [陈航辉], "Artificial Intelligence: Disruptively Changing the Rules of the Game" [人工智能 ：颠覆性改变"游戏规则], China Military Online, March 18, 2016, http://www.81.cn/jskj/2016-03/18/content_6966873_2.htm. Chen Hanghui is affiliated with the Nanjing Army Command College.

8. John R. Allen and Amir Husain, "On Hyperwar," *Proceedings,* 143 no. 7 (July 2017), https://www.usni.org/magazines/proceedings/2017/july/hyperwar.

9. Robert O. Work, "Ending Keynote: Art, Narrative and the Third Offset" (Atlantic Council Global Strategy Forum, Washington, May 2, 2016), https://www.atlanticcouncil.org/unused/webcasts/2016-global-strategy-forum/.

10. Frank Hoffman, "Squaring Clausewitz's Trinity in the Age of Autonomous Weapons," *Orbis*, 63 no. 1 (Winter 2019), 44-63, https://doi.org/10.1016/j.orbis.2018.12.011; and Paul Scharre, "White Walkers and the Nature of War," in *Winning Westeros: How Game of Thrones Explains Modern Military Conflict*, eds. Max Brooks, John Amble, et al. (Lincoln, NE: Potomac Books, 2019), 253-264, https://www.amazon.com/Winning-Westeros-Explains-Military-Conflict/dp/1640122214.

11. Kenneth Payne, *Strategy, Evolution, and War: From Apes to Artificial Intelligence* (Washington: Georgetown University Press, 2018).

12. Commander Colin 'Farva' Price, "Navy F/A-18 Squadron Commander's Take On AI Repeatedly Beating Real Pilot In Dogfight," TheDrive.com, August 24, 2020, https://www.thedrive.com/the-war-zone/35947/navy-f-a-18-squadron-commanders-take-on-ai-repeatedly-beating-real-pilot-in-dogfight. To be fair, the AI in this case had an advantage because it was given full situational awareness about the local environment.

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

14. Michael C. Horowitz, "Artificial Intelligence, International Competition, and the Balance of Power," *Texas National Security Review*, 1 no. 3 (May 2018), 37-57, https://tnsr.org/2018/05/artificial-intelligence-international-competition-and-the-balance-of-power/.

15. Robert Jervis, "Cooperation Under the Security Dilemma," *World Politics*, 30 no. 2 (January 1978), 167-214.

16. Colin S. Gray, *House of Cards: Why Arms Control Must Fail* (Ithaca, NY: Cornell University Press, 1992); and Paul Scharre, "Lethal Autonomous Weapons and Policy-Making Amid Disruptive Technological Change," JustSecurity.org, November 14, 2017, https://www.justsecurity.org/47082/lethal-autonomous-weapons-policy-making-disruptive-technological-change/.

17. Jon Harper, "$98 billion Expected for Military Drone Market," *National Defense Magazine* (January 6, 2020), https://www.nationaldefensemagazine.org/articles/2020/1/6/98-billion-expected-for-military-drone-market.

18. Richard Danzig, "Technology Roulette: Managing Loss of Control as Many Militaries Pursue Technological Superiority" (Center for a New American Security, May 30, 2018), https://www.cnas.org/publications/reports/technology-roulette.

19. Paoli, Vignard, Danks, and Meyer, "Modernizing Arms Control: Exploring responses to the use of AI in military decision-making."

[20] Danielle C. Tarraf et al., "The Department of Defense Posture for Artificial Intelligence" (RAND Corporation, 2019), https://www.rand.org/pubs/research_reports/RR4229.html; "Perspectives on Research in Artificial Intelligence and Artificial General Intelligence Relevant to DoD," JSR-16-Task-003 (JASON Program, January 2017), https://fas.org/irp/agency/dod/jason/ai-dod.pdf; Dario Amodei et al., "Concrete Problems in AI Safety," arXiv preprint arXiv:1606.06565 (2016), https://arxiv.org/abs/1606.06565; Ram Shankar Siva Kumar et al., "Failure Modes in Machine Learning Systems," arXiv preprint, arXiv:1911.11034 (2019), https://arxiv.org/abs/1911.11034; and Michèle A. Flournoy, Avril Haines, and Gabrielle Chefitz, "Building Trust through Testing" (WestExec Advisors, October 2020), https://cset.georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf.

