# TEVV AFF – MNDI HPA

## Introductory Page

#### Thesis of this case:

AI is inevitable. It is a normal part of modernizing our military. Because it is a dual use technology, it will diffuse much more easily, which means there is not a huge military advantage to be the First to develop it. So, we are not in an “Arms Race.”

However, we – along with China and Russia - *believe* that we are in an arms race. We believe we have to develop it first because if China develops it first, we will be at a huge disadvantage. Not true, but that is what we believe. China believes the same thing about us. This is what is called a “Security Dilemma” – our pursuit of our security undermines another nation’s security, and their reaction is to pursue their security in a way that undermines ours, without a natural stopping point. We “create” a race by believing in it, which is a self-fulfilling prophecy.

So, what is the harm of believing that we are in an arms race? This Arms Race Framing or Arms Race Mentality causes us to deploy AI weapons as fast as we can, which causes us to rush to deploy them before they are tested, evaluated, verified and validated (TEVV). Before they are safe and reliable. If we rush, we cut corners, and put out “brittle” AI, which means AI that collapses in the real world when the real world is different from their testing environment. It’s like when EA pushed out Battlefront II way too early to try to beat other game publishers to the market, and it was full of bugs and malfunctions. Except this time, “bugs” in military AI that controls our nuclear weapons could malfunction in ways that cause nuclear war.

So, the plan enhances testing, evaluation, verification and validation (TEVV) processes with NATO for all our AI weapons. This is also called Assurance. It adopts “life cycle” testing, which just means that you continuously test the AI over its time in use, because it evolves over time, due to machine learning. This will make sure that the weapons we put out work as designed, and don’t break down at inopportune times. There is some pretty good evidence that a large commitment to AI TEVV can build confidence globally, reducing the pressure to “race” to deploy weapons.

The advantages are just the different types of AI systems that could break down due to bugs, malfunctions or premature deployment. Nuclear Systems, Cyber Systems, Brittle Systems, etc. There is an Operator Trust scenario, which just says that soldiers won’t Trust their AI weapons if they are not safe and reliable, and that if soldiers and commanders do not Trust their weapons or systems, that breaks down the entire military readiness of NATO – blah blah blah – Russia/Ukraine/escalation something something nuclear war. You can choose one, two, or three of these scenarios to run. Probably not four, but hey – you do you.

This case is well positioned against security kritiks, because it addresses IR from a largely constructivist viewpoint explicitly in the 1AC. It is also well positioned to kritik the impacts to disads that rely on “You lose the Arms Race” arguments as being threat construction.

#### Credits:

There are some Very good cards in here, and they were found and cut by Mira A, Mae D, Lucy D, Austin D, Cali F, Shourya H, Balin K, Tucker M, Louis M-S, and Sofia T, and Josh Z. And they did it with me being mostly on the DL, so any mistakes are mine, while all credit goes to them. Good job, excellent self-leadership, and congratulations on a job well done. I don’t know how to put in a “Thumbs Up” emote-i-con, so just pretend that there is one here!

## 1AC

### 1AC - Contention One

#### We are Not in an AI Arms Race. We think that we are, but those claims are overblown – the AI adoption process is routine continuation of modernization

Horowitz and Scharre, 2021 - Director of the Emerging Capabilities Policy Office in the Office of the Under Secretary of Defense for Policy and  Vice President and Director of Studies at CNAS [Michael, Paul, January 12 2021, “AI and International Stability: Risks and Confidence- Building Measures”, <https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures>, Acc 6/18/22. M.A.]

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](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn2) 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](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn3) 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](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn4) 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.

#### Unfortunately, *Believing* you are in an Arms Race is also dangerous - it creates a Security Dilemma where both states are pressured to deploy AI as soon as they can - weapons are deployed before they are tested, safe, and reliable. This disrupts the balance of power and crisis stability.

Scharre 2021 – Director of the Technology and National Security Program at the Center for a New American Security [Paul, 6/28/21,https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/, “Debunking the AI Arms Race Theory”, 6/18/22, LND]

Even if military AI spending does not rise to the level of an “arms race,” many nations are nevertheless engaged in a security competition in the adoption of military AI, a competition that does pose risks. The situation that states find themselves in with regard to AI competition is much more accurately described as a security dilemma,[16](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/#_ftn16) a more generalized competitive dynamic between states than the more narrowly defined “arms race.” In his 1978 article, “Cooperation Under the Security Dilemma,” Robert Jervis defined the security dilemma as follows: “[M]any of the means by which a state tries to increase its security decrease the security of others.”[17](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/#_ftn17) As Charles Glaser has pointed out, it is not obvious from this definition why it would be intrinsically bad for an increase in one state’s security to come at the expense of another’s security.[18](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/#_ftn18) In fact, decreasing the security of other states could have beneficial effects in enhancing deterrence and reducing the risks of aggression or achieving a favorable balance of power in a region, which could lead to greater political influence. The problem comes in the second- and third-order effects that could develop when another state reacts to having its security reduced. Responses could include counterbalancing with a net effect of no change in security (or worsening security). Glaser argues that there are some situations in which security competition is a rational strategy for a state to pursue even if competitors will arm in response. In other situations, arming may be a suboptimal strategy for a state, which would be better served by restraint or pursuing arms control.[19](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/#_ftn19) Security competition could even leave both states worse off than before. This can occur during a traditional arms race if nations expend vast sums of money in an unsuccessful attempt to gain an advantage over one another, with the result that both nations divert funds from non-defense expenditures. If the outcome of a security competition is the same relative military balance as before, the balance of power may not have meaningfully changed, but both nations could face diminished economic and social well-being at home relative to if they had avoided a security competition. Even absent this “guns vs. butter” tradeoff, however, there are other ways in which security competition can lead to a net negative outcome for both states. One way this could occur is if military innovation and the development of new capabilities alter the character of warfare in a manner that is more harmful, more destructive, less stable, or otherwise less desirable than before. In his 1997 article, “The Security Dilemma Revisited,” Glaser gave the example of military capabilities that shifted warfare to a more offense-dominant regime.[20](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/#_ftn20) There are other ways in which warfare could evolve in a net negative direction as well. For example, in World War I, Germany’s interest in developing and deploying chemical weapons was spurred in part due to fears about France’s developments in poison gas.[21](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/#_ftn21) The result was the introduction of a weapon that increased combatant suffering on both sides, without delivering a significant military advantage to either. The same could occur with AI: It could alter the character of warfare in a way that would be a net negative for all participants. One possibility for how AI could alter warfare in a manner that would leave all states worse off would be if it accelerated the tempo of war past the point of human control, making warfare faster, more violent, and less controllable. There are advantages to adding intelligence into machines, but given the limitations of AI systems today, the optimal model for achieving the highest quality decision-making would be a joint human-machine architecture that combines human and machine decision-making. One way in which machines outperform humans, however, is in speed. It is possible to envision a competitive dynamic in which countries feel compelled to automate increasing amounts of their military operations in order to keep pace with adversaries. Then-Deputy Secretary of Defense Robert O. Work summed up the dilemma when he asked, “If our competitors go to Terminators and we are still operating where the machines are helping the humans and it turns out the Terminators are able to make decisions faster, even if they’re bad, how would we respond?”[22](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/#_ftn22) This is a classic security dilemma. One state’s pursuit of greater automation and faster reaction times undermines other states’ security and leads them to similarly pursue more automation just to keep up. If states fall victim to this trap, it could lead to all states being less secure, since the pursuit of greater automation would not merely be an evolution in weapons and countermeasures that simply leads to the creation of new weapons in the future warfare could shift to a qualitatively different regime in which humans have less control over lethal force as decisions become more automated and the accelerating tempo of operations pushes humans “out of the loop” of decision-making. Some Chinese scholars have hypothesized about a battlefield “singularity,” in which the pace of combat eclipses human decision-making.[23](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/#_ftn23) U.S. scholars have used the term “hyperwar” to refer to a similar scenario.[24](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/#_ftn24) While the speed of engagement necessitates automation in some limited areas today, such as immediate localized defense of ships, bases, and vehicles from rocket and missile attack, expanding this zone of machine control into broader areas of war would be a significant development. Less human control over warfare could lead to wars that are less controllable and that escalate more quickly or more widely than humans intend. Similarly, limiting escalation or terminating conflicts could be more challenging if the pace of operations on the battlefield exceeds human decision-making. Political leaders would have a command-and-control problem in which their military forces are operating “inside” (i.e., faster than) their own decision cycle. The net effect of the quite rational desire for nations to gain an edge in speed could lead to an outcome that is worse for all. Yet, competitive dynamics could nevertheless drive such a result. One state’s pursuit of greater automation and faster reaction times undermines other states’ security and leads them to similarly pursue more automation just to keep up. Financial markets provide an example of this dynamic in a non-military competitive environment. Automation introduced into financial markets, especially high-frequency trading in which trades are executed at super-human speeds in milliseconds, has contributed to unstable market conditions that can lead to “flash crashes,” in which prices rapidly and dramatically shift.[25](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/#_ftn25) Financial regulators have responded by employing “circuit breakers” that automatically halt trading for a pre-determined period of time if the price moves too quickly.[26](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/#_ftn26) Financial markets have the benefit of a regulator who can force cooperative measures on competitors to address suboptimal outcomes. Under conditions of anarchy in the international security environment, any such cooperation would have to come from state v s themselves.

#### The DOD AI Strategy underfunded reliability testing – it focused on ethics, but not safety.

Lewis 2019 - project lead for the DOD’s Joint Lessons Learned studies [Larry, “AI Safety: Charting out the High Road”, War on the Rocks Dec 19, https://warontherocks.com/2019/12/ai-safety-charting-out-the-high-road/ LMSi]

Promise 1: Addressing Safety Risks Unique to AI In its AI strategy, the Department of Defense made a promise to address the safety risks unique to AI technology. This is reflective of America’s long record of commitments to safety and adherence to international laws for armed conflict. For example, all military systems are subject to test and evaluation activities to ensure that they are reliable and safe, as well as legal reviews to ensure they are consistent with international humanitarian law (e.g., the Geneva Conventions). It is not surprising, therefore, that safety is prominent in the defense AI strategy. Though a commendable intention, the strategy has not yet resulted in significant institutional steps to promote safety with regard to AI. The U.S. military has been busy supporting the Defense Innovation Board’s development of AI ethics, with the Joint AI Center also emphasizing the critical role ethics plays in AI applications, yet the pursuit of safety — for example, avoiding civilian casualties, friendly fire, and inadvertent escalation — has not received the same sort of attention. I acknowledge a few steps are being taken towards promoting AI safety. For example, the Defense Advanced Research Projects Agency has a program working to develop explainable AI, to help address challenges with the use of machine learning as a black box technology. Explainability will enhance AI safety: for example, by being able to explain why an AI application does something amiss in testing or operations and to take corrective actions. But such steps, while important, do not make up a comprehensive approach to identify and then systematically address AI safety risks. To that end, our most recent report draws on a risk management approach to AI safety: identifying risks, analyzing them, and then suggesting concrete actions to begin addressing them.

### 1AC - Advantage: Brittle AI

**The Pressure to deploy AI Rapidly makes accidents inevitable – the technology is brittle**

**Arnold and Toner, 2021 – Center for Security and Emerging Threats** [Zachary and Helen, July, CSET Policy Brief. “AI Accidents: An Emerging Threat What Could Happen and What to Do” https://cset. georgetown.edu/wp-content/uploads/CSET-AI-Accidents-An-Emerging-Threat.pdf Acc 6/7/22 TA]

Despite these problems, AI systems are becoming integrated into the real world at a pace that is only expected to accelerate in the next decade.9 These systems may be fragile, but as companies, governments, and militaries decide when and how to deploy them, their huge potential benefits will often overshadow uncertain risks. Leaders in these organizations also may not be fully aware of these risks, and may face pressure from competitors willing to move quickly.10 To be sure, some industries are already deploying AI much faster than others, and a few sensitive sectors may remain “walled off” for some time.11 But eventually, the powerful incentives driving the spread of AI today are likely to make it pervasive. As our economy, security, and health become more and more dependent on AI systems, these systems’ fragilities will put lives at stake. Today, many are worried about AI being misused intentionally. An adversary could attack with swarms of drones; authoritarian governments are already using AI algorithms to discriminate on the basis of race or ideology. These risks are real, and they deserve attention. But unintended, accidental AI disasters are also an urgent concern. AI-related accidents are already making headlines, from inaccurate facial recognition systems causing false arrests to unexpected racial and gender discrimination by machine learning software.12 This is especially striking since AI has so far mostly been deployed in seemingly lower-stakes settings, such as newsfeed rankings, ad targeting, and speech recognition, with less deployment in higher stakes areas such as autonomous driving. Despite these initial accidents, governments, businesses, and militaries are preparing to use today’s flawed, fragile AI technologies in critical systems around the world. Future versions of AI technology may be less accident-prone, but there is no guarantee—and regardless, if rollout continues as expected, prior versions of the technology may already have been deployed at massive scale. The machine learning models of 2020 could easily still be in use decades in the future, just as airlines, stock exchanges, and federal agencies still rely today on COBOL, a programming language first deployed in 1960.13 In retrospect, even the most extreme technological accidents, from the Challenger disaster to the meltdown at Chernobyl, can seem both predictable and preventable.14 History is full of accidents that seem obvious in retrospect, but “no one could have seen coming” at the time. In other cases, known risks are brushed aside, or obvious fixes go unmade. Unless we act, there is no reason to think that the advent of AI will be any different. In fact, there are reasons to think AI could cause more accidents than other technologies that have caused high-profile disasters. Unlike the space shuttle or nuclear power plants, for example, AI will be pervasive throughout society, creating endless opportunities for things to go awry. What’s more, modern AI is so good at some tasks that even sophisticated users and developers can come to trust it implicitly.15 This degree of trust, placed in pervasive, fallible systems without any common sense, could have terrible consequences.

#### Premature deployment of AI weapons risks miscalculation and war - they will fail when they operate outside of their prescribed environments.

Goldfarb and Lindsay, 2022 – Chair in AI and health care at the Univ of Toronto, Prof of Cybersecurity at the Georgia Tech [Avi, Jon, 2/25/22, <https://doi.org/10.1162/isec_a_00425>, “Prediction and Judgement: Why Artificial Intelligence Increases the Importance of Humans in War” 6/18/22, LND]

In contrast with chess, the risks of premature automation are more extreme in the military realm (e.g., fratricide and civilian casualties), but the logic is the same. Militaries abound with standard operating procedures and tactical doctrines that guide the use of lethal capabilities (e.g., instructions for the safe operation of weapon platforms, playbooks for tactical maneuvers, and policies for employing weapons). To the degree that goals and mission parameters can be clearly specified, tactical operations will appear to be attractive candidates for automation. To the degree that combat timelines are expected to be extremely compressed, moreover, automation may appear to be even more urgent.[75](javascript:;) Rapid decision-making would necessitate the pre-specification of goals and payoffs and the coupling of AI prediction to robotic action. Lethal autonomous weapon systems use prediction to navigate complex environments in order to arrive at destinations or follow targets, within constraints that are supplied by human operators.[76](javascript:;) Their targeting systems base their predictions on training data that identify valid targets. Using algorithms, machines may rapidly and accurately identify targets at far greater distances than human visual recognition, and algorithmic target recognition may be collocated with sensors to reduce response times.[77](javascript:;) Many AI-enabled weapons already or imminently exist. The Israeli Harpy loitering munition can search for and automatically engage targets, and China has plans for similar “intelligentized” cruise missiles.[78](javascript:;) Russia is developing a variety of armed, unmanned vehicles capable of autonomous fire or unarmed mine clearance.[79](javascript:;) The United States has been exploring combat applications for AI in all warfighting domains. In the air, the “Loyal Wingman” program pairs an unmanned F-16 with a manned F-35 or F-22 to explore the feasibility of using humans to direct autonomous aircraft, such as the XQ-58A Valkyrie.[80](javascript:;) Air combat algorithms that can process sensor data and plan effective combat maneuvers in the span of milliseconds have already defeated human pilots in some simulators.[81](javascript:;) At sea, the U.S. Navy's LOCUST project explores the feasibility of launching swarms of expendable surface-to-air drones.[82](javascript:;) The Defense Advanced Research Projects Agency's (DARPA) Continuous Trail Unmanned Vessel program is designed to search for enemy missile submarines and automatically trail them for months at a time, reporting regularly on their locations.[83](javascript:;) On land, U.S. Marine “warbot companies” equipped with networks of small robots might provide distributed sensing and precision fire.[84](javascript:;) Automated counter-battery responses, which accurately retaliate against the origin of an attack, could give human commanders leeway to focus on second- and third-order decisions in the wake of an attack.[85](javascript:;) In the cyber domain, AI systems might autonomously learn from and counter cyberattacks as they evolve in real time, as suggested by the performance of the Mayhem system in DARPA's 2016 Cyber Grand Challenge.[86](javascript:;) AI could be especially useful for detecting new signals in the electromagnetic spectrum and reconfiguring electronic warfare systems to exploit or counter them.[87](javascript:;) Space satellites, meanwhile, have been automated from their inception, and space operations might further leverage AI to enhance surveillance and control. Much of the AI security literature is preoccupied with the risks posed by automated weapons to strategic stability and human security.[88](javascript:;) Risks of miscalculation will increase as the operational context deviates from the training data set in important or subtle ways. The risk of deviation increases with the complexity and competitiveness of the strategic environment, while the costs of miscalculation increase with the lethality of automated action. The machine tries to optimize a specific goal, but in the wrong context, doing so can lead to false positives. AI weapons may inadvertently either target innocent civilians or friendly forces or trigger hostile retaliation. In these cases, the AI would have the authority to kill but would not understand the ramifications. The risks are particularly stark in the nuclear arena. Nuclear war is the rarest of rare events—keeping it that way is the whole point of nuclear deterrence—so training data for AI systems is either nonexistent or synthetic (i.e., based on simulation).[89](javascript:;) Any tendency for AI systems to misperceive or miscalculate when confronted with uncertain or novel situations could have catastrophic consequences.[90](javascript:;) In short, autonomous weapon systems that combine prediction with action can quickly make tragic mistakes.

**Premature deployment of AI increases the risk of crisis escalation due to inflexibility and automation bias**

**Horowitz and Kahn, 2021 – Senior and Research Fellows at the Council on Foreign Relations** [Michael and Lauren, The Washington Quarterly 3-19-2021 “Leading in Artificial Intelligence through Confidence Building Measures”  [https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/0163660X.2021.2018794 acc on 6-21-2022](%20https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/0163660X.2021.2018794%20acc%20on%206-21-2022%20) TM]

The Risk of Inadvertent Escalation Third, either insufficient training and education or automation bias could make inappropriate uses of AI-enabled systems by leaders or operators that escalate tense situations more likely.13 The risk of deploying complex algorithms could be magnified when algorithmic systems from different sides interact in a crisis. When tensions are on a hair-trigger, human decision-makers generally look for off-roads from war, which is why accidental war and inadvertent escalation are quite rare despite fears of spiraling conflicts.14 Algorithmic proxies will almost inherently have less flexibility than human commanders and could interpret behavior as more threatening and/or not have the capacity to avoid war in the same way as humans, making inadvertent escalation more likely. Even in the case of human engagement with algorithms, avoiding the potential costs of constrained parameters could be tricky. If not trained properly, operators using AI-based aids can experience cognitive offloading and become over-reliant on the AI and less willing to intervene or deviate when presented with a suggestion or action from the AI—a phenomenon known as automation bias—especially in uncertain, unfamiliar, or time-critical situations, making accidents that could increase the risk of inadvertent escalation.15 Combined, certain characteristics of AI, such as the opacity of algorithms and potential speed of operations, along with procedural difficulties like TEVV for algorithms and training operators to use systems and software correctly, especially in the context of distrust due to great power competition, can increase the likelihood of dangerous unforeseen and undesirable outcomes such as accidents, unintended conflict, and inadvertent escalation.

#### The Impact - Military AI escalates conflicts – it causes accidental escalation, prevents war termination and undermines ceasefires and negotiations

Horowitz and Scharre 2021 - Director of the Emerging Capabilities Policy Office in the Office of the Under Secretary of Defense for Policy and  Vice President and Director of Studies at CNAS [Michael, Paul, January 12 2021, “AI and International Stability: Risks and Confidence- Building Measures”, <https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures>, Acc 6/18/22. M.A.]

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](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn30) 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](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn31) 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](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn32) and even in situations where the algorithms are relatively simple.[33](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn33) 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.

#### US standards for TEVV reduce the risk of accidental conflicts

**Schmidt and Work, 2021 – Chairs of The National Security Commission on Artificial Intelligence** [Eric and Robert, “National Security Commission on Artificial Intelligence Final Report Executive Summary” <https://www.nscai.gov/2021-final-report/> Acc 6/7/22 TA]

Manage risks associated with AI-enabled and autonomous weapons. AI will enable new levels of performance and autonomy for weapon systems. But it also raises important legal, ethical, and strategic questions surrounding the use of lethal force. Provided their use is authorized by a human commander or operator, properly designed and tested AI enabled and autonomous weapon systems can be used in ways that are consistent with international humanitarian law. DoD’s rigorous, existing weapons review and targeting procedures, including its dedicated protocols for autonomous weapon systems and commitment to strong AI ethical principles, are capable of ensuring that the United States will field safe and reliable AI-enabled and autonomous weapon systems and use them in a lawful manner. While it is neither feasible nor currently in the interests of the United States to pursue a global prohibition of AI-enabled and autonomous weapon systems, the global, unchecked use of such systems could increase risks of unintended conflict escalation and crisis instability. To reduce the risks, the United States should (1) clearly and publicly affirm existing U.S. policy that only human beings can authorize employment of nuclear weapons and seek similar commitments from Russia and China; (2) establish venues to discuss AI’s impact on crisis stability with competitors; and (3) develop international standards of practice for the development, testing, and use of AI-enabled and autonomous weapon systems.

### 1AC - Advantage: Nuclear AI

#### An AI Arms Race will increase nuclear instability – nations will rely on AI command and control, which is vulnerable to disinformation and interference

**Fitzpatrick, 2019 - former the IISS Non-Proliferation and Nuclear Policy Programme** [Mark, 21 May, “Artificial Intelligence and Nuclear Command and Control,”  [https://www.iiss.org/blogs/survival-blog/2019/04/artificial-intelligence-nuclear-strategic-stability 6/18/22](https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/00396338.2019.1614782%206/18/22) MD]

The year was 2021. The fictionalised events occurred in an IISS tabletop exercise in London in November 2018, for a project funded by the Carnegie Corporation of New York designed to examine the potential impact of artificial intelligence (AI) on nuclear strategic stability. Experts from China, Russia, the United States, the United Kingdom and the Czech Republic played roles as decision-makers in Beijing, Moscow and Washington as they sought to understand, address and take advantage of the scenarios and variations posed by the IISS control team. The major powers’ nuclear command-and-control systems increasingly rely on AI programmes, or, more precisely, expert systems and machine-learning algorithms, to enhance information flow, situational awareness and cyber security. Such capabilities can provide such systems with a larger window of opportunity in which to respond in the event of a crisis and thereby support de-escalation. Malevolent actors, however, can also use new technologies offensively to deceive, disrupt or impair command-and-control systems and their human controllers. The IISS tabletop exercise sought to explore several of these malfeasance pathways and vulnerabilities in plausible crisis scenarios. It demonstrated how an AI arms race could reduce strategic stability as the nuclear-weapons states become more reliant on AI for strategic warning in relation to nuclear command-and-control functions.[1](https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/00396338.2019.1614782) Escalation via ‘deep fakes’ A key danger played out in the exercise was the potential for third parties to spoof warning systems and embed disinformation to fool human operators. In the scenario, a hitherto unknown non-state actor, the World Peace Guardians, circulated falsified videos and photographs on social-media platforms to create the impression that three US special-operations-forces soldiers had been killed by nerve gas in clashes with Russian military advisers in Syria. Some US pundits argued the legal case for the use of tactical nuclear weapons in response. Doctored videos then appeared on American and Chinese media platforms that showed the families of several prominent US officials hurriedly departing Washington DC. Further eyewitness accounts on social media claimed that missile silos in the western US had gone to high alert, and that the crews of two had even opened the silo doors. As a result of these and other secondary indicators used by Russian AI-driven situational-awareness algorithms, Moscow’s strategic-warning systems began informing the Kremlin leadership that a US missile launch could be imminent. In the first situation report generated by the control team, Chinese President Xi Jinping cautioned the US not to conduct a nuclear strike against any country and warned that if the US did not provide ‘credible evidence’ that it was not mobilising for war, China would have to take unspecified defensive actions. Generally speaking, the scenario was a plausible example of escalation by third parties. In this case, a non-state actor made a deep fake sufficiently believable that it generated a crisis between two nuclear states. A workshop background paper explained how offensive AI capabilities could widen the psychological distance between the attacker and its target. The paper fore-warned that a third-party actor might attempt to use AI-driven adversarial inputs, data-poisoning attacks, and audio and video manipulation to create escalatory effects between nations. While other early-warning systems would eventually discredit the spoof, it would still create high levels of uncertainty and tension in a short period of time. Doctored videos would likely force both parties to put their respective militaries on high alert. They would utilise overhead imagery, signals intelligence or human reporting to determine the reality on the ground. Collecting and processing the intelligence would take precious time during an escalatory crisis in which AI algorithms were urging immediate response actions.

#### Untested AI systems jeopardize nuclear command and control – they are susceptible to hacking by non-state actors

**Fitzpatrick, 2019 - former the IISS Non-Proliferation and Nuclear Policy Programme** [Mark, 21 May, “Artificial Intelligence and Nuclear Command and Control,”  [https://www.iiss.org/blogs/survival-blog/2019/04/artificial-intelligence-nuclear-strategic-stability 6/18/22](https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/00396338.2019.1614782%206/18/22) MD]

Among the most revealing vulnerabilities exposed in the exercise was the way in which private-sector technological breakthroughs might be expropriated by another non-state actor for malevolent purposes, with dire consequences for strategic stability. In the hypothetical two months before the exercise started, a US technology company called QuantumAI, which received significant seed funding from the CIA and won Defense Advanced Research Projects Agency contracts for its advanced magnetometers and gravimeters, partnered with another company to launch four open-source, quantum-sensing satellites that used AI to measure infinitesimal changes in magnetic and gravitational fields on the earth. Subscribers began using the new system to identify new mining opportunities, sunken shipwrecks and toxic metallic effluents from industrial sites. QuantumAI only marketed its product to government-approved researchers from NATO countries, and its technology was export-controlled under both US International Traffic in Arms Regulations restrictions and the Wassenaar Arrangement. In round one, satellite imagery of the South China Sea that rendered the water transparent and revealed the location of three Chinese submarines was put online, allegedly from sources linked to QuantumAI. While the Chinese team suspected that the US government was responsible, the US team worried that whoever put this information online might similarly reveal the location of US submarines. Exactly that transpired in round two. The same website posting locational data regarding Chinese submarines expanded its coverage to include a global map with additional markers for Russian and US nuclear submarines. This posed a clear risk to strategic stability in compromising the second-strike capability that stealthy nuclear-armed submarines provide. Adding the markers for US submarines made it clear that the revelatory website was not affiliated with the US government. Its registration history and hosting service provider in Switzerland suggested it was actually linked to the World Peace Guardians. When US Department of Homeland Security and FBI officials approached large US social-media firms for information regarding World Peace Guardians accounts, the companies replied that they would ‘share relevant information with the government when it becomes available’. The US president did decide that QuantumAI hard drives could be seized, although the CEO of QuantumAI was not fully cooperative. The idea of trying to seize World Peace Guardians servers in Switzerland was discussed, but rejected. The QuantumAI aspect of the exercise was one manifestation of what the background paper described as a ‘black box model extraction’ vulnerability. Such an extraction reverse-engineers an AI system to determine its parameters. An adversary may be able to use this information to enhance the effectiveness of future operations against the system by stealing intellectual property; identifying sensitive or proprietary information related to the system’s training data or objectives; or developing ‘adversarial inputs’ to be covertly introduced into the original AI system. Such inputs confuse the system’s classifiers or its pattern-recognition function, thereby causing it to miscalculate, misclassify or misinterpret elements in its operational environment.

#### AI early warning systems undermine nuclear deterrence and stability – automation bias and false alarms ruin reliability

Horowitz and Scharre 2021 - Director of the Emerging Capabilities Policy Office in the Office of the Under Secretary of Defense for Policy and  Vice President and Director of Studies at CNAS [Michael, Paul, January 12 2021, “AI and International Stability: Risks and Confidence- Building Measures”, <https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures>, Acc 6/18/22. M.A.]

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](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn34) Nuclear stability is an obvious area of concern given the potential consequences of an intentional or unintentional nuclear detonation.[35](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn35) 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](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn36) 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](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn37) Education and training to ensure the responsible use of AI in early warning and forecasting scenarios will be critical.[38](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn38)

**The Impact - this increases the risk of accidental nuclear war because the speed of AI causes countries to adopt Lunch On Warning postures.**

**Shah, 2019 - Research Assistant at the Center for International Strategic Studies** [Syed Sadam CISS Insight Vol.VII, No.2 “The Perils of AI for Nuclear Deterrence” https://journal.ciss.org.pk/index.php/ciss-insight/article/download/10/9 Acc 5/25/22 TA]

The impact of AI on nuclear doctrines The use of autonomous delivery platforms and autonomous detection systems will increase the time urgency and pose several risks to the second strike capability of the states with less AI capabilities. This will compel states to rely on comparatively more ready nuclear arsenal where weapons are placed close to the delivery systems. The less AI capable states may also delegate the authority to use these weapons to lower commanders owing to the time urgent decision making needs. These states may decide to increase the number of their nuclear warheads to ensure survival of a sufficient number of their weapons in an attack situation. Time urgent nuclear doctrines will also be more prone to accidental and unauthorized nuclear launch. However, on the other hand, states with superior AI based detection systems may resort to first strike doctrines, though for different reasons. They might become confident to destroy their enemy in a comprehensive first strike and intercept the remaining retaliatory enemy strike by the timely identification of their AI based detection systems and leaving their missile defences to do the remaining job.

### 1AC - Advantage: Cyber Attacks

#### AI systems are extremely vulnerable to cyberattacks – accidents will escalate quickly

Scharre, 2019 - Vice President and Director of Studies at CNAS[Paul, May-June, “Killer Apps: The Real Dangers of an AI Arms Race,” https://omnilogos.com/killer-apps-real-dangers-of-ai-arms-race/6/18/22 MD]

Equally alarming for U.S. policymakers is the sharp divide between Washington and Silicon Valley over the military use of AI. Employees at Google and Microsoft have objected to their companies' contracts with the Pentagon, leading Google to discontinue work on a project using AI to analyze video footage. China's authoritarian regime doesn't permit this kind of open dissent. Its model of "military-civil fusion" means that Chinese technology innovations will translate more easily into military gains. Even if the United States keeps the lead in AI, it could lose its military advantage. The logical response to the threat of another country winning the AI race is to double down on one's own investments in AI. The problem is that AI technology poses risks not just to those who lose the race but also to those who win it. THE ONLY WINNING MOVE IS NOT TO PLAY Today's AI technologies are powerful but unreliable. Rules-based systems cannot deal with circumstances their programmers did not anticipate. Learning systems are limited by the data on which they were trained. AI failures have already led to tragedy. Advanced autopilot features in cars, although they perform well in some circumstances, have driven cars without warning into trucks, concrete barriers, and parked cars. In the wrong situation, AI systems go from supersmart to superdumb in an instant. When an enemy is trying to manipulate and hack an AI system, the risks are even greater. Even when they don't break down completely, learning systems sometimes learn to achieve their goals in the wrong way. In a research paper last year, a group of 52 AI researchers recounted dozens of times when AI systems showed surprising behavior. An algorithm learning to walk in a simulated environment discovered it could move fastest by repeatedly falling over. A Tetris-playing bot learned to pause the game before the last brick fell, so that it would never lose. One program deleted the files containing the answers against which it was being evaluated, causing it to be awarded a perfect score. As the researchers wrote, "It is often functionally simpler for evolution to exploit loopholes in the quantitative measure than it is to achieve the actual desired outcome." Surprise seems to be a standard feature of learning systems. Machine-learning systems are only ever as good as their training data. If the data don't represent the system's operating environment well, the system can fail in the real world. In 2018, for example, researchers at the MIT Media Lab showed that three leading facial recognition systems were far worse at classifying dark-skinned faces than they were at classifying light-skinned ones. When they fail, machine-learning systems are also often frustratingly opaque. For rules-based systems, researchers can always explain the machine's behavior, even if they can't always predict it. For deep-learning systems, however, researchers are often unable to understand why a machine did what it did. Ali Rahimi, an AI researcher at Google, has argued that much like medieval alchemists, who discovered modern glassmaking techniques but did not understand the chemistry or physics behind their breakthroughs, modern machine-learning engineers can achieve powerful results but lack the underlying science to explain them. Every failing of an AI system also presents a vulnerability that can be exploited. In some cases, attackers can poison the training data. In 2016, Microsoft created a chatbot called Tay and gave it a Twitter account. Other users began tweeting offensive messages at it, and within 24 hours, Tay had begun parroting their racist and anti-Semitic language. In that case, the source of the bad data was obvious. But not all data-poisoning attacks are so visible. Some can be buried within the training data in a way that is undetectable to humans but still manipulates the machine. Even if the creators of a deep-learning system protect its data sources, the system can still be tricked using what are known as "adversarial examples," in which an attacker feeds the system an input that is carefully tailored to get the machine to make a mistake. A neural network classifying satellite images might be tricked into identifying a subtly altered picture of a hospital as a military airfield or vice versa. The change in the image can be so small that the picture looks normal to a human but still fools the AI. Adversarial examples can even be placed in physical objects. In one case, researchers created a plastic turtle with subtle swirls embedded in the shell that made an object identification system think it was a rifle. In another, researchers placed a handful of small white and black squares on a stop sign, causing a neural network to classify it as a 45-mile-per-hour speed-limit sign. To make matters worse, attackers can develop these kinds of deceptive images and objects without access to the training data or the underlying algorithm of the system they are trying to defeat, and researchers have struggled to find effective defenses against the threat. Unlike with cybersecurity vulnerabilities, which can often be patched once they are uncovered, there is no known way of fully inoculating algorithms against these attacks.

#### AI increases threats to cyber security because it increases the dependence on data, while creating new targets of data attacks.

Goldfarb and Lindsay, 2022 – Chair in AI and health care at the Univ of Toronto, Prof of Cybersecurity at the Georgia Tech [Avi, Jon, 2/25/22, <https://doi.org/10.1162/isec_a_00425>, “Prediction and Judgement: Why Artificial Intelligence Increases the Importance of Humans in War” 6/18/22, LND]

In short, there may be more data in modern war, but data management has also become more challenging. Although U.S. weapons may be fast and precise, U.S. wars in recent decades have been protracted and ambiguous.[106](javascript:;) We argue that AI will most likely deepen rather than reverse these trends. Indeed, automation is both a response to and a contributing cause of the increasing complexity of military information practice.

Just as commanders are already preoccupied with C2 architecture, officers in AI-enabled militaries will seek to gain access to large amounts of data that are relevant to specific tasks in order to train and maintain AI systems. Units will have to make decisions about whether they should collect their own data organically or acquire shared data from other units, government agencies, or coalition partners. We expect many relevant databases to be classified and compartmented given the sensitivity of collection techniques or the content itself, which will complicate sharing. Units might also choose to leverage public data sources or purchase proprietary commercial data, both of which are problematic because nongovernmental actors may affect the quality of and access to data. As militaries tackle new problems, or new operational opportunities emerge, data requirements will change, and officers will have to routinely find and integrate new data sources. AI strategy will require militaries to establish data policies, and thus negotiating access to data will be an ongoing managerial—and human—challenge. We contend that militaries will face not only data access but also data relevancy challenges. Heterogenous data-generating processes allow biases and anomalies to creep into databases. Although metadata may help to organize information processing, they are also vulnerable to data friction that only humans can fix.[107](javascript:;) Cleaning and curating data sources will therefore be as important as acquiring them in the first place. To the challenges of producing or procuring data must be added the challenges of protecting data. Just as supply chains become attractive targets in mechanized warfare, data supplies will also become contested. Overall, we expect the rise of AI to exacerbate the already formidable challenges of cybersecurity. Cybersecurity professionals aim to maintain the confidentiality, integrity, and availability of an organization's data. Two of these goals—integrity and availability—capture the AI requirements of unbiased and accessible data, as discussed above. The goal of confidentiality is also important insofar as data provide AI adopters with a competitive advantage. In commerce, AI companies often try to own (rather than buy) the key data that enable their machines to learn.[108](javascript:;) The military equivalent of this is classified information, which is hidden from the enemy to produce a decision advantage.[109](javascript:;) Military organizations will have strong incentives to protect the classified data that military AI systems use to learn. For the same reasons, adversaries will have incentives to steal, manipulate, and deny access to AI learning data. To date, most discussions of AI and cybersecurity have focused on a substitution theory of cybersecurity, that is, using AI systems to attack and defend networks.[110](javascript:;) But we argue that a complementary theory of cybersecurity is just as if not more important. AI will require the entire military enterprise to invest more effort into protecting and exploiting data. If AI systems are trained with classified information, then adversaries will conduct more espionage. If AI enhances intelligence, then adversaries will invest in more counterintelligence. If AI provides commanders with better information, then adversaries will produce more disinformation. Inevitably, different parts of the bureaucracy will tussle among themselves and with coalition partners and nongovernmental actors to access and curate a huge amount of heterogeneous and often classified data. Organizations will also struggle with cyber and intelligence adversaries to maintain control of their own data while also conducting their own campaigns to collect or manipulate the enemy's data. To appreciate the strategic implications of AI, therefore, it is helpful to understand cyber conflict, most of which to date resembles espionage and covert action more than traditional military warfare. In fact, chronic and ambiguous intelligence contests are more common than fast and decisive cyberwar.[111](javascript:;) Military reliance on AI becomes yet another factor abetting the rise of cyber conflict in global affairs, and the (ambiguous, confusing, interminable, gray zone) dynamics of cyber conflict are likely to have a strong influence on the dynamics of AI conflict.

**The Impact - Cyber attacks escalate to nuclear war due to the fog of war – simulations prove**

**Schneider, 2022** - Fellow at the Hoover Institution at Stanford University [Jacquelyn; March 7; Affairs, “The Biggest Cyber Risk in Ukraine? How Russian Hacking Could Threaten Nuclear Stability,” <https://www.foreignaffairs.com/articles/ukraine/2022-03-07/biggest-cyber-risk-ukraine> from Cyber Starter Packet]

Why the apparent restraint? It is almost impossible to know exactly why (or if) the Russians have indeed held back. Perhaps cyber-operations have been attempted and failed; perhaps Russian President Vladimir Putin has held his cyber-capabilities in reserve, saving them for later. Or maybe cyber-operations have taken place, but their effect—which is often virtual and not clearly attributed—will take longer to materialize. What is known is that the conflict is far from over, and the next question becomes whether cyber-operations could play a larger role as the war turns more violent. It is likely that the next stage of conflict will more than ever be defined by planes, tanks, artillery, and soldiers. It seems unlikely, given the amount of indiscriminate damage currently being inflicted by Russia, that cyber-operations will escalate the violence of the campaign within Ukraine. That said, could cyber-operations lead to horizontal escalation, drawing NATO into the fight, for example? Or, given that the United States and Russia are the world’s largest nuclear powers, could cyber-operations escalate to the worst possible outcome—nuclear war? Recent wargaming research suggests that cyber-exploits into nuclear command and control may be enticing for states looking to neutralize a nuclear escalation threat in the midst of a conventional war, and that actors may underestimate the danger of these exploits and vulnerabilities to nuclear stability. GETTING PULLED IN One way cyber-operations could lead to escalation is by pulling the United States or NATO into the conflict. Mark Warner, the Democratic senator from Virginia, warned in late February that potential Russian cyberattacks on critical infrastructure in Ukraine could have accidental spillover effects on NATO countries—for instance if a Russian cyberattack on Ukrainian energy infrastructure caused an outage in a NATO neighbor like Poland. This could inadvertently trip Article 5 of NATO’s founding treaty, which states that an armed attack against one member state will be considered an attack against them all. This would be uncharted waters for NATO, which only recently publicly stated that cyberattacks might invoke Article 5 and is still ambiguous about what types of cyberattack—which range from virtual outages to data manipulations to physical damage (in extremely rare circumstances)—might be serious enough for NATO to respond with conventional retaliation. The Biden administration has warned that the United States would respond to cyberattacks on U.S. critical infrastructure, such as the country’s electrical grid or water supply (although officials stopped short of saying how the United States would respond). So far, the United States has answered previous cyberattacks with either sanctions, law enforcement actions, or the confiscation of cryptoassets. None of these options seem likely to deter Putin at this point, and so the Biden administration may find itself in an unprecedented position of having few credible options to threaten Russia. It is certainly possible that Putin, facing a conventional war that he thinks he might lose, could attack critical infrastructure in the United States or other NATO countries in the hope that their citizens will push their governments to abandon Ukraine. The financial sector, in particular, would seem to be a logical target for Russian cyberattacks, given the damage that Western economic actions have already done to the Russian economy. It is difficult to create widespread and long-lasting effects with cyberattacks, however, and the financial sector is the best equipped and most advanced cyber-defender in the world. Plus, research I’ve conducted with Sarah Kreps, director of the Cornell Tech Policy Lab, finds that the American public views cyberattacks as qualitatively different from conventional means of warfare—more akin to economic sanctions than bombs. Thus, cyberattacks are unlikely to provoke the kind of retaliation or emotional response that would pull the United States or its NATO allies into a war with Russia. What’s more, the United States can probably withstand the short-term damage to critical infrastructure that a Russian cyberattack might create, and such attacks might actually increase resolve to support Ukraine. This means a deliberate choice by Russia to use cyberattacks against the United States or NATO to “escalate to dominate”—deliberately ratcheting up the pressure to force Washington to back off—would likely fail. A more troubling scenario involves accidental escalation from cyber-operations—that is, when critical infrastructure is unintentionally damaged by a cyberattack or when a cyberattack is misattributed to Russia (or the United States). This is especially dangerous for civilian infrastructure that also serves military or security purposes—for example, harming a refugee train by using a cyberattack targeting railroads also used to move troops and supplies to the front. Plus, a jumble of actors has jumped into this space, from criminal syndicates to cyber-militias to hacker collectives such as Anonymous. That increases the chances that one of these players will target civilian infrastructure, and misattribution to either Russia or the United States could needlessly trigger retaliation. WHEN CYBER GOES NUCLEAR By far the most dangerous form of escalation is the possibility that a cyber-operation increases the likelihood of nuclear war. How likely is such a scenario? No one may know if Russia has a cyberweapon that can target nuclear weapons (or, for that matter, whether the United States does), but there are theories and some data about how the cyber-realm might affect nuclear stability. American policymakers have generally recognized that attempting to interfere with nuclear command, control, and communications could lead to dangerous incentives for states to launch nuclear weapons preemptively. Threats to nuclear command and control, for example, could leave states so fearful about their second-strike capability (the ability to launch a nuclear weapon in retaliation against an attacker) that in the midst of a conflict they would feel compelled to use nuclear weapons preemptively. Some scholars have warned that attacks against nuclear command-and-control systems could make it impossible to control nuclear war and keep it limited, leading to inadvertent nuclear Armageddon. Despite these fears about the dangers of attacking nuclear command and control, there was never an agreement between the United States and the Soviet Union (and subsequently Russia) to not attack each other’s nuclear command, control, and communications.