[21] Andrew Lohn, "Estimating the Brittleness of AI: Safety Integrity Levels and the Need for Testing Out-of-Distribution Performance," September 3, 2020, https://arxiv.org/pdf/2009.00802.pdf.

[22] U.S. Department of Defense, *Summary of the 2018 Department of Defense Artificial Intelligence Strategy: Harnessing AI to Advance Our Security and Prosperity* (2019), https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF.

[23] Charles Duhigg, "Did Uber Steal Google's Intellectual Property?," *The New Yorker* (October 22, 2018), https://www.newyorker.com/magazine/2018/10/22/did-uber-steal-googles-intellectual-property; and David Gelles, Natalie Kitroeff, Jack Nicas, and Rebecca R. Ruiz, "Boeing Was 'Go, Go, Go' to Beat Airbus With the 737 Max," *The New York Times*, March 23, 2019, https://www.nytimes.com/2019/03/23/business/boeing-737-max-crash.html.

[24] Charles E. Heller, "Chemical Warfare in World War I: The American Experience, 1917-1918," Leavenworth Papers, No. 10 (Combat Studies Institute, U.S. Army Command and General Staff College, September 1984), 6, https://apps.dtic.mil/dtic/tr/fulltext/u2/a189331.pdf.

[25] Manfred Griehl, *Luftwaffe X-Planes: German Experimental Aircraft of World War II* (London: Greenhill Books, 2004).

[26] This dynamic is akin to Thomas Schelling's "gunslinger" metaphor about the risks of first strike instability but pertaining to a race to field systems. Thomas Schelling, *The Strategy of Conflict* (Cambridge, MA: Harvard University Press, 1980).

[27] Gary Marcus and Ernest Davis, "GPT-3, Bloviator: OpenAI's Language Generator Has No Idea What It's Talking About," *MIT Technology Review* (August 22, 2020), https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/.

[28] Paul McLeary, "US: Iran Shoots Down Global Hawk; Second Drone Down This Month," BreakingDefense.com, June 20, 2019, https://breakingdefense.com/2019/06/us-iran-shoots-down-global-hawk-second-drone-down-this-month/; Julian Borger, "US shoots down second Iran-made armed drone over Syria in 12 days," *The Guardian*, June 20, 2017, https://www.theguardian.com/us-news/2017/jun/20/us-iran-drone-shot-down-syria; Terri Moon Cronk, "Chinese Seize U.S. Navy Underwater Drone in South China Sea," U.S. Department of Defense, December 16, 2016, https://www.defense.gov/Explore/News/Article/Article/1032823/chinese-seize-us-navy-underwater-drone-in-south-china-sea/; Brian Everstine, "Air Force: Lost Predator was shot down in Syria," *Air Force Times*, June 29, 2015, https://www.airforcetimes.com/news/pentagon-congress/2015/06/29/air-force-lost-predator-was-shot-down-in-syria/; and Dan Gettinger, "'An Act of War': Drones Are Testing China-Japan Relations," The Center for the Study of the Drone at Bard College, November 8, 2013, https://dronecenter.bard.edu/act-war-drones-testing-china-japan-relations/.

[29] Soviet Lieutenant Colonel Stanislav Petrov's decision to ignore a false alarm of a U.S. nuclear attack in 1983 is one of the more famous examples of the value of human decisionmaking in crisis situations.

[30] Paul Scharre, *Army of None: Autonomous Weapons and the Future of War* (New York: W.W. Norton & Co., 2018).

[31] For more on these risks, see Paul Scharre, "Autonomous Weapons and Stability" (doctoral thesis, King's College London, 2020), https://kclpure.kcl.ac.uk/portal/files/129451536/2020_Scharre_Paul_1575997_ethesis.pdf.

[32] U.S. Commodity Futures Trading Commission and U.S. Securities and Exchange Commission, *Findings Regarding the Market Events of May 6, 2010* (September 30, 2010), 2, http://www.sec.gov/news/studies/2010/marketevents-report.pdf; and Maureen Farrell, "Mini flash crashes: A dozen a day," CNNMoney, March 20, 2013, http://money.cnn.com/2013/03/20/investing/mini-flash-crash/index.html.

[33] Michael Eisen, "Amazon's $23,698,655.93 book about flies," it is NOT junk blog on WordPress.com, April 22, 2011, https://notjunk.wordpress.com/2011/04/22/amazons-23698655-93-book-about-flies/.

[34] This analysis builds on work done over the past few years by these scholars on different aspects of these risks, and in many cases there are more fulsome treatments of these specific challenges in other published research. Vincent Boulanin, Lora Saalman, Petr Topychkanov, Fei Su, and Moa Peldán Carlsson, "Artificial Intelligence, Strategic Stability and Nuclear Risk" (Stockholm International Peace Research Institute, June 2020), https://www.sipri.org/publications/2020/other-publications/artificial-intelligence-strategic-stability-and-nuclear-risk; T4GS, "AI and the Military: Forever Altering Strategic Stability" (T4GS Reports, February 13, 2019), https://securityandtechnology.org/wp-content/uploads/2020/07/ai_and_the_military_forever_altering_strategic_stability__IST_research_paper.pdf; Forrest E. Morgan, Benjamin Boudreaux, Andrew J. Lohn, Mark Ashby, Christian Curriden, Kelly Klima, and Derek Grossman, "Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World" (RAND Corporation, 2020), https://www.rand.org/pubs/research_reports/RR3139-1.html; Michael C. Horowitz, Paul Scharre, and Alexander Velez-Green, "A Stable Nuclear Future? The Impact of Autonomous Systems and Artificial Intelligence," 2019, https://arxiv.org/abs/1912.05291; and Edward Geist and Andrew J. Lohn, "How Might Artificial Intelligence Affect the Risk of Nuclear War?" (RAND Corporation, 2018), https://www.rand.org/pubs/perspectives/PE296.html.

[35] Boulanin, Saalman, Topychkanov, Su, and Peldán Carlsson, "Artificial Intelligence, Strategic Stability and Nuclear Risk"; T4GS, "AI and the Military: Forever Altering Strategic Stability"; Morgan, Boudreaux, Lohn, Ashby, Curriden, Klima, and Grossman, "Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World"; Horowitz, Scharre, and Velez-Green, "A Stable Nuclear Future? The Impact of

Autonomous Systems and Artificial Intelligence"; Michael C. Horowitz et al., "Policy Roundtable: Artificial Intelligence and International Security," Texas National Security Review, June 2, 2020, https://tnsr.org/roundtable/policy-roundtable-artificial-intelligence-and-international-security/; Melanie Sisson, Jennifer Spindel, Paul Scharre, and Vadim Kozyulin, "The Militarization of Artificial Intelligence," Stanley Center for Peace and Security, June 2020, https://stanleycenter.org/publications/militarization-of-artificial-intelligence/; Imbrie and Kania, "AI Safety, Security, and Stability Among Great Powers: Options, Challenges, and Lessons Learned for Pragmatic Engagement"; and Geist and Lohn, "How Might Artificial Intelligence Affect the Risk of Nuclear War?"

36. Ben Buchanan, "A National Security Research Agenda for Cybersecurity and Artificial Intelligence," CSET Issue Brief (Center for Security and Emerging Technology, May 2020), https://cset.georgetown.edu/research/a-national-security-research-agenda-for-cybersecurity-and-artificial-intelligence/.

37. Scharre, *Army of None: Autonomous Weapons and the Future of War*.

38. Michael C. Horowitz and Lauren Kahn, "The AI Literacy Gap Hobbling American Officialdom," War on the Rocks, January 14, 2020, https://warontherocks.com/2020/01/the-ai-literacy-gap-hobbling-american-officialdom/.

39. On building trust and confidence in AI, see Horowitz, Kahn, and Mahoney, "The Future of Military Applications of Artificial Intelligence: A Role for Confidence-Building Measures?," 528-543; and Imbrie and Kania, "AI Safety, Security, and Stability Among Great Powers: Options, Challenges, and Lessons Learned for Pragmatic Engagement."

40. Paoli, Vignard, Danks, and Meyer, "Modernizing Arms Control: Exploring responses to the use of AI in military decision-making."

41. The list below is adapted from Marie-France Desjardins, *Rethinking Confidence-Building Measures* (New York: Routledge, 1997), 5.