#### TEVV is key to protect military AI from cyber-attacks – AI systems are prone to attack and failure – TEVV helps promote innovations to prevent these attacks.

**Flournoy, Haines et al 2020 - former Under Secretary of Defense for Policy and Director of National Intelligence** [Michèle and Avril, Gabrielle Cheftiz Special Assistant to the Under Secretary of Defense for Policy October, “Building Trust through Testing Adapting DOD’s Test & Evaluation, Validation & Verification (TEVV) Enterprise for Machine Learning Systems, including Deep Learning Systems” https://cset. georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf Acc 6/23/22 JZ]

Second, the context in which DOD operates means these technologies are prone to adversary attack and system failure, with very real consequences. Machine learning systems have an increased potential for failure modes relative to other systems, such as bias due to a distribution shift in data, as well as novel vulnerabilities to attacks ranging from data poisoning to adversarial attacks. One could easily imagine an image classifier that accidentally classifies a civilian school bus as a tank or an adversary exfiltrating a model processing sensitive intelligence, surveillance, and reconnaissance or communications data. Image classification algorithms developed for one environment (e.g., the desert) could turn out to work incorrectly in another environment (e.g., cities). Third, with an effective TEVV system, the United States can reduce barriers to innovation and facilitate U.S. leadership in ML/DL technologies. As most of the innovation in ML/DL will come from the private sector, unless the U.S. government is able to effectively draw on private sector work in this arena, it will not be able to leverage the best cutting-edge technology. Research on new TEVV methods and organizational reforms to adapt the current system is simply not keeping pace with private sector development. Without urgent reforms and prioritized investment in new research and infrastructure, the Defense Department will lose its chance to shape industry’s approach to ML/DL development in a manner consistent with DOD standards for safety, reliability, and accountability. It will lose the opportunity to take advantage of new private sector developments, while allowing other nations without such standards to adopt the latest innovations. It is critical that the U.S. government not only shape its own U.S. industry standards but also promote compatible global standards and norms. Fourth, adversary advancements will likely increase pressure to field AI-enabled systems faster, even if testing and assurance are lacking. China has elevated AI to be a major national priority across sectors and is already exporting armed drones with varying degrees of autonomy.3 Russia is also pursuing R&D on AI for military 5 purposes4 and fields AI-enabled robotic systems in Syria with little regard for ethical considerations.5 However, it shouldn’t be a race against our competitors to field AI systems at any cost. It’s a race to field robust, lawful, and secure AI systems that can be trusted to perform as intended. Finally, high standards for robustness, assurance, interpretability, and governability can ultimately be a tremendous source of strategic advantage, incentivizing industry to harden systems to adversary attack. Taken together, these risks and opportunities suggest that devising an effective, efficient, and ethical TEVV process is critical for maintaining the U.S. military and economic competitive edge, as well as deploying reliable and trustworthy ML/DL systems.

### 1AC - Advantage: Operator Trust

**Brittle AI undermines operator trust – soldiers need to know that their weapons work.**

**Flournoy, Haines et al 2020 - former Under Secretary of Defense for Policy and Director of National Intelligence** [Michèle and Avril, Gabrielle Cheftiz Special Assistant to the Under Secretary of Defense for Policy October, “Building Trust through Testing Adapting DOD’s Test & Evaluation, Validation & Verification (TEVV) Enterprise for Machine Learning Systems, including Deep Learning Systems” https://cset. georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf Acc 6/23/22 JZ]

All of these challenges undermine the establishment of trust between operator and system, which is essential given the U.S. military is likely to deploy DL as part of human-machine teams. Critical to building this trust will be the ability to accurately characterize the bounds of a system’s behavior—that is, when it will work and when it will not. If DOD has an image classifier that only works in a desert environment and operators know it will only work in a desert environment, then they are more likely to trust it. Operators don’t need to know exactly how a system works, only under what conditions it will and won’t work. This will require new methods of testing and assurance to predict system failure and govern system performance.

#### Operator trust is essential to effective military operations – NATO deterrence fails if the military does not trust its weapons systems

Trabucco and Stanley-Lockman, 2022 – prof of Political Science, University of Copenhagen and prof of Defense and Strategic Studies, Nanyang Technological University [Lena and Zoe, The Oxford Handbook of AI Governance, March, “NATO’s Role in Responsible AI Governance in Military Affairs” https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780197579329.001.0001/oxfordhb-9780197579329-e-69 Acc 4/16/22 TA]

Safety and security For humans to meet ethical and legal commitments when developing and deploying AI, the systems themselves must be safe, secure, and reliable. More simply put, if humans and institutions interacting with AI do not have confidence that the systems will perform as expected, then they cannot assure that its development and deployment are responsible. This makes safety and security a key pillar of responsible AI governance for any actor.82 As this section explores for NATO in particular, safety and security are indispensable to the Alliance’s stated goals to focus its approach to EDTs in the areas of “deterrence and defense, capability development, legal and ethical norms, and arms control aspects.”83 Politically, democratic militaries using AI cannot be accountable to their citizenries nor their coalition partners if they lack mechanisms to trace and explain how their systems are reliable. Accidents and interference with AI systems could likewise create political risks for the Alliance. For example, if deepfakes and micro-targeted information attacks compromise confidence in the integrity of information used to build a common operating picture, then the operational difficulties could also erode political trust between Allies in a few key ways. In the North Atlantic Council, disagreement about the integrity of information could slow the decision-making body’s ability to react to fast-changing operational realities.84 Further, compromised AI systems may not only make it harder for forces to prevent harm to non-combatants, but also to prevent friendly fire. In this way, coalition forces arguably face even higher obligations to coordinate on the reliability of their systems, relative to adversaries and near-peer competitors that tend to operate alone. As such, responsible AI governance is not purely technical; policy alignment and strategic planning are likewise necessary to draw attention to risk management above the tactical level.

#### Focusing on AI safety is key to US Military Effectiveness because it is key to allied cooperation and interoperability

Lewis 2019 - project lead for the DOD’s Joint Lessons Learned studies [Larry, “AI Safety: Charting out the High Road”, War on the Rocks Dec 19, https://warontherocks.com/2019/12/ai-safety-charting-out-the-high-road/ LMSi]

Safety Is Strategically Smart Why is it a problem if the United States does not rise to establish concrete steps to emphasize AI safety? After all, current and former U.S. government leaders have spoken about how neither Russia nor China will be slowing down their AI efforts in order to address ethical or safety issues. Does it matter? A focus on safety and care in the conduct of operations has served the United States well. During the second offset, Washington developed precision capabilities to help counter the Soviet Union’s advantages in troop numbers. These developments then enabled the United States to take additional steps to promote safety in the form of reduced civilian casualties: developing and fielding new types of munitions for precision engagements with reduced collateral effects, developing intelligence capabilities for more accurately identifying and locating military targets, and creating predictive tools to help estimate and avoid collateral damage. These steps had strategic as well as practical benefits, enhancing freedom of action and boosting legitimacy of U.S. actions while enabling steps to reduce the civilian toll of recent operations. If AI is leveraged to help promote safety on the battlefield, it can yield similar strategic and practical benefits for the United States. AI safety also has relevance to U.S. allies. Unlike its peer competitors, Russia and China, the United States almost always operates as part of a coalition effort. This is a significant advantage for the United States politically, numerically, and in terms of additional capabilities that can be brought to bear. But ally cooperation, and the interoperability of partners within a coalition, will depend on what capabilities our allies are willing to adopt or to operate in the same battlespace with. It is important that the United States be able to convince would-be allies of both the effectiveness and the safety of its military AI. AI and America’s Future The revelation that China used AI to violate human rights contrasts strongly to U.S. promises to take the high road with regard to its military applications of AI. Eric Schmidt and Bob Work have declared that the United States could easily lose its leadership in AI if it does not act urgently. The leadership the United States has shown in AI ethics is commendable, but to fully be the leader it needs to be for our national security and for our prosperity, America must lead the way on safety too. The opportunity to develop AI of unrivaled precision is historic. If it can build AI that is lethal, ethical, and safe, the United States would have an edge in both future warfare and the larger climate of competition that surrounds it. Developing safer AI would once again show the world that there is no better friend and no worse enemy than the United States.

#### The Impact – an effective NATO is key to containing the Ukraine conflict – weakness will cause Russia to escalate to a great power war.

Graham, 2022 – Fellow at the Center for Preventive Action and Europe Program at the Council on Foreign Relations [Thomas, March 8 Contingency Planning Memorandum No. 38 “Preventing a Wider European Conflict” [https://www.cfr.org/report/preventing-wider-european-conflict](https://www.cfr.org/report/preventing-wider-european-conflict%20) Starter Packet]

The large-scale Russian invasion of Ukraine now underway could quite plausibly precipitate a wider conflict in Europe. The United States is focused primarily on raising the costs to Russia with punishing sanctions and reassuring North Atlantic Treaty Organization (NATO) allies neighboring Russia of its commitment to collective defense. Less attention has been given to containing the war to Ukraine and preventing its escalation into a broader European conflict. The stakes are enormous. The ripple effects of a wider conflict in Europe would spread across the globe, stressing the geopolitical, economic, and institutional foundations of the international order the United States has fashioned and underwritten since the end of the Second World War. It would test the resilience of the U.S. global system of alliances, the international financial system, global energy markets, arms control regimes, and global institutions in the face of ever more violent great power competition. No region of the world would be spared, although developments on the Eurasian supercontinent, the other locus of world power and economic might outside North America, would bear the gravest consequences for U.S. interests. NATO (North Atlantic Treaty Organization) The Russian military intervention in Ukraine could easily escalate into a larger conflict stretching from the Baltic to the Black Sea and further west into Europe. Although Russia, wielding massive military superiority, might overrun Ukrainian forces in a matter of weeks, stabilizing and pacifying the country will likely prove to be a grueling and costly affair. A significant Ukrainian resistance movement is almost certain to emerge. With sustained Western support, it could prolong the warfare for months, if not years. The first wave of sanctions that Washington has levied on Moscow could be followed by others in a continuing effort to raise the cost to Moscow and force it to yield. A negotiated end to the conflict will not come easily, since Washington has framed it in Manichean terms as a world historical struggle between the democratic West and the aggressive, malevolent, and autocratic Russia. Anything short of “victory” will be decried as surrender or appeasement in the West, while Russia will not capitulate on a matter it considers vital to its security and prosperity. The stage is thus set for an escalating cycle of violence, with Moscow seeking to stamp out a Ukrainian insurgency and retaliate against Western efforts to stop Russia’s advance. If the conflict wears on, Moscow could be increasingly tempted to expand its military operations further into Europe to achieve its goals. As a first option, Russia could intensify pressure on states neighboring Ukraine (e.g., Hungary, Poland, Romania, and Slovakia) that could provide safe havens for insurgents or the inevitable government-in-exile. It will doubtless reinforce its military presence in Kaliningrad and elsewhere in the Baltics and patrol the Baltic Sea more aggressively. It could deploy hybrid-war tactics—cyberattacks, disinformation campaigns, and economic sabotage—to destabilize countries providing safe havens. If those actions did not sufficiently degrade the resistance, Moscow could even launch direct attacks on insurgents and their supporters outside Ukraine, as well as attempt to assassinate leading figures in the government-in-exile, akin to the attacks it has made on Chechen rebels and Federal Security Service (FSB) defectors in Europe in recent years. Such steps could, at a minimum, draw frontline NATO states directly into the military conflict with Russia, obligating the United States and other allies to come to their defense. To build up further pressure, Moscow could also “weaponize” the inevitable refugee flows into neighboring states. Refugees, who would likely number in the millions, would move first into unoccupied Ukrainian territory but eventually into adjacent European states, which have shown little tolerance for outsiders. Moscow could use harsh military and police tactics that would increase the number of refugees and seek to guide them into countries where they would create the greatest socioeconomic stress, such as Moldova. In addition, Moscow could increase the tension by pushing Belarusian President Aleksandr Lukashenko to again seek to push thousands of Middle Eastern migrants across the borders into Poland and Lithuania. That could lead to border clashes, as it almost did on occasion last fall, with Russia supporting its ally, Belarus, and NATO states coming to the defense of allies under attack. A second option Moscow could pursue is opening up a second front in the Balkans. In recent years, Russia has taken a number of destabilizing actions in the region, seeking to weaken Montenegro after its accession to NATO, exacerbate tensions between Serbs and Bosniaks in Bosnia-Herzegovina, and undermine relations between Serbia and Kosovo. As it fought in Ukraine, Russia could encourage Republika Srpska leader Milorad Dodik to press for separation from Bosnia, threatening to reignite the bitter wars of the 1990s in the former Yugoslavia. A Balkans war would complicate the security calculus of all countries in the region, as well as that of Germany and France, which have significant interests there. To quell the fighting, NATO countries could decide to use military force against Bosnian Serb forces enjoying Russian support. If the conflict wears on, Moscow could be increasingly tempted to expand its military operations further into Europe to achieve its goals. A third, riskier, option would be to directly attack the United States, the country that Moscow believes is orchestrating a larger anti-Russia campaign. In response to Western sanctions designed to crater Russia’s financial system and undermine critical industries, Moscow could launch major cyberattacks against U.S. critical infrastructure. If a cyberattack were to take down a major financial institution or corrupt its records, the ensuing havoc in U.S. markets could prompt overwhelming public and congressional pressure for a forceful response. The U.S. and NATO response to Russian actions will impact Moscow’s decisions on the conduct of the conflict. Both a weak response and an excessively harsh one could lead to escalation. In the first case, Moscow could be tempted to press militarily even further into Europe to enlarge its sphere of influence. Vladimir Putin has demanded that NATO withdraw its forces back to the lines they held in 1997, when the NATO-Russia Founding Act was signed and the first wave of post−Cold War expansion remained in the future. His remarks announcing the start of hostilities against Ukraine hinted at a broader effort to restore Russia’s control over all of the former Soviet Union. That could include military action against the Baltic states, especially Lithuania, through which Moscow could try to carve out a land corridor to Kaliningrad, a Russian exclave on the Baltic Sea. NATO would have little choice but to provide military aid to those states if it did not want to forfeit its role as the central pillar of European security. Crippling sanctions, meanwhile, could provoke Putin to lash out with greater violence. If Putin felt cornered, he could escalate the conflict either horizontally to other countries or vertically to the nuclear level in a desperate effort to save himself, his regime, and, in his mind, Russia itself. And he could find considerable public support for such a reaction. Already, some Russians believe that U.S. and EU sanctions are aimed not simply at the leaders behind the war but, by cratering the economy, at all Russians. Warning Indicators As is the case with the current crisis in Ukraine, Moscow’s intentions will remain ambiguous. The indicators of an approaching escalation in the conflict beyond Ukraine are likely to fall into three categories. The first indicators that political and military conditions are increasing the risk of broader conflict include a breakdown in channels of communication with Moscow. The absence of active diplomatic ties would preclude a negotiated resolution of the conflict in Ukraine. An end to U.S.-Russian military-to-military channels would undermine any effort to avoid direct military conflict between the two countries. Another indicator would be major insurgent successes that dramatically increase Russian casualties. Moscow would be tempted to move more aggressively against insurgent safe havens rather than capitulate on what it considers to be its vital interest in Ukraine. A wider European conflict would pose the stiffest challenge to the global standing of the United States since the end of the Cold War and to the international system it has built and underwritten for decades longer. Second are the indicators that Moscow is preparing for a broader conflict, which it would undoubtedly argue had been forced by Western actions. Such signs include Kremlin efforts to prepare the Russian public for a wider conflict, which could entail official statements, greater media focus on escalating Western “aggression,” an increased pace of civil defense drills, and mobilization of reserves. Another indicator includes the massing of Russian forces in the Baltic region. It could include such moves as aggressive hybrid actions to destabilize Poland and the Baltic states, coupled with efforts to rally indigenous ethnic Russian communities against their governments. Third are the indicators that Moscow is intentionally seeking to widen the conflict. This could include greater support for Bosnian Serb leader Dodik, such as diplomatic and financial backing, and provision of weapons. They could also encourage Serb leaders to more assertively pursue their grievances against Kosovo. Implications for the United States A wider European conflict would pose the stiffest challenge to the global standing of the United States since the end of the Cold War and to the international system it has built and underwritten for decades longer. It would test the durability of its global system of alliances and the efficacy of international regimes and institutions that have guarded world peace, security, and prosperity. The challenge would come at a time when the United States itself is in immense disarray, as a deeply polarized polity confronts massive domestic problems—the pandemic, inflation, racial justice, and cultural wars—that leave less time and fewer resources for foreign matters. The United States will be tested to see whether it can muster the will, energy, and creativity to execute an effective policy toward the unfolding crisis in Europe. At home, public attention has been focused on developments in and around Ukraine, but the Joe Biden administration cannot ignore the home front. In response to U.S.-levied sanctions, Russia can be expected to step up its cyber operations against the United States. It will more actively sow disinformation, seek to exacerbate domestic tensions, and paralyze critical infrastructure. The severity of the attacks will likely rise in proportion to the harshness of the sanctions Washington levies on Moscow. Abroad, the fate of the transatlantic community, a central pillar of U.S. security and prosperity, would be a stake. One of the Biden administration’s priorities, as laid out in the Interim National Security Strategy Guidance released in March 2021, is repairing U.S. alliances—especially with Europe—after four disruptive years under President Donald Trump. Although relations are more cordial, significant substantive differences remain and the willingness of allies to align behind a common purpose for the long haul remains questionable. The United States’ allies have rallied behind a harsh set of sanctions in response to Russia’s invasion of Ukraine, but preserving unity as the conflict drags on remains a challenge, especially if sacrifice is spread unevenly across NATO, as will most likely be the case. Putin will seek to exploit divisions through differentiated levels of pressure on NATO members, targeted energy cutoffs, offers of negotiation, and the like to advance two long-standing Russian goals: the end of NATO as a collective defense organization and the erosion of the foundations of the EU. Should he succeed, the new order that would emerge in Europe is far from certain. But Russia would undoubtedly play a central role in its formulation, and almost any conceivable new order would diminish the power and role of the United States on the continent. A similar situation obtains in the Indo-Pacific region. The Biden administration spent 2021 bolstering relations with its allies and partners—energizing the Quad (the United States, Australia, India, and Japan), and cutting a submarine deal with the United Kingdom and Australia—to meet the growing strategic challenge posed by China. A major, prolonged European distraction could undo further efforts to pivot to Asia, raise doubts among allies and partners about the credibility of the U.S. commitment, and free China to pursue its objectives with greater vigor. The United States could avoid this outcome by pursuing lesser goals in Europe—leading to the quicker development of a new order less favorable to American interests—or by a massive buildup of its military capabilities that would enable it to play a major, perhaps decisive, role in both regions. The latter would have to come at the cost of the Biden administration’s domestic priorities. Whether the Biden administration could muster sufficient domestic political support, if it decided to move in this direction, is far from certain. The United States’ allies have rallied behind a harsh set of sanctions in response to Russia’s invasion of Ukraine, but preserving unity as the conflict drags on remains a challenge, especially if sacrifice is spread unevenly across NATO. In addition to regional challenges, a major European conflict would also stress critical international regimes and institutions. One of the first victims would likely be the arms control regime that has served as the foundation of strategic nuclear stability for the past fifty-plus years. The United States withdrew from some central elements—including the Anti-Ballistic Missiles (ABM) and the Intermediate-Range Nuclear Forces (INF) treaties—but two critical elements have remained in place: the New START treaty and the Nonproliferation Treaty (NPT). A wider conflict in Europe would all but guarantee that the United States and Russia could not agree to a follow-on treaty to the New START treaty before it expires in 2026, and the NPT review conference tentatively scheduled for August 2022 would fall by the wayside. As a consequence, the incipient arms race now underway, fueled by new technologies—hypersonics, cyber tools, and artificial intelligence—would accelerate. A new wave of nuclear proliferation could ensue, especially if U.S. allies and partners lose faith in America’s commitment to extended deterrence. Mutually assured destruction, which for better or worse has anchored strategic stability since the early 1970s, would be severely stressed in a multipolar nuclear landscape with Russia and the United States fighting at least a proxy war. Likewise, a broader conflict in Europe would stress, perhaps to the breaking point, the United Nations and many of its auxiliary organizations. Already stymied by a growing rift between the Western permanent members and Russia and China, the Security Council would have failed in its primary reason for being—to prevent the outbreak of a major conflict in Europe. It could continue to exist as a forum for the airing of grievances and acrimonious debate, but it would serve little purpose as a platform for addressing major global issues. Finally, the humanitarian costs of a wider conflict in Europe would be staggering, particularly given the destructiveness of modern weapons. Beyond the physical destruction and loss of life, untold numbers of refugees would flow across borders not only into Central East Europe but perhaps further West depending on the scale of the fighting. The strain on the socioeconomic systems—coming on top of the stress of the two-year-old pandemic, economic dislocation, and mounting inflation—could bring some close to collapse. Preventive Options U.S. policy toward Russia has traditionally been a combination of deterrence and diplomacy. The Biden administration deployed both as it tried to dissuade Russia from invading Ukraine. Both have a role to play in reducing the risk of a wider European conflict, now that Russia has invaded. Many of the steps that the Biden administration is now taking to counter Russia could be accelerated and expanded to deter it from expanding its military operations beyond that country. They would likely prove more effective due to NATO’s Article 5 collective defense guarantee, which does not apply to Ukraine. The Biden administration could: With its NATO allies, accelerate and expand its current augmentation of forces in vulnerable allies along the frontier with Russia to reassure them—and convince Moscow—of the alliance’s commitment to collective defense. Step up its already intensive schedule of consultations with allies to maintain alliance unity in the face of a burgeoning Russian threat. Develop a long-term plan to reduce Europe’s dependence on imported Russian gas, building on the stopgap measures it is already putting in place to deal with a near-term decision by Moscow to stop flows of gas westward. Consider cutting off energy imports from Russia, and asking the Europeans to do the same, but only after it has prepared the American public for the economic hardship (rising energy costs, inflation) such a step would entail. Accelerate efforts to harden American and allied critical infrastructure against cyber intrusions. The Biden administration could also resume its diplomatic efforts to find a negotiated solution. To that end, it could: Resist the temptation to cut off channels of communication, as past administrations have done in reaction to Russian aggression. White House−to-Kremlin and military-to-military channels will be critical to reducing misunderstandings that could lead to direct military confrontation between the two countries. In addition, a White House−to-Kremlin link could provide a platform for negotiating an end to the conflict before it spreads beyond Ukraine. Carefully recalibrate its rhetoric to ensure that the confrontation does not turn into an existential one, where victory, whatever that might mean, is the only acceptable outcome. Such a posture would ignore the reality that Russia is unlikely to capitulate in a matter of vital interest—and would escalate rather than surrender. Talk of regime change and possible war crimes charges would probably prove counterproductive and fuel public support for escalation, especially at a moment when polls suggest the war effort enjoys the backing of the vast majority of the Russian population. Avoid appearances that the United States and NATO are waging a conflict against the Russian people. Releasing constructive proposals for resolving the conflict (including provisions for the lifting of sanctions), and urging the Ukrainians to publish reasonable negotiating terms, would be more likely than bellicose warnings to turn the Russian elites and public against the war. Russians need to be persuaded that the United States and Europe are not seeking a punitive peace but are open to a renewal of relations should their country act to end the conflict. Accelerate efforts to get information to the Russian people that would give them a more accurate portrayal of the brutal, unnecessary conflict their leaders are waging allegedly on their behalf. Students and young professionals would be particularly receptive to such information and inclined to protest. The mitigating options identified below, with the exception of invoking Article 5, could also be taken now to induce Russia to de-escalate and withdraw from Ukraine and to prevent it from expanding its military operations beyond Ukraine. Mitigating Options Should the conflict spread beyond Ukraine despite U.S. efforts, the task will be to bring it to an end on terms favorable to the United States as quickly as possible. Washington could consider diplomatic initiatives, defensive steps, and sanctions. Diplomatically, Washington could invoke Article 5 of the North Atlantic Treaty to make clear NATO’s determination to come to the aid of members under Russian attack. It could call for an urgent session of the UN Security Council to focus on the threat posed by Russia to international peace and security. The debate would doubtlessly be acrimonious, but the United States needs to make a concerted effort to shape public opinion and isolate Russia as the aggressor. Washington could also propose a P5 (the five permanent members of the UN Security Council: the United States, China, France, Russia, and the United Kingdom) meeting to discuss steps to reduce the risk of nuclear war. To avoid turning this into a two-bloc standoff between Russia and China and the Western powers, India could be added to the discussion. But New Delhi could resist being drawn into an East-West conflict, as it has in the past. The Biden administration should take care not to provoke severe Russian retaliation or produce spillover effects that cause undue harm to its or its allies’ interests. With regard to defensive measures, Washington could enlarge NATO de facto to coordinate strategy and tactics with Sweden and Finland, with an eye to their de jure membership in the near future. It could also send a small NATO contingent to the Balkans (Albania, Croatia, Montenegro, and North Macedonia) to warn Serbia and Republika Srpska against aggressive actions against Kosovo and Bosnia-Herzegovina. Concerning sanctions, Washington could build on the sanctions it had already levied to raise the costs further. However, the Biden administration should take care not to provoke severe Russian retaliation or produce spillover effects that cause undue harm to its or its allies’ interests. Recommendations The Biden administration is already taking steps to prevent the spread of conflict in Europe and harden the resilience of allies and partners in the face of Russia’s invasion of Ukraine: further augmentation of NATO forces, including a greater presence of American troops and equipment, along the entire Russia/NATO frontier stretching from the Baltic to the Black Sea; frequent consultations with allies and partners; steps to handle the large-scale exodus of refugees from Ukraine; organization of fuel shipments to Europe from various sources to cover gaps in the event of a Russian cutoff of gas exports; measures to harden U.S. and allied computer networks against attacks. The task is to turn those expedient measures into strategies to fortify the transatlantic community against a prolonged threat from the East, which Russia will continue to pose even if the current crisis is somehow defused in the near future. In particular, the United States needs to work with its European allies to drastically reduce Europe’s dependence on Russian gas. The goal should be to cut that dependence in half by the end of the decade by fully using the regasification facilities in place, building more, and accelerating work on renewables. In addition, even as the United States and European Union are dealing with the war in Ukraine, they need to recommit themselves to sorting out the continuing problems in Bosnia-Herzegovina and between Serbia and Kosovo to reduce the opportunities for destabilizing Russian interference in the Balkans. Finally, to ease the burden on states bordering Ukraine, the United States should be working with the UNHCR and its allies to develop plans for the long-term resettlement of Ukrainian refugees throughout Europe in case of a long period of instability in Ukraine. The United States should confront the urgent crisis in Europe without unduly sacrificing focus on the strategic challenge in the Indo-Pacific. All these steps, however, do not go far enough to deal with the enduring Russia challenge. The Biden administration needs to do more, ideally as part of a larger effort to reposition the United States strategically on the global stage. Critically, the United States should confront the urgent crisis in Europe without unduly sacrificing focus on the strategic challenge in the Indo-Pacific, and to prepare for a major change in the geopolitics of the Eurasian supercontinent. A tall order but not an impossible task. There are three core elements to this task: rethinking NATO, enhancing the U.S. presence in the Indo-Pacific, and creating a security forum to enhance allies’ support for U.S. policy across Eurasia. Rethinking NATO. The strategic goal should be the achievement of a near perfect overlap in NATO and EU membership among European states. That would provide the foundation for the development of a united European pillar inside NATO, in a sense resolving the tension between NATO and the EU (if not necessarily between the United States and Europe). The European pillar would assume ever greater responsibility for the defense of the continent, backed up by the American strategic deterrent, thus freeing up American forces to deal with the growing challenges in the Indo-Pacific region. The alliance’s new strategic concept, to be adopted at the Madrid Summit this coming June, provides an opportunity to articulate this goal, as well as to lay out the full breadth and enduring nature of the Russia challenge. The United States should consider pressing for the following steps: Fortify NATO’s eastern border. The alliance should abandon the pledge of the 1997 NATO-Russia Founding Act not to deploy permanent substantial combat forces to new members. It should augment its forces in vulnerable member states, as long as there is no agreement between NATO and Russia to mutually restrict force levels in border zones. Prepare for the eventual membership of Finland and Sweden to reinforce the northern flank. In the face of Russian conduct, the populations of these two countries are reconsidering their long-standing traditions of neutrality. While staying out of the domestic debate, the United States and other allies should indicate that they would welcome the two countries into the alliance and articulate clearly the changing nature of the security environment in the Baltic region brought on by a more aggressive Russia. Repair relations with Turkey. This is a matter primarily for the United States, which has levied sanctions on its ally for its purchase of S-400s, an advanced Russian air defense system. The United States could take a first step by approving the sale to Turkey of the F-16s it has requested. Washington should also look for an opportunity amidst deteriorating relations with Moscow to persuade Ankara to reconsider its purchase of S-400s. Forego expansion into the former Soviet space for an extended period. No one believes that any former Soviet state will be ready for membership for years to come. Without necessarily abandoning the Open Door policy, the alliance should make clear that it will not expand eastward while it focuses on its own consolidation.

#### TEVV is key to build trust in commanders and operators. This improves our military readiness - broken weapons don’t win wars

**Flournoy, Haines et al 2020 - former Under Secretary of Defense for Policy and Director of National Intelligence** [Michèle and Avril, Gabrielle Cheftiz Special Assistant to the Under Secretary of Defense for Policy October, “Building Trust through Testing Adapting DOD’s Test & Evaluation, Validation & Verification (TEVV) Enterprise for Machine Learning Systems, including Deep Learning Systems” https://cset. georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf Acc 6/23/22 JZ]

The Pentagon cannot let TEVV become a barrier to fielding AI-enabled systems in an operationally relevant time frame, but must do so in a manner that engenders trust in such systems and is consistent with U.S. values and principles. The ultimate goal of any TEVV system 4 should be to build trust—with a commander who is responsible for deploying a system and an operator who will decide whether to delegate a task to such system—by providing relevant, easily understandable data to inform decision-making. Fielding AI systems before our competitors may not matter if DOD systems are brittle and break in an operational environment, are easily manipulated, or operators consequently lose faith in them. Military operations present a challenging environment. The Defense Department needs ML/DL systems that are robust and secure. They need to be able to function in a range of environmental conditions, against adversaries who are adaptive and clever, and in a manner that engenders trust by the warfighter.

### 1AC - Plan Text

#### Plan: The United States federal government should increase its security cooperation with NATO, expanding standardized testing, evaluation, verification, and validation programs for artificial intelligence systems, including life cycle testing.

### 1AC - Contention Two

#### Expanding TEVV prevents AI accidents – an Arms Race dilemma rushes deployment of unreliable weapons, which undermines effectiveness and safety.

Scharre 2021 – Director of the Technology and National Security Program at the Center for a New American Security [Paul, 6/28/21,https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/, “Debunking the AI Arms Race Theory”, 6/18/22, LND]

A related risk of a “racing” dynamic among competitors could come from an acceleration, not of the pace of operations on the battlefield, but of the process of fielding new AI systems. AI systems today have a host of safety and security problems that can make them brittle, unreliable, and insecure.[29](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn29) Because machine learning in particular can create new ways in which systems can fail, militaries face novel challenges in adopting AI systems.[30](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn30) Militaries will have to adopt new methods to test, evaluate, verify, and validate AI systems (also known as TEVV).[31](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn31) Such concerns related to autonomy are well known in the U.S. defense community,[32](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn32) although at present they have not been solved to a satisfactory degree. Machine learning introduces additional challenges with regard to testing, evaluation, verification, and validation. A rush to field AI systems before they are fully tested could result in a “race to the bottom” on safety, with militaries fielding accident-prone AI systems. There are strong bureaucratic and institutional imperatives for militaries to field systems that are robust and secure. Indeed, designing systems to military specification standards often means making them more robust for a wider range of environmental conditions and shocks than comparable commercial systems, even at the expense of other aspects of performance, such as size, weight, or usability. AI presents novel challenges, however, in achieving the robustness needed for operating in the complex, hazardous, and adversarial environments that often characterize military operations. Certain AI methods today, such as deep learning, remain relatively immature with significant reliability challenges. A 2017 Department of Defense report by the JASON scientific advisory group explained that deep neural networks are immature as regards the “illities”, including reliability, maintainability, accountability, validation and verification, debug-ability, evolvability, fragility, attackability, and so forth. … Further, it is not clear that the existing AI paradigm is immediately amenable to any sort of software engineering validation and verification. This is a serious issue, and is a potential roadblock to DoD’s [Department of Defense’s] use of these modern AI systems, especially when considering the liability and accountability of using AI in lethal systems.[33](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn33) The Defense Department’s 2018 AI strategy calls for building AI systems that are “resilient, robust, reliable, and secure.”[34](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn34) Yet, the current state of technology makes achieving this goal particularly difficult for AI systems that incorporate deep learning, a subfield of AI that has seen significant growth and attention in recent years. While there is active research underway to improve AI safety and security, militaries will have to adapt to the technology as it currently is, at least for the time being. An ideal process would be for militaries to engage in experimentation, prototyping, and concept development, but also to subject AI systems to rigorous TEVV under realistic operational conditions before deployment. Taking shortcuts on testing and evaluation and fielding a system before it is fully tested could lead to accidents, which, in some settings, could undermine international stability. In evaluating new technologies, militaries may be relatively accepting of the risk of accidents, which may lead them to tolerate the deployment of systems that have reliability concerns. In building and fielding new capabilities, militaries have to weigh the possibility of an accident occurring against other concerns, such as forgoing valuable military capabilities. The military operational environment is fraught with risk, in both training and real-world operations. Military institutions balance managing this risk with other factors, such as the need for training, developing new capabilities, or accomplishing the mission. Military institutions view casualties from training accidents or testing new capabilities as a tragic but unavoidable part of the business of preparing for war. Militaries expect high performance from their forces, often while they are performing dangerous tasks, but militaries neither demand nor expect accident-free operations in most settings.[35](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn35) From 2006 to 2020, over 5,000 U.S. servicemembers were killed in non-war related accidents, the majority of which occurred within the United States. Accidents overall accounted for nearly 32 percent of U.S. servicemember deaths during this period, and even accounted for a significant portion of servicemember deaths in Iraq (19 percent) and Afghanistan (16 percent).[36](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn36) These accident rates are not unusual for the U.S. armed forces. This is business as usual. Accidents draw the attention of senior military and civilian officials when a spate of accidents occur in a short amount of time — such as a series of aircraft crashes,[37](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn37) ship collisions,[38](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn38) or training accidents.[39](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn39) Yet, as one report on naval accidents from 1945 to 1988 notes, “peacetime naval accidents are a fact of life.”[40](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn40) The same is true of military air and ground operations. Other nations’ militaries may do an even poorer job of managing risk when it comes to accidents than the U.S. military. For example, the Soviet/Russian submarine community has a much higher accident rate than the U.S. submarine community.[41](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn41) New technologies in particular present an increased risk of accidents, yet militaries may press ahead out of a desire to develop and field what they perceive to be a valuable capability. For example, the V-22 Osprey tiltrotor aircraft suffered four crashes during development, killing 30 U.S. servicemembers in total, yet the Defense Department continued development.[42](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn42) The V-22 program manager cited a rush to develop the technology as a factor in the accidents, stating, “Meeting a funding deadline was more important than making sure we’d done all the testing we could.”[43](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn43) Taking shortcuts on testing in particular appears to have been a factor in at least one fatal crash. According to a Government Accountability Office investigation of the V-22 program, “schedule pressures” led the program to conduct only 33 of 103 planned tests of an aerodynamic phenomenon called a “vortex ring state,”[44](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn44) a phenomenon that later caused an April 2000 crash that killed 19 servicemembers.[45](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn45) Absent competitive dynamics, militaries may be able to manage the challenges of fielding safe AI systems to a more-or-less satisfactory degree, albeit with some risk of an accident occurring. However, out of a desire to field AI capabilities ahead of competitors, militaries may be more willing to accept risk than they might otherwise be and to field systems that are prone to mishaps.[46](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn46) Similar competitive dynamics may have played a role in accidents with self-driving cars and commercial airline autopilot technology, as companies rushed to beat others to market.[47](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn47) These dynamics, while not an arms race, could lead militaries to engage in a “race to the bottom” on safety. This risk could become particularly acute in wartime. New technologies in particular present an increased risk of accidents, yet militaries may press ahead out of a desire to develop and field what they perceive to be a valuable capability. Managing these risks is challenging because assessing them can be difficult, especially when it comes to new technologies. Accident rates may be well-known for mature technologies, but they are unknown for technologies still in development. In the case of the V-22 Osprey development, for example, it is not as though the Defense Department knew that developing it would lead to multiple crashes and 30 fatalities but decided that achieving the capability was worth the cost. Engineers, testers, and program managers are flying in the dark when it comes to new technologies — that is, after all, the point of testing new systems. The concern is not only that organizations may take measured risks to field new capabilities, but also that institutional and bureaucratic imperatives may lead organizations to distort their own perceptions of risk, further contributing to accidents. This sociological phenomenon has been cited as a cause in the 1986 Space Shuttle Challenger explosion, for example.[48](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn48)

#### The US is Key – funding of TEVV signals to allies the importance of updated AI Testing

Scharre 2021 – Director of the Technology and National Security Program at the Center for a New American Security [Paul, 6/28/21,https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/, “Debunking the AI Arms Race Theory”, 6/18/22, LND]

The United States has done more to date than any other nation to advance norms surrounding the responsible use of AI, although the Department of Defense could be more deliberate in its approach to addressing the risks of military AI competition. U.S. defense leaders have focused primarily on implementing and demonstrating AI applications in an effort to prove AI’s value in military operations. This is understandable. The Department of Defense has many practical challenges to fielding AI systems even in relatively low-risk applications, including problems with data, computing infrastructure, contracting, and funding.[60](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn60) Nevertheless, it can and should do more to ensure that, as it competes in AI, it does so in a way that does not generate unnecessary risks or undermine international stability. Costly signals, such as investing in AI safety research or TEVV processes and infrastructure, may be even more effective in demonstrating to other nations that a state values fielding safe AI systems that operate under effective human control. The most important step that defense leaders could take in the near term to mitigate the risks stemming from AI competition would be to implement the necessary internal processes to ensure adequate TEVV of AI systems. A 2019 congressionally mandated independent assessment of the Defense Department’s AI efforts conducted by the RAND Corporation found that current TEVV processes were “nowhere close to ensuring the performance and safety of AI applications, particularly where safety-critical systems are concerned,” and issued recommendations for addressing this gap.[61](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn61) Similarly, a 2020 independent study led by Michèle Flournoy and Avril Haines identified a number of actionable steps that the department could take to improve its AI TEVV.[62](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn62) The National Security Commission on AI also concluded that “TEVV of traditional legacy systems is not sufficient” at providing adequate assurance for AI systems, and that “an entirely new type of TEVV will be needed.”[63](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn63) The report issued a raft of recommendations to improve AI TEVV and establish “justified confidence in AI systems.”[64](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn64) The Defense Department should adopt these reports’ recommendations to improve AI TEVV, a step that would not only bolster the safety of its AI systems but also their effectiveness. In addition to increasing resources and getting the attention of senior leadership, improving TEVV will require shifting how senior Defense Department leaders think about building robust, reliable, and effective AI systems. At times, senior defense leaders have characterized safety and ethics as an encumbrance the United States has to deal with that its adversaries do not.[65](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn65) While it is undoubtedly true that Russia and China are less concerned about ethics, safety, or international law than the United States, ensuring that military AI systems operate effectively and in a way that is consistent with human intent is a strength in the long run, even if more rigorous TEVV processes are needed in the near term to achieve that goal. Another element of mitigating the risks of military AI competition is with regard to how states characterize AI. U.S. messaging has been consistent and strong on the need for the responsible, lawful, ethical, and safe use of AI.[66](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn66) In their messaging, however, U.S. policymakers have frequently refrained from highlighting the risks of military AI competition, such as those outlined in this article. At times, they have emphasized a desire for speed that could feed into security dilemma concerns about a race-to-fielding that could undermine safety. In his 2020 Wired article, Will Roper wrote: “Our nation must wake up fast. The only thing worse than fearing AI itself is fearing not having it.”[67](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn67) While he called for using AI “safely and effectively,” his overriding emphasis was for the Defense Department to move faster.[68](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn68) Quite understandably, U.S. policymakers working to accelerate the adoption of AI in a slow-moving and sclerotic bureaucracy may be loath to portray the technology as immature, unready, or unreliable. Additionally, U.S. policymakers may fear that highlighting the risks of military AI could contribute to AI engineers balking at working with the military, even if such fears are unfounded.[69](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn69) Nevertheless, just as other emerging technologies such as computer networks opened up novel strategic challenges in the form of cyber operations, defense analysts should begin to think now about ways that AI may complicate international stability. A forthright acknowledgment of the risks of military AI competition is the first step toward a strategy of competing smartly while addressing those risks. Acknowledging the risks of military AI competition does not mean that the United States should refrain from adopting AI any more than acknowledging the stability risks of competing in space, cyberspace, or nuclear weapons necessitates unilateral disarmament in those spheres. America’s response to these risks should not be to refrain from AI competition, but rather to shape the character of the competition so that states are, at a minimum, mindful of these risks.