42. There is an arms control dilemma whereby the more useful that capabilities are perceived to be for fighting and winning wars, the harder it can become to craft effective regulation. This could create challenges for effective arms control surrounding military applications of AI, as it frequently has for other technologies such as submarines, air-delivered weapons, or nuclear weapons.

43. U.S. Department of State, Bureau of International Security and Nonproliferation, *Memorandum of Understanding Between the United States of America and the Union of Soviet Socialist Republics Regarding the Establishment of a Direct Communications Link*, entered into force June 20, 1963, https://2009-2017.state.gov/t/isn/4785.htm.

44. Sean M. Lynn-Jones, "A Quiet Success for Arms Control: Preventing Incidents at Sea," *International Security*, 9 no. 4 (1985), 154-84, doi:10.2307/2538545.

45. Desjardins, *Rethinking Confidence-Building Measures*.

46. Ralph A. Cossa, "Security Implications of Conflict in the South China Sea: Exploring Potential Triggers of Conflict" (Pacific Forum CSIS, 1998), http://www.southchinasea.org/files/2012/03/Cossa-Security-Implications-of-Conflict-in-the-S.ChinaSea.pdf.

47. Russell Goldman, "India-China Border Dispute: A Conflict Explained," *The New York Times*, June 17, 2020, https://www.nytimes.com/2020/06/17/world/asia/india-china-border-clashes.html; and Jeffrey Gettleman, "Shots Fired Along India-China Border for First Time in Years," *The New York Times*, September 8, 2020, https://www.nytimes.com/2020/09/08/world/asia/india-china-border.html.

48. Christian Ruhl, Duncan Hollis, Wyatt Hoffman, and Tim Maurer, "Cyberspace and Geopolitics: Assessing Global Cybersecurity Norm Processes at a Crossroads" (Carnegie Endowment for International Peace, February 26, 2020), https://carnegieendowment.org/2020/02/26/cyberspace-and-geopolitics-assessing-global-cybersecurity-norm-processes-at-crossroads-pub-81110.

49. Horowitz, Kahn, and Mahoney, "The Future of Military Applications of Artificial Intelligence: A Role for Confidence-Building Measures?"

50. Cossa, "Security Implications of Conflict in the South China Sea: Exploring Potential Triggers of Conflict."

51. Defense Innovation Board, *AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense* (2019), https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF.

52. U.S. Department of Defense, *Summary of the 2018 Department of Defense Artificial Intelligence Strategy: Harnessing AI to Advance Our Security and Prosperity*, 8.

53. U.S. Department of Defense, *Summary of the 2018 Department of Defense Artificial Intelligence Strategy: Harnessing AI to Advance Our Security and Prosperity,* 15.

54. Imbrie and Kania, "AI Safety, Security, and Stability Among Great Powers: Options, Challenges, and Lessons Learned for Pragmatic Engagement."

55. Paoli, Vignard, Danks, and Meyer, "Modernizing Arms Control: Exploring responses to the use of AI in military decision-making."

56. Beth Simmons, "Treaty Compliance and Violation," *Annual Review of Political Science*, 13 (2010), 274-296, 10.1146/annurev.polisci.12.040907.132713; Jana von Stein, "Making Promises, Keeping Promises: Democracy, Ratification and Compliance in International Human Rights Law," *British Journal of Political Science*, 46 no. 3 (July 2016), 655-679, https://www.cambridge.org/core/journals/british-journal-of-political-science/article/abs/making-promises-keeping-promises-democracy-

ratification-and-compliance-in-international-human-rights-law/D34262A0BD1EDC7451B4506E3A718A51; and Andrew Moravcsik, "The Origins of Human Rights Regimes: Democratic Delegation in Postwar Europe," *International Organization*, 54 no. 2 (Spring 2000), 217-252, https://www.princeton.edu/~amoravcs/library/origins.pdf.

57. Sometimes expressed as test and evaluation, verification and validation (TEVV). Flournoy, Haines, and Chefitz, "Building Trust through Testing."

58. Desjardins, *Rethinking Confidence-Building Measures.*

59. United States Army, *Legal Review of Weapons and Weapon Systems*, Army Regulation 27-53 (2019), https://armypubs.army.mil/epubs/DR_pubs/DR_a/pdf/web/ARN8435_AR27-53_Final_Web.pdf.