#### NATO cooperation on TEVV is necessary for AI safety and reliability because they have specialized staff and testing centres.

**Stanley-Lockman and Christie, 2021 - Innovation Officer in the Emerging Security Challenges Division in NATO and former Deputy Head of Innovation in NATO’s International Staff**  [Zoe and Edward, NATO Review, October “An Artificial Intelligence Strategy for NATO” <https://www.nato.int/docu/review/articles/2021/10/25/an-artificial-intelligence-strategy-for-nato/index.html>, BK]

For NATO, the common commitment to these principles has practical advantages as well, providing a coherent common basis for both NATO and Allies to design and develop AI applications while also supporting interoperability goals. As such, NATO can foster the necessary interlinkages between safety, security, responsible use, and interoperability. This can be seen across the principles. For instance, it is important to ensure that AI systems are adequately robust and reliable for their intended use, not only so that they can be expected to function in accordance with legal obligations, but also to mitigate the risks of the system’s defects or limitations being exploited by nefarious actors. Putting Principles into Practice These enduring principles are also foundational to the discussion and adoption of more detailed best practices and standards. Allies and NATO can leverage NATO’s consultative mechanisms and NATO’s specialised staff and facilities to work actively towards that goal. NATO’s own standardization and certification efforts can also be bolstered by coherence with relevant international standard-setting bodies, including for civilian AI standards. In addition to best practices and standards, these principles can also be operationalized via other mechanisms including review methodologies, risk and impact assessments, and security certification requirements like threat analysis frameworks and audits, among others. Further, NATO’s cooperative activities provide the basis to test, evaluate, validate, and verify (TEVV) AI-enabled capabilities in various different contexts. More specifically, NATO’s experience not only in operations, but also in trials, exercises, and experimentation provide several avenues in which Allies and NATO can test principles against intended use cases. This is further reinforced by NATO’s scientific and technical communities, which have worked on issues such as trust, human-machine and machine-machine interactions, and human-systems integration, among many others. In addition to these existing activities, the implementation of the AI Strategy will also benefit from connections with NATO’s forthcoming Defence Innovation Accelerator for the North Atlantic (DIANA). Allied Test Centres affiliated with DIANA could be used to fulfil the aims set out in the definitions of the principles. In the future, use of these Test Centres can help ensure that AI adoption and integration are tested for robustness and resilience. For example, to ensure that AI is Traceable, Reliable and Bias-mitigating, Test Centres could synthesise how AI systems perform in different simulated environments and on different testing data, or provide independent validation and verification to assess compliance with standards that focus on responsible engineering practices.

#### Public commitments to cooperation on AI Testing and Evaluation promote global norms that mitigate competitive pressures that cause accidents and miscalculation

Horowitz and Scharre 2021 - Director of the Emerging Capabilities Policy Office in the Office of the Under Secretary of Defense for Policy and  Vice President and Director of Studies at CNAS [Michael, Paul, January 12 2021, “AI and International Stability: Risks and Confidence- Building Measures”, <https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures>, Acc 6/18/22. M.A.]

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](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn57) 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. 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.

#### A US commitment to cooperation on standards for AI TEVV is a confidence building measure that reduces the risk of inadvertent conflicts and builds international support for safety

**Horowitz and Kahn, 2021 – Senior and Research Fellows at the Council on Foreign Relations** [Michael and Lauren, The Washington Quarterly 3-19-2021 “Leading in Artificial Intelligence through Confidence Building Measures”  [https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/0163660X.2021.2018794 acc on 6-21-2022](%20https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/0163660X.2021.2018794%20acc%20on%206-21-2022%20) TM]

Reducing Risks through Confidence-Building Measures Confidence-building measures could prove promising to reduce these risks of accidents, unintentional conflict, and inadvertent escalation associated with military applications of AI. The United Nations defines CBMs as “planned procedures to prevent hostilities, to avert escalation, to reduce military tension, and to build trust between countries.”16 In their simplest form, they are cross-national plans and procedures put in place to handle certain military and crisis events. Formal CBMs emerged from the 1975 Helsinki Final Act, the third meeting of the Conference on Security and Cooperation in Europe (CSCE), due to shared need and interest in preventing nuclear war.17 CBMs were specifically designed for European states’ “exchange of information, notification, and observation of major military activities,” with the aim of decreasing the risk of accidental war during the Cold War.18 Historically, CBMs work best when states fear that a lack of information could increase pressure to escalate: some mechanisms exist to mitigate lack of information; states have sufficient control over their militaries to ensure procedures are adhered to and orders are followed; and finally, when adversaries share at least some interests such as preventing accidental war. In the case of shared interests, going back to Thomas Schelling on Cold War strategic competition, cooperation between potential adversaries becomes more plausible when both sides can benefit and are incentivized to pursue the same end goal.19 We now describe four different CBMs that are applicable to the military AI arena: standardization, dialogues on AI safety, a statement on positive human control over nuclear weapons, and an autonomous incidents agreement. All focus specifically on issues surrounding accidents and miscalculation—areas where countries have shared interests in ensuring that military applications of AI do not lead to unintended conflict or escalation of tensions. Standard-Setting for Military AI, Including AI Safety Shared standards for safe uses of military AI could reduce the probability of accidents. Standards help to “herald market access, interoperability, and connectivity,” and while they might appear to be technical, standards can permit the “baking-in” of fundamental human rights protections when “ethical AI principles are translated into technical standards.”20 Furthermore, in the long run, standardization will facilitate coordination between states and ease interoperability with allied systems as well. Therefore, standard-setting for AI should start with something the United States is already committed to—following international humanitarian law. For example, including a commitment to strong TEVV requirements for AI enabled systems could further help create norms of responsible behavior surrounding the development of military uses of AI, or setting parameters on training data that can help reduce issues of bias.21 A critical element of standard setting could be commitments to follow expert-designed standards on AI safety in the TEVV process for military applications of AI.22 Key principles could include a testing framework that specifies error rates relative to existing systems, continual testing—even after deployment—rather than one-shot testing, testbeds to provide more realistic testing environments, and fail-safes built into systems in case something goes wrong. For the United States, standard-setting to enhance AI safety will only succeed if the United States does not share information that could directly help potential US adversaries. Sharing details about how the United States does TEVV, or how it is planning on doing TEVV for AI-enabled systems, could provide potential adversaries with information they would use to make their AI-enabled systems more effective. Thus, standard-setting and dialogue in these areas should focus more on themes and best practices rather than on policy details. A final element of common standards could be an emphasis on familiarizing operators with the limits of algorithms, an approach designed to reduce the potential for automation bias and even, on the other hand, automation aversion.23 An informal multilateral agreement could be proposed and opened for signature to all nations. If the United States leads, American allies and partners would be likely to sign on, both due to the impact of American leadership in shaping attitudes and the likely perception that following US-led principles would facilitate defense cooperation in this area.

#### DOD leadership is necessary to advance TEVV – it secures funding, provides visibility, and coordinates leadership. This signals the importance of reliability and safety.

Tarraf, Shelton, and Parker 2019 - Senior Information Scientist, Senior Engineer, and physical Scientist at the RAND Corporation [ Danielle C., William, Edward, et al RAND Cooperation, “The Department of Defense Posture for Artificial Intelligence”, [file:///Users/MiraAgarwal/Downloads/RAND\_RR4229%20(2).pdf](file:///C:\Users\MiraAgarwal\Downloads\RAND_RR4229%20(2).pdf), Acc 6/18/22, M.A.]

DoD should advance the science and practice of VVT&E of AI systems, working in close partnership with industry and academia. The JAIC, working closely with USD(R&E), USD(A&S), and operational test and evaluation, should take the lead in coordinating this effort both internally and with external partners. VVT&E is a critical consideration for DoD (see Chapter Three and Chapter Four), a significant challenge for the entire AI community (see Appendixes B and C), and a challenge that needs to be addressed, particularly as DoD looks to employ safety-critical AI systems. VVT&E has multiple facets that need to advance: from foundational research to establish the theory and science of V&V and the theoretical underpinnings of T&E, to the development of standards, guidelines, and engineering best practices for the VVT&E of systems being fielded and operated. Because of this, multiple entities within DoD and the government have stakes in VVT&E and roles to play. Indeed, DARPA and the service labs have significant roles to play in establishing the science and its foundations, while the JAIC has a role to play in setting guidelines and institutionalizing best practices in DoD. NIST also has a role to play in developing nationwide standards. Likewise, multiple entities outside government have a stake and a role, including academics researching the science and developing the foun- 74 The Department of Defense Posture for Artificial Intelligence dations and industry seeking to continue to leverage AI at scale.9 In an ideal world, the theory and the science of VVT&E would come first. In practice, AI systems are currently being deployed, and it is therefore important to develop and institute practical alternatives, including best practices and guidelines, while waiting for the theory and science to mature.10 Leadership and open cooperation among the many stakeholders, at various levels, are required to overcome the challenge of advancing the science and practice of VVT&E. It behooves DoD to focus significant efforts and resources in this direction; to take a leadership role nationally in view of its reliance on safety-critical systems; to seek engagements and partnerships with all stakeholders in industry and academia to advance the science and practice of VVT&E; and to keep Congress informed of the status of DoD’s activities, engagements, and partnerships in this regard. The JAIC should take the lead in spearheading coordination and keeping Congress informed of this national effort, both internally and with external partners. Recommendation T-5: All funded AI efforts should include a budget for AI VVT&E. This recommendation, in support of Recommendation S-2, is a forcing function that is relatively simple to implement and that can help ensure that the consideration of VVT&E is baked into the R&D of AI techniques and the design of AI solutions rather than considered as an afterthought further down the line. Although VVT&E during early R&D phases should be commonplace, we make this recommendation to explicitly reinforce its critical importance, to highlight the present lack of foundations for VVT&E in AI, and the importance of developing that science. 9 Indeed, as we note in Chapter Three and Appendix C, the Partnership on AI—a technology-industry consortium focused on establishing best practices for AI—appears to be moving toward establishing engineering guidelines for certification of ML. 10 Some of our industry interviewees indicated taking steps, internally and in industry partnerships, toward developing best practices (see Appendix C).

#### Cooperation on TEVV reduces the pressure for an Arms Race mentality – it builds confidence and encourages modelling.

Scharre 2021 – Director of the Technology and National Security Program at the Center for a New American Security [Paul, 6/28/21,https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/, “Debunking the AI Arms Race Theory”, 6/18/22, LND]

What can states do to avoid a race to the bottom on safety or an acceleration of the tempo of war beyond human control? In both instances, there are countervailing incentives that push against these trends. Militaries desire trusted systems on the battlefield and effective control over their own forces. There are several actions states can take to strengthen these incentives toward ensuring robust, secure, and controllable AI systems in their own institutions, as well as those of other countries. First, states should invest in adequate internal processes to test, evaluate, verify, and validate AI systems, in order to ensure that the systems they are fielding are robust and secure.[55](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/#_ftn55) States should similarly strengthen their internal processes — doctrine, training, system design and testing, human-machine interfaces, etc. — to retain effective human control over combat operations.[56](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/#_ftn56) Second, states should take specific measures to encourage other states to do likewise in order to mitigate perverse incentives to cut corners on testing or cede human control to machines where it would otherwise not be preferable. Such actions could include voluntary transparency measures about TEVV processes, although there will no doubt be technical details that states are unwilling to share. States could also communicate the importance of AI safety and reliability and of maintaining human control over combat operations, both publicly and in international diplomatic channels such as the Convention on Certain Conventional Weapons.[57](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/#_ftn57) For example, in 2020 the U.S. Department of Defense released a set of ethical principles for AI.[58](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/#_ftn58) Costly signals, such as investing in AI safety research or TEVV processes and infrastructure, may be even more effective in demonstrating to other nations that a state values fielding safe AI systems that operate under effective human control. States should avoid messages that may incentivize other states to take shortcuts on these processes, such as claims of an “AI arms race.” Lastly, states should explore opportunities to take cooperative measures that might mitigate these risks. Getting adversaries to cooperate is inherently challenging, but states have succeeded in the past in regulating the conduct of war in a variety of ways to mitigate mutual harm. Joint declarations, codes of conduct, or confidence-building measures may help to reduce the greatest dangers of AI competition and encourage states to adopt AI responsibly.[59](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/#_ftn59)

#### Reforming the DOD TEVV system will allow coordination and continuous lifecycle testing – these are essential for AI

**Flournoy, Haines et al 2020 - former Under Secretary of Defense for Policy and Director of National Intelligence** [Michèle and Avril, Gabrielle Cheftiz Special Assistant to the Under Secretary of Defense for Policy October, “Building Trust through Testing Adapting DOD’s Test & Evaluation, Validation & Verification (TEVV) Enterprise for Machine Learning Systems, including Deep Learning Systems” https://cset. georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf Acc 6/23/22 JZ]

For ML/DL, the Defense Department will need to replace its classic approach to TEVV of formulating a T&E Master Plan for a given capability up front with a more automated, iterative, and continuous approach to testing in line with DevSecOps. Assuring that ML/DL systems function as expected and do not engage in behaviors outside their intended use and operational parameters will require testing across the system’s entire life cycle—from development to operational deployment to sustainment. It will also require new methods for capturing lessons learned and integrating these into iterative development and testing. Because of the difficulty predicting and binding system performance, one should consider every deployment of an ML/DL system as an experiment and opportunity to collect data and insight on performance.16 To support this approach, DOD will need to expand coordination between program managers and testers to ensure testing milestones are built in throughout the acquisition program. Program managers often see TEVV as an obstacle to be surmounted at the end of the development process, rather than a necessary process to be integrated throughout the development life cycle. Of course, this problem is not unique to ML/DL programs, but it is exacerbated when it comes to emerging technologies that do not yet have established testing methodologies. Further, the Pentagon needs to invest in and scale automated TEVV capabilities for operational platforms, such as the Navy’s automatic test and retest program,17 which will significantly speed up the testing process.

## Contention One Extensions

### Extend Inherency – DOD

#### DOD has not committed to improve TEVV – this prevents the US from adopting safe AI

**Flournoy, Haines et al 2020 - former Under Secretary of Defense for Policy and Director of National Intelligence** [Michèle and Avril, Gabrielle Cheftiz Special Assistant to the Under Secretary of Defense for Policy October, “Building Trust through Testing Adapting DOD’s Test & Evaluation, Validation & Verification (TEVV) Enterprise for Machine Learning Systems, including Deep Learning Systems” https://cset. georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf Acc 6/23/22 JZ]

The Defense Department needs to reform its existing testing and verification system—its methods, processes, infrastructure, and workforce—in order to help decision-makers and operators understand and manage the risks of developing, producing, operating, and sustaining AI-enabled systems. Several DOD reports and policy documents identify TEVV as a barrier to AI adoption and call for increased research into new methodologies, including the Pentagon’s AI Ethics Principles1 and AI Strategy,2 which states, “we will invest in the research and development of AI systems that are resilient, robust, reliable, and secure; we will continue to fund research into techniques that produce more explainable AI; and we will pioneer approaches for AI test, evaluation, verification, and validation.” However, DOD has yet to translate this stated goal into a real plan of action. Advancing the Defense Department’s TEVV enterprise for ML/DL systems is critical for several reasons. First, developing an effective TEVV approach that is sufficiently predictive of performance is critical to building the trust in these systems necessary to deploy and leverage these capabilities at scale. The United States has already seen this dynamic with nuclear power, for example, where lost trust in the technology has prevented policymakers from harnessing nuclear power for clean energy

#### The DOD AI Strategy fails due to a lack of a value metric to assess progress

Tarraf, Shelton, and Parker 2019 - Senior Information Scientist, Senior Engineer, and physical Scientist at the RAND Corporation [ Danielle C., William, Edward, et al RAND Cooperation, “The Department of Defense Posture for Artificial Intelligence”, [file:///Users/MiraAgarwal/Downloads/RAND\_RR4229%20(2).pdf](file:///C:\Users\MiraAgarwal\Downloads\RAND_RR4229%20(2).pdf), Acc 6/18/22, M.A.]

DoD lacks baselines and metrics in conjunction with its AI vision. Baselines and metrics are important for two reasons. First, they are a means of assessing and enhancing progress toward DoD’s vision and of managing expectations (see section “Adoption and Scaling of Unmanned Aircraft Systems” in Appendix D and section “AI History in DoD” in Appendix D). Second, metrics are needed to demonstrate value and secure continued leadership support as progress is made in institutional transformation (see section “Industry: Organization” in Appendix C); this is particularly important in ensuring continued support at the highest levels of decisionmaking, both within and outside DoD. The summary of the 2018 DoD AI strategy lays out a vision for institutional transformation through AI, the assignment of an organization to be a focal point for that transformation, and a set of activities associated with the vision. However, the strategy does not articulate baselines, metrics, or quantifiable measures of value or success.

### Extend Inherency – TEVV

**There is not enough funding for TEVV now – the 2021 Defense Authorization proves**

**Arnold and Toner, 2021 – Center for Security and Emerging Threats** [Zachary and Helen, July, CSET Policy Brief. “AI Accidents: An Emerging Threat What Could Happen and What to Do” https://cset. georgetown.edu/wp-content/uploads/CSET-AI-Accidents-An-Emerging-Threat.pdf Acc 6/7/22 TA]

4. What to do AI accidents are already happening.45 If we do not act, they will become far more common and destructive. Improvements in AI technology and bottom-up market pressure from consumers may help make AI safer and less accident-prone, but they are unlikely to do enough on their own. Policy has an essential role to play. Smart policy can drive research into less accident-prone AI technologies, bring the AI community together to reduce risks, and provide incentives for private actors to use AI safely, saving lives and livelihoods in the future. Today, the policy effort around AI safety and accident risk is only beginning. There are several federal actions that will be central to any policy agenda. These include: ● Facilitate information sharing about AI accidents and near misses. To make AI safer, we need to know when and how it fails. In many other technological domains, shared incident reporting contributes to a common base of knowledge, helping industry and government track risks and understand their causes. Models include the National Transportation Safety Board’s database for aviation incidents and the public-private cyber intelligence platforms known as Information Sharing and Analysis Centers.46 The government should consider creating a similar repository for AI accident reports. As part of this effort, policymakers should explore different ways of encouraging the private sector to actively disclose the details of AI accidents. For example, the government could offer confidentiality protections for sensitive commercial information in accident reports, develop common standards for incident reporting, or even mandate disclosure of certain types of incidents.47 ● Invest in AI safety research and development (R&D). The federal government and private industry invest billions in AI R&D every year, but almost none of this funding goes to AI safety research.48 Federal R&D funding has led to critical safety and security innovations in many other contexts, from cryptographic protocols that enable secure communication to the sensors behind modern airbags.49 It will be crucial to make similar investments in AI safety, including research aiming to solve the problems of robustness, specification, and assurance described above, as well as investing in the development of AI engineering as a more rigorous discipline.50 The 2021 National Defense Authorization Act (NDAA) made a good start in this direction by including provisions calling for the National Science Foundation and the Department of Energy to invest in research into “trustworthy AI.”51 However, it remains to be seen how much funding will actually be invested in these areas.

#### Current TEVV is not enough – assurance is an afterthought for most AI programs

Batarsheh et al. 2021 - Professor of Electrical and Computer Engineering, Virginia Polytechnic Institute [Feras, Laura Freeman Director, Intelligent Systems Division – National Security Institute Research Associate Professor, Chih Hao Huang, Data Analyst at Turing Research. April 26 “A survey on artificial intelligence assurance” https://link-springer-com.proxy.lib.umich.edu/article/10.1186/s40537-021-00445-7#Sec8]

Applications of AI such as facial recognition using deep learning have become commonplace. Deep learning models are often exposed to adversarial inputs (such as deep-fakes), thus limiting their adoption and increasing their threat [145]. Unlike conventional software, aspects such as explainability (unveiling the blackbox of AI models) dictate how assurance is performed and what is needed to accomplish it. Unfortunately however, similar to the software engineering community’s experience with testing, ensuring a valid and verified system is often an afterthought. Some of the classical engineering approaches would prove useful to the AI assurance community, for instance, performing testing in an incremental manner, involving users, and allocating time and budget specifically to testing, are some main lessons that ought to be considered. A worthy recent trend that might aid majorly in assurance is using AI for testing AI (i.e., deploying intelligence methods for the testing and assurance of AI methods). Additionally, from a user’s perspective, recent growing questions in research that are relevant to assurance pose the following concerns: how is learning performed inside the blackbox? How is the algorithm creating its outcomes? Which dependent variables are the most influential? Is the AI algorithm dependable, safe, secure, and ethical? Besides all the previously mentioned assurance aspects, we deem the following foundational concepts as highly connected, worthy of considering by developers and AI engineers, and essential to all forms of AI assurance: (1) Context: refers to the scope of the system, which could be associated with a timeframe, a geographical area, specific set of users, and any other system environmental specifications (2) Correlation: the amount of relevance between the variables, this is usually part of exploratory analysis, however, it is key to understand which dependent variables are correlated and which ones are not, (3) Causation: the study of cause and effect; i.e., which variables directly cause the outcome to change (increase or decrease) in any fashion, (4) Distribution: whether a normal distribution is assumed or not. Data distribution of the inputted dependent variables can dictate which models are best suited for the problem at hand, and (5) Attribution: aims at allocating the variables in the dataset that have the strongest influence on the outcomes of the AI algorithm.

#### Current TEVV fails – new techniques needed to address complexities in AI tech

Scharre, 2016 -- Vice President and Director of Studies at CNAS [Paul, Feb 2016, Center for New American Security, “Autonomous weapons and operational risk”, https://s3.us-east-1.amazonaws.com/files.cnas.org/documents/CNAS\_Autonomous-weapons-operational-risk.pdf?mtime=20160906080515&focal=none, Acc 6/21/22, M. A.]

System failures: System failures occur not from the breakdown of any one given part, but from unanticipated interactions between elements of a system. Verifying all possible combinations of the internal workings of the system becomes increasingly difficult as the system’s complexity increases. o A recent report on autonomy by the U.S. Air Force Office of the Chief Scientist highlighted the need for new techniques for the verification and validation of autonomous software as a “critical” issue for the Air Force. “Traditional methods … fail to address the complexities associated with autonomy software … There are simply too many possible states and combination of states to be able to exhaustively test each one.”14 11 Emergent behavior could come from individual systems or from groups or swarms of simpler systems coordinating their actions together, similar to ants, termites, or bees. For more on military applications of swarming, see Paul Scharre, “Robotics on the Battlefield – Part II: The Coming Swarm,” Center for a New American Security, October 2014,

### Extend No Arms Race

#### Claims of an AI arms race are overblown – the AI adoption process is routine continuation of modernization

Horowitz and Scharre, 2021 - Director of the Emerging Capabilities Policy Office in the Office of the Under Secretary of Defense for Policy and  Vice President and Director of Studies at CNAS [Michael, Paul, January 12 2021, “AI and International Stability: Risks and Confidence- Building Measures”, <https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures>, Acc 6/18/22. M.A.]

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](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn2) 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](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn3) 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](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn4) 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.

#### AI is not an Arms Race – spending is not enough to warrant the Arms Race title – it is just military modernization

Scharre 2021 – Director of the Technology and National Security Program at the Center for a New American Security [Paul, 6/28/21,https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/, “Debunking the AI Arms Race Theory”, 6/18/22, LND]

The scale of military AI spending, at least at present, is nowhere near large enough to warrant the title of “arms race.” Of course, AI can also be used for weapons. Militaries around the world are actively working to adopt AI to improve their military capabilities. Yet the militarization of AI does not, at present, meet the traditional definition of an arms race, despite the rhetorical urgency of many national leaders. Michael D. Wallace, in his 1979 article “Arms Races and Escalation,” defined an arms race as “involving simultaneous abnormal rates of growth in the military outlays of two or more nations” resulting from “the competitive pressure of the military itself, and not from domestic forces exogenous to this rivalry.” Wallace further stated that the concept of an arms race only applied “between nations whose foreign and defense policies are heavily interdependent” and who have “roughly comparable” capabilities.[11](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/#_ftn11) AI is being adopted by many countries around the globe.[12](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/#_ftn12) Arguably at least some of the dyads, such as the United States and China, meet Wallace’s definition in terms of being nations with “roughly comparable” capabilities, locked in competition, “whose foreign and defense policies are heavily interdependent.” However, AI fails the arms race test in the critical area of spending. Wallace distinguished arms races from the normal behavior of states to improve their military forces. A state that adopts a new technology and modernizes its military forces is not automatically in an arms race, under Wallace’s definition, even if the modernization is aimed at competition with another country. The decisive factor in qualifying as an arms race, according to Wallace, is the rate of growth in defense spending. Wallace characterized arms races as resulting in abnormally large growth rates in defense spending, beyond the historical average of 4 to 5 percent annual growth (in real dollars). In an arms race, annual growth rates are above 10 percent or even as high as 20 to 25 percent.[13](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/#_ftn13) Other scholars define arms races using different quantitative thresholds — and some definitions lack clear quantitative thresholds at all — but the existence of rapid increases in defense spending or military forces above normal levels is a common criterion in the scholarly literature on arms races.[14](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/#_ftn14) Arms races result in situations in which two or more countries are locked in spiraling defense spending, grabbing ever-greater shares of national treasure often with little to no net gain in relative advantage over the other. Classic historical examples include the Anglo-German naval arms race prior to World War I and the U.S.-Soviet nuclear arms race during the Cold War. Military AI spending today clearly does not meet these criteria of abnormally large growth rates in defense spending. AI defense spending is difficult to calculate due to the general-purpose nature of AI technology. Unlike ships or ballistic missiles, AI systems cannot be easily counted. Nevertheless, even crude estimates of defense spending show that military AI investments are nowhere near large enough to constitute an arms race. An independent estimate by Bloomberg Government of U.S. defense spending on AI identified $5 billion in AI-related research and development in fiscal year 2020, or roughly 0.7 percent of the Department of Defense’s over $700 billion budget.[15](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/#_ftn15) The scale of military AI spending, at least at present, is nowhere near large enough to warrant the title of “arms race.” (Adding in private sector spending, which constitutes the bulk of AI investment, would lead to larger figures but would further belie the claim of an “arms” race since most private sector AI investment is not in weapons.)

**There is no global sprint or arms race – “Arms Race” claims overexaggerate normal military behavior and ignore actual spending.**

**Perry and Scharre, 2020 - Project Coordinator at the Future of Life Institute and Director of the Technology and National Security Program at the Center for a New American Security** [Lucas and Paul; March 16 "AI Alignment Podcast: On Lethal Autonomous Weapons with Paul Scharre," <https://futureoflife.org/2020/03/16/on-lethal-autonomous-weapons-with-paul-scharre/?cn-reloaded=1> Acc 2/2/21 TA]

Paul Scharre: If there’s an arms race, it’s a very strange one because no one is building the weapons. We see militaries advancing in robotics and autonomy, but we don’t really see sort of this rush to build autonomous weapons. I struggle to point to any programs that I’m aware of in militaries around the globe that are clearly oriented to build fully autonomous weapons. I think there are lots of places where much like these incremental advancements of autonomy in cars, you can see more autonomous features in military vehicles and drones and robotic systems and missiles. They’re adding more autonomy. And one might be violently concerned about where that’s going. But it’s just simply not the case that militaries have declared their intention. We’re going to build autonomous weapons, and here they are, and here’s our program to build them. I would struggle to use the term arms race. It could happen, maybe worth a starting line of an arms race. But I don’t think we’re in one today by any means. It’s worth also asking, when we say arms race, what do we mean and why do we care? This is again, one of these terms, it’s often thrown around. You’ll hear about this, the concept of autonomous weapons or AI, people say we shouldn’t have an arms race. Okay. Why? Why is an arms race a bad thing? Militaries normally invest in new technologies to improve their national defense. That’s a normal activity. So if you say arms race, what do you mean by that? Is it beyond normal activity? And why would that be problematic? In the political science world, the specific definitions vary, but generally, an arms race is viewed as an increase in defense spending overall, or in a particular technology area above normal levels of modernizing militaries. Now, usually, this is problematic for a couple of reasons. One could be that it ends up just in a massive national expenditure, like during the case of the Cold War, nuclear weapons, that doesn’t really yield any military value or increase anyone’s defense or security, it just ends up net flushing a lot of money down the drain. That’s money that could be spent elsewhere for pre K education or healthcare or something else that might be societally beneficial instead of building all of these weapons. So that’s one concern. Another one might be that we end up in a world that the large number of these weapons or the type of their weapons makes it worse off. Are we really better off in a world where there are 10s of thousands of nuclear weapons on hair-trigger versus a few thousand weapons or a few hundred weapons? Well, if we ever have zero, all things being equal, probably fewer nuclear weapons is better than more of them. So that’s another kind of concern whether in terms of violence and destructiveness of war, if a war breakout or the likelihood of war and the stability of war. This is an A in an area where certainly we’re not in any way from a spending standpoint, in an arms race for autonomous weapons or AI today, when you look at actual expenditures, they’re a small fraction of what militaries are spending on, if you look at, say AI or autonomous features at large.

#### AI is different from other Arms Races – it is not exclusively military, which means there is less advantage to moving first

Horowitz, 2018 – Professor of Political Science at UPen**n** [Michael, September “The Algorithms of August: The AI arms race won't be like previous competitions, and both the United States and China could be left in the dust,” https://foreignpolicy.com/2018/09/12/will-the-united-states-lose-the-artificial-intelligence-arms-race/ 6/18/22 MD]

AN [ARTIFICIAL INTELLIGENCE](https://go-gale-com.proxy.lib.umich.edu/ps/i.do?p=AONE&u=umuser&id=GALE%7CA556838648&v=2.1&it=r) ARMS RACE IS COMING. It is Unlikely to play out in the way that the mainstream media suggest, however: as a faceoff between the United States and China. That's because AI differs from the technologies, such as nuclear [weapons](https://go-gale-com.proxy.lib.umich.edu/ps/i.do?p=AONE&u=umuser&id=GALE%7CA556838648&v=2.1&it=r) and battleships, that have been the subject of arms races in the past. After all, AI is software--not hardware. Because AI is a general purpose [technology](https://go-gale-com.proxy.lib.umich.edu/ps/i.do?p=AONE&u=umuser&id=GALE%7CA556838648&v=2.1&it=r)--more like the combustion engine or electricity than a weapon--the competition to develop it will be broad, and the line between its civilian and military uses will be blurry. There will not be one exclusively military AI arms race. There will instead be many AI arms races, as countries (and, sometimes, violent nonstate actors) develop new algorithms or apply private sector algorithms to help them accomplish particular tasks. In North America, the private sector invested some $15 billion to $23 billion in AI in 2016, according to a McKinsey Global Institute report. That's more than 10 times what the U.S. government spent on unclassified AI programs that same year. The largest share came from companies such as Google and Microsoft, as well as a number of smaller private firms, not from government-funded defense research. This reverses the dynamic from the Cold [War](https://go-gale-com.proxy.lib.umich.edu/ps/i.do?p=AONE&u=umuser&id=GALE%7CA556838648&v=2.1&it=r), when government investments led to private sector innovation and produced technologies such as GPS and the internet. China says it already holds more than 20 percent of patents in the field and plans to build its AI sector to be worth $150 billion by 2030. But while Beijing and Washington are the current leaders in this race, they are not the only competitors. Countries around the world with advanced technology sectors, from Canada to France to Singapore, also have the potential to make great strides in AI (or build on lower-level advances made by others). While this diffusion means that many more countries will have a stake in the regulation of AI, it also means that many more governments will have incentives to go it on their own. UNLIKE THE DEVELOPMENT of a Stealth bomber, which has only military applications, basic AI research has both military and civilian uses, which makes it much harder to keep research secret and thereby sustain a large first-mover advantage. The dual-use character of many developments in AI creates an incentive to promote their release and spread to the general public. That means companies can co-opt advances made by market leaders--especially lower-level advances that do not require significant computing hardware.

**Autonomous weapons won’t cause an Arms Race – just because countries develop them doesn’t mean that they are in a hostile race.**

**Horowitz, 2019 - Professor of Political Science, University of Pennsylvania** [Michael C. May 2“When Speed Kills: Autonomous Weapon Systems, Deterrence, and Stability” https://ssrn.com/abstract=3348356]

Arms Races or Proliferation? It is important to distinguish arms races from the proliferation of military technologies. The scientists in the example above compared an LAWS arms race to the spread of Kalashnikovs, but there was never an arms race in Kalashnikovs – they just spread rapidly because they were cheap, easy to produce, and useful.39 Huntington’s classic work on arms races distinguishes between proliferation and arms races. Countries can increase their arms acquisitions due to an “absolute need” that exists “regardless of the actions of other states”, or for economic reasons.40 Huntington argues that many things described as arms races are simply general buildups due to military necessity (or for economic reasons), rather than a specific buildup due to a particular disagreement between states.41 For example, after their debut in World War I, countries around the world acquired tanks in the 1920s and 1930s. Tanks then became a critical part of ground warfare in World War II and subsequent conventional ground combat operations. Yet few would call the spread of tanks in the 1920s and 1930s an arms race. It was simply proliferation that was not possible to stop, as large-caliber guns, caterpillar tracks, and the combustion engine were available to countries around the world.42 To the extent that those concerned about an arms race in autonomous weapon technologies are actually concerned with proliferation, some degree of proliferation may be inevitable simply due to the underlying factors involved in the production of LAWS. As Horowitz argues, military capabilities diffuse faster when, on the technology acquisition side, there is underlying commercial demand and the unit costs are low.43 Commercial markets are already driving the integration of artificial intelligence into several areas of the US and global economies, through deep learning and machine learning applications.44 From Google search to Macy’s shopping assistance for customers, artificial intelligence is increasingly embedded in commercial sectors of modern society.45 Narrow applications of AI could become increasingly integrated into most economic sectors, with many algorithms, once developed, now available to many actors.46 Export controls are unlikely to stop basic narrow AI capabilities from spreading, even if more advanced applications are beyond the capacity of most companies and governments; government regulations generally lag emerging technologies. Finally, AI innovation is occurring around the world, not just in the United States or even the West. Machine learning capabilities designed for commercial purposes could also have spillovers with useful applications to the military realm. This would reverse the Cold War dynamic in the West, where US civilian economic innovations such as GPS often spun out of military development programs.47 Artificial intelligence, as described above, is more an enabler such as the combustion engine than a weapon. It is therefore different than a platform like stealth technology, which really only has military purposes.

#### Framing AI as an arms race assumes that AI is only a weapon – it is beneficial in many other contexts

Scharre 2021 – Director of the Technology and National Security Program at the Center for a New American Security [Paul, 6/28/21,https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/, “Debunking the AI Arms Race Theory”, 6/18/22, LND]

As Heather Roff has written, the arms race framing “misrepresents the competition going on among countries.”[5](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/#_ftn5) To begin with, AI is not a weapon. AI is a general-purpose enabling technology with myriad applications. It is not like a missile or a tank. It is more like electricity, the internal combustion engine, or computer networks.[6](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/#_ftn6) General-purpose technologies like AI have applications across a range of industries. Wired magazine co-founder Kevin Kelly has argued that it “will enliven inert objects, much as electricity did more than a century ago. Everything that we formerly electrified we will now cognitize.”[7](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/#_ftn7) Nations may very well be in a technology race to adopt AI across a range of industries. AI will help to improve economic productivity and, by extension, economic and military power. During the industrial revolution, early adopters of industrial technology significantly increased their national power. From 1830 to 1890, Britain and Germany, which were both early industrializers, more than doubled their per capita gross national product while Russia, which lagged in industrialization, increased its per capita gross national product by a mere 7 percent over that 60-year period.[8](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/#_ftn8) These technological advantages led to increased economic and military power, most notably for Europe relative to the rest of the world. In 1790, Europe (collectively), China, and India (including what is now Pakistan and Bangladesh) held roughly the same shares of global manufacturing output, with Europe and India each holding about one-quarter of global manufacturing output and China holding roughly one-third. They all had approximately equivalent levels of per capita industrialization at that time. But the industrial revolution skyrocketed European economic productivity. By 1900, Europe collectively controlled 62 percent of global manufacturing output, while China held only six percent and India less than two percent. These economic advantages translated into military power. By 1914, Europeans occupied or controlled over 80 percent of the world’s land surface.[9](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/#_ftn9) Being ahead of the curve in adopting AI is likely to lead to significant national advantages. Although AI can increase military capabilities, the more consequential advantages over the long term may come from non-military AI applications across society. Long-term benefits from AI could include increased productivity, improved healthcare outcomes, economic growth, and other indicators of national well-being. Increasing productivity is especially significant because it has a compounding effect on economic growth. Over the long term, technological progress is the main driver of economic growth.

### AT DOD AI Strategy Solves Now

#### The US has not committed to AI safety – there were promises but no concrete action – this creates the risk of accidental war.

Lewis 2019 - project lead for the DOD’s Joint Lessons Learned studies [Larry, “AI Safety: Charting out the High Road”, War on the Rocks Dec 19, https://warontherocks.com/2019/12/ai-safety-charting-out-the-high-road/ LMSi]

Concerns about how governments can leverage AI also extend to the waging of war. Two major concerns about the application of AI to warfare are ethics (is it the right thing to do?) and safety (will civilians or friendly forces be harmed, or will the use of AI lead to accidental escalation leading to conflict?). With the United States, Russia, and China all signaling that AI is a transformative technology central to their national security strategy, with their militaries planning to move ahead with military applications of AI quickly, should this development raise the same kinds of concerns as China’s use of AI against its own population? In an era where Russia targets hospitals in Syria with airstrikes in blatant violation of international law, and indeed of basic humanity, could AI be used unethically to conduct war crimes more efficiently? Will AI in war endanger innocent civilians as well as protected entities such as hospitals? BECOME A MEMBER To be clear, any technology can be misused in war. A soldier could use a rock to commit a war crime. A simple, low-tech land or sea mine can be used indiscriminately and endanger civilians if it is used in the wrong way. A transformative technology like AI can be used responsibly and safely, or it could fuel a much faster race to the bottom. The United States has declared it will take the high road with military applications of AI. For example, the Department of Defense AI strategy has “AI ethics and safety” as one of the its fundamental lines of effort. And this is not an empty promise: The Defense Innovation Board just released its principles for ethical military use of AI, marking a year-long, deliberate initiative drawing in AI experts, ethicists, and the general public. By this laudable effort, the United States has shown leadership in the responsible and principled use of this technology in war. But there is something missing: The AI strategy commitment was to ethics and safety. To date, the Department of Defense has not shown a similar, concerted focus on AI safety. Despite commitments made to the international community and in its own AI strategy, the Pentagon has done little to act on promises to address safety risks unique to the technology of AI or to use AI to enhance safety in conflict. My recent research has shown that this inaction creates risks to those on the battlefield, for civilians and combatants alike, and increases the likelihood of accidental escalation and conflict. In an era where the technology of AI can so easily be exploited by governments to violate the principles of humanity, the United States can demonstrate the high road is possible, but to do so it needs to keep its promises: to address safety risks intrinsic to AI and to search for ways to use AI for good.

### AT DODD 3000.09 Solves Now

**DODD 3000.19 does not implement TEVV – it has vague mandates for testing**

**Flournoy, Haines et al 2020 - former Under Secretary of Defense for Policy and Director of National Intelligence** [Michèle and Avril, Gabrielle Cheftiz Special Assistant to the Under Secretary of Defense for Policy October, “Building Trust through Testing Adapting DOD’s Test & Evaluation, Validation & Verification (TEVV) Enterprise for Machine Learning Systems, including Deep Learning Systems” https://cset. georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf Acc 6/23/22 JZ]

The Department has developed policy and ethical guidance on autonomous systems and AI, but these guidelines have yet to be translated into TEVV implementation guidance. DoD has established important foundational policy guidance for the use of autonomous systems and artificial intelligence with DoD Directive 3000.09 on Autonomy in Weapon Systems and the Defense Innovation Board’s AI Ethics principles, adopted by DoD in February 2020. These policy documents have important implications for testing and evaluating of ML/DL. For example, 3000.09 calls for systems to go through “rigorous hardware and software [verification and validation] and realistic system developmental and operational T&E, including analysis of unanticipated emergent behavior resulting from the effects of complex operational environments on autonomous or semiautonomous systems.” It also states that interfaces should be “readily understandable to trained operators,”22 making explainability an important component of implementing this policy. Meanwhile, the AI Ethics Principles commit DoD to develop and deploy AI that is traceable (including with transparent and auditable methodologies, data sources, and design procedure and documentation), reliable (explicit, well-defined uses, with the safety, security, and effectiveness of such capabilities subject to testing and assurance across their entire life cycle), and governable (with the ability to detect and avoid unintended consequences, as well as disengage or deactivate deployed systems that demonstrate unintended behavior). These policies are an important start and provide a useful framework for driving TEVV for AI and autonomous systems. However, these goals are incredibly broad, and many are currently technologically infeasible, given existing testing methodologies. DoD needs to develop TEVV implementation guidance for both 3000.09 and AI ethics principles. In particular, these principles must inform ML/DL design and be incorporated into the standards, specifications, and requirements against which systems will be tested.