60. For example, see OECD Principles on AI, May 2019, https://www.oecd.org/going-digital/ai/principles/.

61. James Kraska, "Prize Law," Max Planck Encyclopedia of Public International Law, 2011, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1876724.

62. "Rule 65. Perfidy," Customary IHL - ICRC, https://ihl-databases.icrc.org/customary-ihl/eng/docs/v1_rul_rule65.

63. "Rule 46. Orders or Threats that No Quarter Will Be Given," Customary IHL - ICRC, https://ihl-databases.icrc.org/customary-ihl/eng/docs/v1_rul_rule46.

64. U.S. Department of State, *Maritime Matters: Military Safety – Agreement Between the United States of America and the People's Republic of China*, (January 19, 1998), https://www.state.gov/wp-content/uploads/2019/02/12924-China-Maritime-Matters-Misc-Agreement-1.19.1998.pdf; David Griffiths, "U.S.-China Maritime Confidence Building: Paradigms, Precedents, and Prospects," Study No. 6 (U.S. Naval War College's China Maritime Studies Institute, July 2010); *Memorandum of Understanding Between the Department of Defense of the United States of America and the Ministry of National Defense of the People's Republic of China Regarding the Rules of Behavior for Safety of Air and Maritime Encounters* (November 2014), https://archive.defense.gov/pubs/141112_MemorandumOfUnderstandingRegardingRules.pdf; and Yeganeh Torbati, "Despite agreements, risks linger of U.S.-China naval mishaps," Reuters, October 30, 2015, https://www.reuters.com/article/us-southchinasea-usa-communications/despite-agreements-risks-linger-of-u-s-china-naval-mishaps-idUSKCN0SO0E220151030. Both nations are also signatories to the 2014 multinational Code for Unplanned Encounters at Sea. "Document: Code for Unplanned Encounters at Sea," USNI News, August 22, 2016, https://news.usni.org/2014/06/17/document-conduct-unplanned-encounters-sea.

65. Zhou Bo, "China and America Can Compete and Coexist," *The New York Times,* February 3, 2020, https://www.nytimes.com/2020/02/03/opinion/pla-us-china-cold-war-military-sea.html.

66. This would hedge against the risk of transparency CBMs being abused as outlined in Desjardins, *Rethinking Confidence-Building Measures*.

67. Nations, of course, can and do often delegate self-defense authority to crewed ship/aircraft/vehicle commanders. What is different in those circumstances is that human control is still retained over use-of-force decisions, albeit delegated to a lower-level commander. Delegating use-of-force decisions to an autonomous system would raise novel stability concerns in crises or militarized disputes due to the brittle nature of machine decisionmaking and the inability of machines (at the current state of technology) to exercise judgment, understand context, or apply "common sense." For more on these risks, see Scharre, "Autonomous Weapons and Stability."

68. Jon L. Jacobson, "The Law of Submarine Warfare Today," *International Law Studies,* 64 (1991), 207-208, https://digital-commons.usnwc.edu/cgi/viewcontent.cgi?article=1756&context=ils; and Howard Levie, "Submarine Warfare: With Emphasis on the 1936 London Protocol," *International Law Studies,* 70 (1998), https://digital-commons.usnwc.edu/ils/vol70/iss1/12/.

69. Many thanks to Helen Toner for raising this point.

70. William Langewiesche, "The Human Factor," *Vanity Fair* (October 2014), http://www.vanityfair.com/news/business/2014/10/air-france-flight-447-crash; and "Final Report: On the accident of 1st June 2009 to the Airbus A330-203 registered F-GZCP operated by Air France flight 447 Rio de Janeiro – Paris" (Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile, [English translation], 2012), http://www.bea.aero/docspa/2009/f-cp090601.en/pdf/f-cp090601.en.pdf. This also may have been a factor in the U.S. Army's shoot-down of a Navy F-18 aircraft in 2003 with the Patriot air and missile defense system; and Scharre, "Autonomous Weapons and Stability," 185.

71. Horowitz, Scharre, and Velez-Green, "A Stable Nuclear Future? The Impact of Autonomous Systems and Artificial Intelligence."
72. Headquarters, U.S. Air Force, *RPA Vector: Vision and Enabling Concepts, 2013-2038* (February 17, 2014), 54, http://www.globalsecurity.org/military/library/policy/usaf/usaf-rpa-vector_vision-enabling-concepts_2013-2038.pdf.

73. Hope Hodge Seck, "Air Force Wants to Keep 'Man in the Loop' with B-21 Raider," DefenceTech.org, September 19, 2016, http://www.defensetech.org/2016/09/19/air-force-wants-to-keep-man-in-the-loop-with-b-21-raider/.

74. Sydney Freedberg Jr., "No AI for Nuclear Command & Control: JAIC's Shanahan," BreakingDefense.com, September 25, 2019, https://breakingdefense.com/2019/09/no-ai-for-nuclear-command-control-jaics-shanahan/.

75. The Soviet Perimeter system was reportedly a semiautomated "dead hand" system to ensure nuclear retaliation if Soviet leadership was wiped out in a decapitation attack. According to reports, the system would retain human control over launch decisions in the hands of a relatively junior Soviet officer who would retain final launch authority. The semiautomated nature of the system would bypass the senior-level approvals normally required to authorize a nuclear launch. The system's functionality and the extent to which it was built and deployed

remain disputed. Nicholas Thompson, "Inside the Apocalyptic Soviet Doomsday Machine," *WIRED* (September 21, 2009), https://www.wired.com/2009/09/mf-deadhand/; Vitalii Leonidovich Kataev, interviewed by Ellis Mishulovich, May 1993, http://nsarchive.gwu.edu/nukevault/ebb285/vol%20II%20Kataev.PDF; Varfolomei Vladimirovich Korobushin, interviewed by John G. Hines, December 10, 1992, http://nsarchive.gwu.edu/nukevault/ebb285/vol%20II%20Korobushin.PDF; Andrian A. Danilevich, interview by John G. Hines, March 5, 1990, 62-63, http://nsarchive.gwu.edu/nukevault/ebb285/vol%20iI%20Danilevich.pdf; and Viktor M. Surikov, interview by John G. Hines, September 11, 1993, 134-135, http://nsarchive.gwu.edu/nukevault/ebb285/vol%20II%20Surikov.PDF.

[76.] Desjardins, *Rethinking Confidence-Building Measures*.

[77.] Details surrounding the reported Russian Status-6 "Poseidon" uninhabited underwater vehicle are murky. Depending on its design and intended use, if deployed it could constitute an uninhabited nuclear launch platform of the kind we describe. H.I. Sutton, "Poseidon Torpedo," Covert Shores, February 22, 2019, http://www.hisutton.com/Poseidon_Torpedo.html; Russian President Vladimir Putin, "Presidential Address to the Federal Assembly," (Manezh Central Exhibition Hall, Moscow, March 1, 2018), http://en.kremlin.ru/events/president/news/56957; H.I. Sutton, "Countering Russian Poseidon Torpedo," Covert Shores, August 15, 2018, http://www.hisutton.com/Countering_Russian_Poseidon_Torpedo.html; U.S. Department of Defense, *Nuclear Posture Review 2018*, 8-9; and Michael Kofman, "Emerging Russian Weapons: Welcome to the 2020s (Part 2 – 9M730?, Status-6, Klavesin-2R)," Russia Military Analysis blog at WordPress.com, March 6, 2018, https://russianmilitaryanalysis.wordpress.com/2018/03/06/emerging-russian-weapons-welcome-to-the-2020s-part-2-9m730-status-6-klavesin-2r/.

[78.] "Russia Could Deploy Unmanned Bomber After 2040 - Air Force," RIA Novosti, August 2, 2012, http://www.globalsecurity.org/wmd/library/news/russia/2012/russia-120802-rianovosti01.htm.

[79.] Putin, "Presidential Address to the Federal Assembly"; U.S. Department of Defense, *Nuclear Posture Review 2018*, 8-9; Sutton, "Poseidon Torpedo"; Sutton, "Countering Russian Poseidon Torpedo"; and Kofman, "Emerging Russian Weapons: Welcome to the 2020s (Part 2 – 9M730?, Status-6, Klavesin-2R)."

[80.] The Intermediate-Range Nuclear Forces Treaty and Anti-Ballistic Missile Treaty are both examples of treaties adopted during the Cold War to address strategic stability risks but revisited after the Cold War era when the strategic environment had changed.

[81.] Defense Innovation Board, *AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense*.