### AT NATO Guidelines Solve Now

#### NATO’s guidelines are just principles – they do not implement policies that back them up – that is the plan.

**Stanley-Lockman and Christie, 2021 - Innovation Officer in the Emerging Security Challenges Division in NATO and former Deputy Head of Innovation in NATO’s International Staff**  [Zoe and Edward, NATO Review, October “An Artificial Intelligence Strategy for NATO” <https://www.nato.int/docu/review/articles/2021/10/25/an-artificial-intelligence-strategy-for-nato/index.html>, BK]

Adopting AI in the defense and security context also calls for effective and responsible governance, in line with the common values and international commitments of Allied nations. To that end, Allied governments have committed to Principles of Responsible Use as a key component of NATO’s AI Strategy. Allies and NATO commit to ensuring that the AI applications they develop and consider for deployment will be in accordance with the following six principles: Lawfulness: AI applications will be developed and used in accordance with national and international law, including international humanitarian law and human rights law, as applicable. Responsibility and Accountability: AI applications will be developed and used with appropriate levels of judgment and care; clear human responsibility shall apply in order to ensure accountability. Explainability and Traceability: AI applications will be appropriately understandable and transparent, including through the use of review methodologies, sources, and procedures. This includes verification, assessment and validation mechanisms at either a NATO and/or national level. Reliability: AI applications will have explicit, well-defined use cases. The safety, security, and robustness of such capabilities will be subject to testing and assurance within those use cases across their entire life cycle, including through established NATO and/or national certification procedures. Governability: AI applications will be developed and used according to their intended functions and will allow for: appropriate human-machine interaction; the ability to detect and avoid unintended consequences; and the ability to take steps, such as disengagement or deactivation of systems, when such systems demonstrate unintended behaviour. Bias Mitigation: Proactive steps will be taken to minimise any unintended bias in the development and use of AI applications and in data sets. Having agreed to adopt these mutually reinforcing principles, the task now turns to translating them into principled action. As such, NATO’s role in operationalising these principles will involve efforts that similarly tackle different aspects of the technology’s lifecycle. Building the principles of responsible use into the front end of AI development is important because, the later they are considered, the harder it may be to ensure they are upheld. Ensuring a full life-cycle approach also depends on multi-stakeholder engagement because responsibility is diffused amongst the policymakers, designers, developers, and testers, as well as operational end users that engage in AI development and use. For NATO, this is relevant because various entities play an active role in AI integration, and because the Alliance can encourage coherence with national AI developments.

### AT JAIC Solves Now

#### JAIC does not solve now – it lacks authority for budgeting.

Tarraf, Shelton, Parker 2019 - Senior Information Scientist, Senior Engineer, and physical Scientist at the RAND Corporation [ Danielle C., William, Edward, et al RAND Cooperation, “The Department of Defense Posture for Artificial Intelligence”, [file:///Users/MiraAgarwal/Downloads/RAND\_RR4229%20(2).pdf](file:///C:\Users\MiraAgarwal\Downloads\RAND_RR4229%20(2).pdf), Acc 6/18/22, M.A.]

The JAIC lacks the authorities to carry out its present role. At its core, the JAIC’s overarching mission can be distilled to this: Scale AI and its impact across DoD. This mission and its present scope—as defined by the summary of the DoD AI strategy and the memo establishing the JAIC—are extensive, while the JAIC’s current authorities are limited. In particular, the JAIC is expected to synchronize DoD AI activities and coordinate AI initiatives totaling more than $15 million annually. It is unclear whether the JAIC has any mechanisms for enforcing these directives, because it does not have the authorities to direct investments or to halt programs or activities that are deemed to be misaligned with DoD’s strategy (a fact we learned through multiple DoD interviews). In short, the JAIC does not have directive or budget authorities, and that critically limits its ability to synchronize and coordinate DoD-wide AI activities to enact change. Currently, it can catalogue these activities, but it is unclear how doing so would help scale AI across DoD. Of course, that assumes that what constitutes an AI activity is known. However, it is not currently clear how the determination of what constitutes an AI initiative or activity is made, by whom, and whether that determination is consistent across DoD.8

#### Enhancing the JAIC is necessary for solvency – they coordinate across different services and maintain focus on the message and mission for AI

Tarraf, Shelton, Parker 2019 - Senior Information Scientist, Senior Engineer, and physical Scientist at the RAND Corporation [ Danielle C., William, Edward, et al RAND Cooperation, “The Department of Defense Posture for Artificial Intelligence”, [file:///Users/MiraAgarwal/Downloads/RAND\_RR4229%20(2).pdf](file:///C:\Users\MiraAgarwal\Downloads\RAND_RR4229%20(2).pdf), Acc 6/18/22, M.A.]

Organization We begin by revisiting DoD’s vision for AI and the means of achieving that vision. As we noted in Chapter Four, there is a lack of clarity about the raison d’être of the JAIC, and how the specific mandate and roles it has been assigned—and the visibility and authorities it has been given—support that. DoD needs to be consistent in its intent, actions, and messaging. Our first strategic recommendation therefore addresses DoD’s vision for AI and the governance structures that would support this vision, as articulated in the DoD AI strategy. Recommendation S-1: DoD should adapt AI governance structures that align authorities and resources with the mission of scaling AI. 66 The Department of Defense Posture for Artificial Intelligence As we noted in Chapter Four, DoD’s vision for AI and the scale, urgency, and unity of efforts conveyed in that vision are at odds with the visibility, authorities, and resources it has provided to the JAIC— the focal point of DoD AI. DoD needs to develop governance and organizational structures that align authorities and resources with its vision of scaling AI across DoD. The insights gathered from our industry interviews and the supporting change management literature (see section “Industry: Organization” in Appendix C) lead us to believe there is value in, if not strict necessity of, a centralized effort supported at the highest levels with long-term funding commitments to institute organizational change and scale AI across DoD. Indeed, one of our industry interviewees even noted that centralization at onset was key to their organization’s success, and premature decentralization of effort would have likely been detrimental (see section “Industry: Organization” in Appendix C).1 Starting from that premise, we highlight two possible options for organizational and governance structures, and the rationale for them, although noting that other options might be viable as well, subject to further study. The first option would likely require congressional support to execute, while the second can be executed without, as it aligns with current DoD procedures and organizational structures. Option 1: Enhance the visibility and authorities of the JAIC to enable it to carry out its mission of scaling AI and its impact across DoD, including budgetary and workforce authorities over the military services. Option 2: Take a two-pronged organizational approach as follows: • Establish a JAIC council chaired by the JAIC director and consisting of one AI leadership representative from each service.2 1 We note that this comment, and those of other industry interviewees, imply that if the effort to scale AI across an organization is successful, its natural ending point might be the sunset of the centralized entity that drove the transformation as AI capabilities are diffused across the organization. We therefore expect the JAIC’s role to evolve, though we expect that to happen along a longer timeline (ten or more years), based on our assessment of the state of AI technologies in Chapter Three. 2 Regardless of which governance structure is instituted, if any, establishment of a JAIC council, as described, could facilitate coordination between the JAIC and the services. Recommendations 67 • Establish or reinforce a centralized AI coordination and investment organization within each of the services, with appropriate visibility and authorities, to facilitate scaling AI and its impact across the service, and to promote mandated coordination with the JAIC. In either option: • The DSD should provide the JAIC director with opportunities, at least annually, to present and be heard at the Deputy’s Management Action Group (DMAG) forum (or whichever Deputy Secretary–level forum performs the functions of the DMAG). 3 The rationale for Option 1 is as follows: There is evidence to support that DoD has taken the right approach in establishing the JAIC as a centralized focal point for DoD’s AI strategy (see Chapter 4, Organization). The evidence also suggests that the JAIC, pending initial success, will need to continue in that role for several years because of the expected timeline for AI deployment across enterprise, missionsupport, and operational AI (Chapter Three).

### AT UN Solves Now

#### UN actions failed to adopt real action on autonomous weapons – The US and Russia blocked action

Farge 2021 - Senior Correspondant, Reuters Geneva [Emma, 12/17/21, Reuters, <https://www.reuters.com/world/un-talks-adjourn-without-deal-regulate-killer-robots-2021-12-17/#:~:text=The%20Convention%20on%20Certain%20Conventional,artificial%20intelligence%20and%20facial%20recognition>, , “U.N. talks adjourn without deal to regulate ‘killer robots’”, 6/18/22, LND]

GENEVA, Dec 17 (Reuters) - Countries taking part in U.N. talks on autonomous weapons stopped short of launching negotiations on an international treaty to govern their use, instead agreeing merely to continue discussions. The International Committee of the Red Cross and several NGOs had been pushing for negotiators to begin work on an international treaty that would establish legally-binding new rules on the machine-operated weapons. Unlike existing semi-autonomous weapons such as drones, fully-autonomous weapons have no human-operated "kill switch" and instead leave decisions over life and death to sensors, software and machine processes. Opponents say they raise the risks for civilians, pose problems for accountability and increase the chances of conflict escalation. The Geneva talks, ongoing for eight years, have taken on new urgency since a U.N. [panel report in March](https://undocs.org/S/2021/229) that said the first autonomous drone attack may have already occurred in Libya. "It's a real missed opportunity and not in our view what is needed to respond to the risks posed by autonomous weapons," Neil Davison, a policy adviser in the Legal Division at ICRC, said of the outcome of the week-long talks. Many countries also expressed disappointment in the outcome. "At the present rate of progress, the pace of technological development risks overtaking our deliberations," Switzerland's Disarmament Ambassador, Felix Baumann, said. The Convention on Certain Conventional Weapons which has 125 parties has been discussing possible limits on the use of lethal autonomous weapons, or LAWS, which are fully machine-operated and use new technology such as artificial intelligence and facial recognition. U.N. Secretary General Antonio Guterres had called for countries to come up with an "ambitious plan" on new rules. [read more](https://www.reuters.com/world/un-chief-urges-action-killer-robots-geneva-talks-open-2021-12-13/) Sources following the talks said that Russia, India and the United States were among the countries who expressed doubts about the need for a new LAWS treaty. Washington has previously pointed to their possible benefits, such as precision. Clare Conboy of campaign group Stop Killer Robots said the outcome was one that "keeps the minority of militarised states investing in developing these weapons very happy". She said she expected the many countries in favour of a new law such as New Zealand or Austria to begin negotiations outside of the United Nations.

## “Arms Race” Framing

### 2AC Kritik of Disads with Arms Race Impacts

#### The Negative reinforces the Security Dilemma described in the 1AC – their discourse labelling AI as an “arms race” increases the risk of conflict – hyperbolic rhetoric escalates rivalries and makes AI brittle.

Roff 2019 – Fellow in the Brookings Foreign Policy Program [Heather, 4/26/19, <https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/00963402.2019.1604836>, “The Frame problem: The Ai ‘Arms Race’ isn’t one’, 6/18/22, LND]

Often,[12](https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/00963402.2019.1604836) one hears the phrase “AI arms race,”[13](https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/00963402.2019.1604836) especially in regard to competition[14](https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/00963402.2019.1604836) between major powers. Yet an AI arms race seems a particularly unfortunate and misleading phrase. Since 1957,[15](https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/00963402.2019.1604836) myriad scholars have attempted to understand arms races, including how one can identify and measure[16](https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/00963402.2019.1604836) them and what the potential effects of one are. While there have certainly been academic and policy disputes over the last 62 years, there is one important consensus to note about arms races: looking at one particular dimension of a state’s behavior is not sufficient. Arms races deal with military build ups, arms expenditures, rivalry, alliances, territorial disputes, economic policies, and more. As yet, however, there has been no coherent or comprehensive discussion about the so-called AI arms race. For example: How is AI by itself a weapon? Or is the “race” merely military modernization efforts that include automation, autonomy, or AI enabled military systems? How would we even begin to find, label and disaggregate the numbers to claim that there is an arms race between rivals regarding only AI? Indeed, the discussion of an AI arms race is reminiscent of the hyperbolic and mislabeled rhetoric surrounding “cyber bombs”[17](https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/00963402.2019.1604836) in the battle again ISIS. Of course, there are quite spectacular claims about AI’s potential benefits and risks. Russian President Vladimir Putin[18](https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/00963402.2019.1604836) fanned that fire when he claimed that “whoever becomes the leader in this space [AI] will become the ruler of the world.” But talking about technological competition – in research, adoption, and deployment – in all sectors of multiple economies and in warfare is not really an arms race. Indeed, to frame this competition in military terms risks the adoption of policies or regulations that could escalate rivalry between states and increase the likelihood of actual conflict. More accurately stated, the current situation is one of AI competition, with variations of technological proliferation and diffusion. In some cases, countries may want to limit the amount and kinds of specific AI systems that proliferate to other countries or non-state actors. In these instances, it will more than likely be particular kinds of components or platforms that are at issue. The International Traffic in Arms Regulations (ITAR), the Export Administration Regulations (EAR), the Wassenaar Arrangement on Export Controls for Conventional Arms and Dual-Use Goods and Technologies, the Australia Group, or the Missile Technology Control Regime are but a few regimes that are meant to deal with such proliferation issues.

#### Reject their framing of an “Arms Race” – changing our discourse in conjunction with plan are necessary to head off proliferation, escalation and constructed insecurity

Roff 2019 – Fellow in the Brookings Foreign Policy Program [Heather, 4/26/19, <https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/00963402.2019.1604836>, “The Frame problem: The Ai ‘Arms Race’ isn’t one’, 6/18/22, LND]

Diffusion, however, is a different animal. Widely available commercial off the shelf components, a widely available and open source knowledge base, and wide access to large amounts of data make limiting the diffusion of AI knowledge almost impossible. There are necessary ingredients for AI, including access to computing power and sufficient amounts of data. Some actors possess more of these ingredients than others, but that advantage does not preclude individuals, groups, companies, or states from obtaining access to and knowledge about AI. As greater emphasis is placed on the economic and security benefits of using AI systems, AI will become more diffused because the incentive structure rewards diffusion. Managing the risks of proliferation and diffusion of a knowledge base is an entirely different enterprise than restraining an arms race. In the case of artificial intelligence, rather than looking at what sorts of actions are required for deterrence, what balances may affect conflict onset or escalation, world leaders should turn their focus to how to foster responsible competition. Mitigating the risks associated with AI is not a single-shot activity. Certainly there are technological solutions, such as researching new ways of testing, verifying, and validating systems that include the technologies that fall under the AI umbrella. There even may be novel ways of constraining unwanted system behaviors by generating new architectures or safety controls. Ultimately, however, reducing the risks that AI will be abused requires a reframing of the way in which we think, talk, write about, and deploy AI. The problems of AI misuse are human problems; they are problems exhibited by all dual-use technologies, not just AI. Control of artificial intelligence lies with humans, because they are the moral agents responsible for the design, development, and deployment of AI.

### “Arms Race” Rhetoric Links

#### The Rhetoric of an AI Arms Race sparks a race to the bottom – states will deploy unsafe weapons to get them out before their competitors

Scharre, 2019 - Vice President and Director of Studies at CNAS[Paul, May-June, “Killer Apps: The Real Dangers of an AI Arms Race,” https://omnilogos.com/killer-apps-real-dangers-of-ai-arms-race/6/18/22 MD]

The nation that leads in the development of artificial intelligence will, Russian President Vladimir Putin proclaimed in 2017, "become the ruler of the world." That view has become commonplace in global capitals. Already, more than a dozen governments have announced national AI initiatives. In 2017, [China](https://go-gale-com.proxy.lib.umich.edu/ps/i.do?p=AONE&u=umuser&id=GALE|A585763278&v=2.1&it=r) set a goal of becoming the global leader in AI by 2030. Earlier this year, the White House released the American AI Initiative, and the U.S. Department of Defense rolled out an AI strategy. But the emerging narrative of an "AI [arms race](https://go-gale-com.proxy.lib.umich.edu/ps/i.do?p=AONE&u=umuser&id=GALE|A585763278&v=2.1&it=r)" reflects a mistaken view of the risks from AI--and introduces significant new risks as a result. For each country, the real danger is not that it will fall behind its competitors in AI but that the perception of a race will prompt everyone to rush to deploy unsafe AI systems. In their desire to win, countries risk endangering themselves just as much as their opponents. AI promises to bring both enormous benefits, in everything from health care to transportation, and huge risks. But those risks aren't something out of science fiction; there's no need to fear a robot uprising. The real threat will come from humans. Right now, AI systems are powerful but unreliable. Many of them are vulnerable to sophisticated attacks or fail when used outside the environment in which they were trained. Governments want their systems to work properly, but competition brings pressure to cut corners. Even if other countries aren't on the brink of major AI breakthroughs, the perception that they're rushing ahead could push others to do the same. And if a government deployed an untested AI weapons system or relied on a faulty AI system to launch cyberattacks, the result could be disaster for everyone involved.Policymakers should learn from the history of computer networks and make security a leading factor in AI design from the beginning. They should also ratchet down the rhetoric about an AI arms race and look for opportunities to cooperate with other countries to reduce the risks from AI.A race to the bottom on AI safety is a race no one would win.

#### “Arms Race” rhetoric creates a self-fulfilling prophecy by causing states to cut corners on AI development and testing

Scharre, 2019 - Vice President and Director of Studies at CNAS[Paul, May-June, “Killer Apps: The Real Dangers of an AI Arms Race,” https://omnilogos.com/killer-apps-real-dangers-of-ai-arms-race/6/18/22 MD]

When it comes to applying AI to national security, government agencies will have to reconsider their traditional approaches to testing new systems. Verifying that a system meets its design specifications isn't enough. Testers also need to ensure that it will continue to function properly in the real world when an adversary is trying to defeat it. In some cases, they can use computer simulations to tease out bugs, as manufacturers now do for autonomous cars. On top of that, the Departments of Defense and Homeland Security and the intelligence community should create red teams--groups that act as attackers to test a system's defenses--to ferret out vulnerabilities in AI systems so that developers can fix them before the systems go live. Government officials should also tone down their rhetoric about an AI arms race, since such talk could easily become self-fulfilling. At a conference in 2018, Michael Griffin, the chief Pentagon official for research and engineering, said, "There might be an artificial intelligence arms race, but we're not yet in it." Militaries are certainly going to adopt AI, but Griffin's statement was missing any concern for--or even awareness of--the risks that come with it. Talk of an arms race encourages adversaries to cut corners on safety. Government officials should emphasize not only the value of AI but also the importance of guaranteeing reliability and security.

### Security Dilemma Links

#### US AI deployment is driven by the fear that China will take the lead.

Scharre, 2019 - Vice President and Director of Studies at CNAS[Paul, May-June, “Killer Apps: The Real Dangers of an AI Arms Race,” https://omnilogos.com/killer-apps-real-dangers-of-ai-arms-race/6/18/22 MD]

WHAT AI WILL DO Whichever country takes the lead on AI will use it to gain economic and military advantages over its competitors. By 2030, AI is projected to add between $13 trillion and $15 trillion to the global economy. AI could also accelerate the rate of scientific discovery. In 2019, an artificial neural network significantly outperformed existing approaches in synthetic protein folding, a key task in biological research. AI is also set to revolutionize warfare. It will likely prove most useful in improving soldiers' situational awareness on the battlefield and commanders' ability to make decisions and communicate orders. AI systems can process more information than humans, and they can do it more quickly, making them valuable tools for assessing chaotic battles in real time. On the battlefield itself, machines can move faster and with greater precision and coordination than people. In the recent AI-versus-human StarCraft match, the AI system, AlphaStar, displayed superhuman abilities in rapidly processing large amounts of information, coordinating its units, and moving them quickly and precisely. In the real world, these advantages will allow AI systems to manage swarms of robots far more effectively than humans could by controlling them manually. Humans will retain their advantages in higher-level strategy, but AI will dominate on the ground. Washington's rush to develop AI is driven by a fear of falling behind China, which is already a global powerhouse in AI. The Chinese technology giants Alibaba, Baidu, and Tencent rank right alongside Amazon, Google, and Microsoft as leading AI companies. Five of the ten AI startups with the most funding last year were Chinese. Ten years ago, China's goal of becoming the global leader in AI by 2030 would have seemed fanciful; today, it's a real possibility.

#### US deployment of Autonomous Weapons is driven by the perception of “falling behind” in an AI arms race

Sosanya, 2022 - AI researcher and a policy analyst at the Day One Project [Andrew, Jan 3, Peace Review A Journal of Social Justice “Autonomous Weapons Are Here to Stay” <https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/10402659.2021.1998856> TM]

Great powers cannot afford to be idle in the face of technological innovation. Today, clear signs that leading militaries are preparing for the inevitable warfare of tomorrow exist: a future with autonomous weapons. With promising technology that is still in the research and development (R&D) phase, states act cautiously as the future has become doubly uncertain: they cannot predict their adversaries’ intentions, nor can they predict the usefulness of the weapon. In the United States, Department of Defense (DoD) leaders have indicated their willingness to deploy them. In 2017, Lieutenant General Jack Shanahan, the then-Director of DoD’s Joint AI Center, stated that he “does not want to see a future where our potential adversaries have a fully AI-enabled force and we do not.” The DoD has upstarted numerous programs in the last decade dedicated to autonomy, covering robotics, swarms, target recognition, machine learning, and more. Shanahan had cited China and Russia’s military AI R&D as a reason for the DoD’s increased focus on improving their own militarized artificial intelligence. Former Secretary of Defense Bob Work, who oversaw the 2017 revision of the DoD’s autonomous weapons policy, corroborated Shanahan’s thinking, that the United States would deploy autonomous weapons if their adversaries do first.

### “Arms Race” causes Premature Deployment

**The Arms Race context pressures states to adopt AI before safe, increasing the risk of unintentional conflict and escalation of crises**

**Horowitz and Kahn, 2021 – Senior and Research Fellows at the Council on Foreign Relations** [Michael and Lauren, The Washington Quarterly 3-19-2021 “Leading in Artificial Intelligence through Confidence Building Measures”  [https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/0163660X.2021.2018794 acc on 6-21-2022](%20https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/0163660X.2021.2018794%20acc%20on%206-21-2022%20) TM]

Risks exist for AI uses across society, in civilian and military sectors alike.7 Survey research shows that many in AI and machine learning (ML) communities worry about whether militaries can be trusted to develop safe and reliable AI.8 Given the current geopolitical context, there is fear that competitive pressures to be first in deploying AI systems could overshadow safety and ethics efforts, incentivizing militaries to take “short cuts” that could increase the risk of lethal accidents, unintentional conflict (conflict between states that results from miscommunication or accidents), and inadvertent escalation (when states commit intentional actions that unintentionally cause escalation by an adversary). The Risk of Accidents First, military applications of AI may not work as intended due to flaws in the algorithms themselves. Inadequate training data, the algorithm’s complexity, biased data, biased coding, or intentional data sabotage or poisoning by adversaries can make accidents more likely when systems are deployed. For example, a malfunctioning AI-enabled targeting system in a conflict could attack friends instead of foes, or friends and foes alike. Or an algorithmic decision aid could derive a flawed estimate of the risk of a particular operational plan, leading to a use of force that either unnecessarily fails or succeeds at a greater cost than required. These risks are amplified when systems are employed outside their design context, which may be more likely due to competitive dynamics.9 The Risk of Unintentional Conflict Second, even if algorithms work as designed, they could unintentionally make conflict more likely. Uncertainty about how algorithms will work on the battlefield could create challenges for signaling adversaries through posture and deployments in a crisis or within a conflict. Actors might be uncertain about how A Ienabled systems deployed by an adversary will behave. They might not believe that AI-enabled autonomous systems are programmed as described because the black-box nature of algorithms could lead to increased skepticism of claims made about the systems. Even if a state is telling the “truth” about a specific system and how it was used, there might not be a way for a third party or another state to validate those claims. Mistrust due to great power competition could exacerbate the situation. There is still substantial uncertainty about the reliability of many AI methods, especially deep learning and related techniques when they generate opaque outputs, meaning there is not an available chain of logic which explains why the algorithm recommended or engaged in a particular action.10 Even with sufficient training data, hedges against biases, and protections from data poisoning and hacking, current DoD systems may not be prepared for testing, evaluation, validation, and verification (TEVV) of algorithmic systems, particularly AI-enabled autonomous systems that will continually learn while operating.11 Furthermore, in operating at machine speed, AI-enabled autonomous systems could cause adversaries in a conflict to fear losing a war so quickly that they feel compelled to escalate.12 For an adversary with nuclear weapons, this could create pressure for launch postures such as pre-delegation and launch on warning that increase the chance of nuclear use.

#### The global misperception of an “Arms Race” forces states to rush deployment which inevitably leads to accidents. There is no arms race currently

Scharre 2021 – Director of the Technology and National Security Program at the Center for a New American Security [Paul, 6/28/21,https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/, “Debunking the AI Arms Race Theory”, 6/18/22, LND]

In 2015, a group of prominent AI and robotics researchers signed an open letter warning of the dangers of autonomous weapons. “The key question for humanity today,” they wrote, “is whether to start a global AI arms race or to prevent it from starting. If any major military power pushes ahead with AI weapon development, a global arms race is virtually inevitable.”[1](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/#_ftn1) Today, many nations are working to apply AI for military advantage, and the term “AI arms race” has become a catchphrase used by both critics and proponents of AI militarization. In 2018, then-Under Secretary of Defense Michael Griffin, calling for the United States to invest more in AI, stated, “There might be an artificial intelligence arms race, but we’re not yet in it.”[2](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/#_ftn2) In a 2020 Wired article, Will Roper, then chief acquisition officer for the U.S. Air Force, warned of the risks of falling behind in a “digital arms race with China.”[3](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/#_ftn3) The so-called AI arms race has become a common feature in news headlines,[4](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/#_ftn4) but the arms race framing fails to match reality. While nations are clearly competing to develop and adopt AI technology for military use, the character of that competition does not meet the traditional definition of an arms race. Military AI competition nevertheless does pose risks. The widespread adoption of military AI could cause warfare to evolve in a manner that leads to less human control and to warfare becoming faster, more violent, and more challenging in terms of being able to manage escalation and bring a war to an end. Additionally, perceptions of a “race” to field AI systems before competitors do could cause nations to cut corners on testing, leading to the deployment of unsafe AI systems that are at risk of accidents that could cause unintended escalation or destruction. Even if fears of an “AI arms race” are overblown, military AI competition brings real risks to which nations should attend. There are concrete steps nations can take to mitigate some of these dangers.

#### “Arms Race” mentality pressures rapid deployment of weapons – militaries skip critical testing and evaluation.

Horowitz and Scharre 2021 - Director of the Emerging Capabilities Policy Office in the Office of the Under Secretary of Defense for Policy and  Vice President and Director of Studies at CNAS [Michael, Paul, January 12 2021, “AI and International Stability: Risks and Confidence- Building Measures”, <https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures>, Acc 6/18/22. M.A.]

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](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn23) 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](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn24) 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](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn25) 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](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn26) 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.

#### Focusing on “winning” the arms race will undermine effective oversight and testing – we will rush into AI development without safeguards against their abuse.

Bob, 2019 - intelligence, terrorism and legal analyst at the Jerusalem Post[Yonah Jeremy, 13 February Jerusalem Post, “Top cybersecurity expert: Crisis worse than 9/11 could emerge from AI arms race.” <https://www.proquest.com/docview/2182568907?parentSessionId=BxSQxuLDub9YJx%2BOeIXcOzvsMV848Jfpxd5%2B%2B3GON8c%3D&pq-origsite=primo&accountid=14667>, 6/18/22 MD]

Winning the artificial intelligence arms (AI) race between the US, Israel and their adversaries may need to take precedent over balancing the risks at this stage, a former government agency chief technology officer has told The Jerusalem Post. Those risks could include crises even worse than 9/11, said Amit Meltzer, now a top cybersecurity consultant. Discussing the issue on Tuesday, merely a day after US President Donald Trump issued the first-ever executive order to bolster US efforts in the artificial intelligence arms race, he said that winning such a competition would not go well with careful oversight of negative consequences and potential abuses. While US Senate Vice Intelligence Committee Chairman Mark Warner (D-VA) applauded aspects of the order, he criticized the lack of developing oversight for the misuse of AI. Warner supported the order’s provisions for opening US federal data sets to non-federal entities to enhance cooperation, but said that Trump’s order 'reflects a laissez-faire approach to AI development that... will have the US repeating the mistakes it has made in treating digital technologies as inherently positive forces, with insufficient consideration paid to their misapplication.' Meltzer called Trump’s declaration more symbolic rather than concretely operational, but said that it still was crucial 'to give strong backing to academic and government institutions… which suffer from a chronic lack of personnel' and resources. Furthermore, he said that concerns that certain companies would quickly gain domination of the AI sector and abuse their standing economically were possible. But, he said that it was nearly impossible to square 'the national necessity for the US' or Israel to 'strengthen and maintain leadership in the industry' with policies that encourage caution and that new technologies should only be rolled out after any danger was carefully examined. He compared the need to act with trying to keep up in the field of artificial intelligence with the US’s sale of weapons to any dictatorship that is not in a direct conflict with it, in order to make weapon development economically sustainable. However, Meltzer said that neither Trump nor Warner honed-in on the true potential dangers of AI.Once AI starts taking over a large number of societal functions, damage to a such a network could have far more disastrous and broader consequences and at a more rapid pace, he explained. By raising the specter of a disaster the size of the stock market crash of 1987, he said that economic or social-psychological warfare influence campaigns that dwarf current threats would become possible. Another nightmare scenario could be using an AI algorithm to make four million Toyota cars all crash at the same time worldwide, leaving countless dead and wounded in a tragedy 'that would be worse than 9/11.' The worst part of these dangers is that artificial intelligence could make such attacks infinitely easier than with advanced existing cyber hacking capabilities. Despite these threats, he predicted that the US needs to keep up with China, Russia and Israel’s need to stay ahead of its adversaries, meaning that rapidly developing new capabilities would continue to be put before protecting civil liberties. 'Afterwards, they will fix things and ask forgiveness from the victims,' he said of moving ahead too fast without proper oversight.

The Rhetoric of an AI Arms Race sparks a race to the bottom – states will deploy unsafe weapons to get them out before their competitors

Scharre, 2019 - Vice President and Director of Studies at CNAS[Paul, May-June, “Killer Apps: The Real Dangers of an AI Arms Race,” https://omnilogos.com/killer-apps-real-dangers-of-ai-arms-race/6/18/22 MD]

The nation that leads in the development of artificial intelligence will, Russian President Vladimir Putin proclaimed in 2017, "become the ruler of the world." That view has become commonplace in global capitals. Already, more than a dozen governments have announced national AI initiatives. In 2017, [China](https://go-gale-com.proxy.lib.umich.edu/ps/i.do?p=AONE&u=umuser&id=GALE|A585763278&v=2.1&it=r) set a goal of becoming the global leader in AI by 2030. Earlier this year, the White House released the American AI Initiative, and the U.S. Department of Defense rolled out an AI strategy. But the emerging narrative of an "AI [arms race](https://go-gale-com.proxy.lib.umich.edu/ps/i.do?p=AONE&u=umuser&id=GALE|A585763278&v=2.1&it=r)" reflects a mistaken view of the risks from AI--and introduces significant new risks as a result. For each country, the real danger is not that it will fall behind its competitors in AI but that the perception of a race will prompt everyone to rush to deploy unsafe AI systems. In their desire to win, countries risk endangering themselves just as much as their opponents. AI promises to bring both enormous benefits, in everything from health care to transportation, and huge risks. But those risks aren't something out of science fiction; there's no need to fear a robot uprising. The real threat will come from humans. Right now, AI systems are powerful but unreliable. Many of them are vulnerable to sophisticated attacks or fail when used outside the environment in which they were trained. Governments want their systems to work properly, but competition brings pressure to cut corners. Even if other countries aren't on the brink of major AI breakthroughs, the perception that they're rushing ahead could push others to do the same. And if a government deployed an untested AI weapons system or relied on a faulty AI system to launch cyberattacks, the result could be disaster for everyone involved. Policymakers should learn from the history of computer networks and make security a leading factor in AI design from the beginning. They should also ratchet down the rhetoric about an AI arms race and look for opportunities to cooperate with other countries to reduce the risks from AI. A race to the bottom on AI safety is a race no one would win.

**Competitive races pressure militaries to deploy AI before it is safe, increasing the risk of catastrophic accidents**

**Arnold and Toner, 2021 – Center for Security and Emerging Threats** [Zachary and Helen, July, CSET Policy Brief. “AI Accidents: An Emerging Threat What Could Happen and What to Do” https://cset. georgetown.edu/wp-content/uploads/CSET-AI-Accidents-An-Emerging-Threat.pdf Acc 6/7/22 TA]

As AI is integrated into more and more critical systems, the dangers of AI accidents will grow. In practice, we expect these accidents will be more likely and more severe in some situations than in others. Identifying these risky situations ahead of time is challenging, but based on AI accidents that have already occurred and historical accidents involving other technologies, we expect risk factors for severe AI accidents will include: ● Competitive pressure. When not using AI could mean falling behind competitors or losing profits, companies, militaries, and governments are more likely to deploy buggy AI systems, use them in reckless ways, or cut corners on testing and operator training.39 The infamous Boeing 737 MAX, though it does not use machine learning, is an example of this dynamic. The aircraft was developed, tested, and certified under extreme time pressure, aiming to compete with a comparable Airbus system.40 Ultimately, this haste led to two crashed planes and hundreds of deaths.

#### An “Arms Race” mentality creates pressure to deploy brittle, untested AI – this causes weapons failures in real world environments

Sharre, 2018 - director of the technology and national security program at the Center for a New American Security [Paul, 12 September, “Ultrafast computing is critical to modern warfare. But it also ensures a lot could go very wrong, very quickly.” <https://foreignpolicy.com/2018/09/12/a-million-mistakes-a-second-future-of-war/> ST]

The problem is that automated systems—at least those using current technology—tend to be brittle. Machines are good at handling routine tasks under predictable circumstances, such as flying a commercial airliner. But automation can sometimes fail dramatically in new situations, which is a major reason why self-driving cars, which must contend with extremely dynamic and uncontrolled environments, have proved so much harder to develop than self-flying planes. Learning systems, which focus on a set of rules for processing and incorporating data into behavior instead of strict mandates, exhibit more flexibility but are still limited by the quality of information. Machine learning can fail if the real world proves different from training models—a major problem in the military context, where adversaries are unlikely to offer easy access to their tactics and hardware. Human intelligence is robust and adaptable in ways that machine intelligence is not—yet. The most effective militaries will thus be those that find ways to successfully marry human and machine intelligence into joint cognitive systems—an approach that defense analysts call centaur warfighting. Despite humans’ advantages in decision-making, an arms race in speed may slowly push humans out of the OODA loop. Militaries are unlikely to knowingly field weapons they cannot control, but war is a hazardous environment and requires balancing competing risks. Faced with the choice of falling behind an adversary or deploying a new and not yet fully tested weapon, militaries are likely to do what they must to keep pace with their enemies. As mentioned above, automated stock trading provides a useful window into the perils of this dynamic. In 2010, the Dow Jones Industrial Average lost nearly 10 percent of its value in just minutes. The cause? A sudden shift in market prices driven in part by automated trading, or what’s come to be known as a flash crash. In the last decade, financial markets have started to suffer such crashes, or at least miniature versions of them, on a regular basis. The circuit breakers installed by regulators to pull a stock offline can’t prevent incidents from occurring, but they can stop flash crashes from spiraling out of control. Circuit breakers are still regularly tripped, though, and on Aug. 24, 2015, more than 1,200 of them went off across multiple exchanges after China suddenly devalued the yuan. In competitive environments such as stock markets and battlefields, unexpected interactions between algorithms are natural. The causes of the 2010 flash crash are still disputed. In all likelihood, there were a range of causes, including an automated sell algorithm interacting with extreme market volatility, exacerbated by high-frequency trading and deliberate spoofing of trading algorithms. To prevent the military equivalent of such crises, in which autonomous weapons become trapped in a cascade of escalating engagements, countries will have to balance advantages in speed with the risk of accidents. Yet growing competition will make that balancing act ever more difficult. In 2016, Robert Work, then-U.S. deputy defense secretary, colorfully summed up the problem this way: “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?”

### “Arms Race” Impacts

#### The Arms Race mentality limits our thinking about AI – it prevents global collaboration, which is critical to get the most out of the technology

Huang and Scott, 2018 - Chief Executive Officer and Chief Technology Officer at Malong Technologies [Dinglong and Matt July 21 World Economic Forum “Who will win the AI race? If countries work together, then the answer could be all of us” https://www.weforum.org/agenda/2018/06/ai-arms-race-global-collaboration/, BK]

Why, then, the fixation? We suggest three reasons, starting with the most obvious: China’s 2017 declaration of support for AI as a national priority got the world’s attention, and other nations have followed suit. Second, the great places to work for the world’s best talent now include cities in China. We’re in Shenzhen, Beijing and Shanghai, and can observe our peer AI companies also thriving in these cities. Third, the amount of seed capital and venture capital investment available in China to support great companies is robust and according to some studies, exceeds that of the US for the first time. This need not be something to fear. The most promising near-term AI applications are those that are supported by globally relevant data, and will help people wherever they happen to live. Take medical technology. Everyone looks the same on the inside, and everyone can benefit from AI-enabled solutions that lead to faster, more accurate diagnoses and more effective treatment. Or take manufacturing and agriculture. AI-driven improvements and efficiencies will bring benefits where the factories and farms are. We aren’t arguing for nations to abandon homegrown efforts to support AI research, investment and entrepreneurship. In fact, we see them as something to celebrate and encourage. We expect great people, ideas and companies to continue to show up all over the world, with all the attendant benefits in terms of job creation and economic growth. There are indeed amazing things going on not only in the US and China, but also in Canada and Europe, Korea and Japan. It’s genuinely an international phenomenon. Problems come when we limit our thinking to nation-vs-nation competition. It bogs down our progress. There cannot be a Chinese AI vs an American AI vs a French AI, and so on. Instead, those national efforts must contribute to global cooperation and collaboration if AI is to advance and bring benefits to all communities worldwide. Let the companies do the competing.

#### Engaging in an AI arms race undermines US security and stability, and drains resources from more important causes.

Garcia, 2021 - Vice-chair of the International Committee for Robot Arms Control[Denise, May 13, 2021, Nature.com, “Stop the emerging AI cold war,” <https://www.nature.com/articles/d41586-021-01244-z#:~:text=Proliferating%20military%20artificial%20intelligence%20will,on%20ethics%20and%20global%20cooperation.&text=Denise%20Garcia%20is%20a%20professor,Committee%20for%20Robot%20Arms%20Control>., 6/18/22 MD]

A race to militarize artificial intelligence is gearing up. Two years ago, the US Congress created the [National Security](https://go-gale-com.proxy.lib.umich.edu/ps/i.do?p=ITOF&u=umuser&id=GALE|A661476451&v=2.1&it=r) Commission on Artificial Intelligence (NSCAI). This March, it recommended that the [United States](https://go-gale-com.proxy.lib.umich.edu/ps/i.do?p=ITOF&u=umuser&id=GALE|A661476451&v=2.1&it=r) must accelerate artificial-intelligence (AI) technologies to preserve national security and remain competitive with China and Russia. This will undermine the United States' ability to lead emerging global norms on AI. In April, the European Commission published the first international legal framework for making AI secure and ethical; in January, the European Parliament issued guidelines stating that military AI should not replace human decisions and oversight. By contrast, the NSCAI recommendations advocate "the integration of AI-enabled technologies into every facet of war-fighting". Enhancing AI war-fighting capacity will decrease security in a world where the biggest threats are instability -- political, social, economic and planetary. The NSCAI should heed the research community. Some 4,500 AI and [robotics](https://go-gale-com.proxy.lib.umich.edu/ps/i.do?p=ITOF&u=umuser&id=GALE|A661476451&v=2.1&it=r) researchers have declared that AI should not make the decision to take a human life -- aligning with the European Parliament guidelines and the European Union regulation. The NSCAI resurrected disastrous ideas from the cold war and framed its report in terms of winning a competition for AI-enabled warfare. During the cold war, the drive to stay ahead in the technological race led to the accumulation of 70,000 nuclear weapons and today's global arsenal of 13,100 warheads. This brought extortionate costs: US$70 billion is spent annually to maintain nuclear weapons globally. Other threats demand similar investments: in 2019, climate-induced natural disasters displaced 25 million people, and decentralized conflicts forced 8.6 million to move. Still more threats affect infrastructure, such as the ransomware attack on 8 May that shut down a 8,850-kilometre US fuel pipeline. The NSCAI does not prioritize international cooperation to create new regulations. Indeed, it speaks against a global ban on autonomous weapons, saying that other countries cannot be trusted to comply. But an AI-militarization race would be profoundly destabilizing. Unlike nuclear arms, AI is already ubiquitous in civilian spheres, so the dual-use risks of, say, flying drones or computer night [vision](https://go-gale-com.proxy.lib.umich.edu/ps/i.do?p=ITOF&u=umuser&id=GALE|A661476451&v=2.1&it=r) are much higher. Since 2014, I have been an observer and adviser at United Nations meetings, and I testified in 2017 as part of the International Panel on the Regulation of Autonomous Weapons. In my view, rather than focusing on counting weapons or on particular weapons systems, policies should specify human intention and human-machine interaction, obligating countries to maintain human control over military force. Other agreements could mitigate malicious uses of AI, such as using facial recognition to oppress citizens or biased data to guide decisions about employment or incarceration. The world's people need protection from cyberattacks to infrastructure -- such as those on US hospitals in 2020 or those that hit national electrical grids. The NSCAI report calls for international standards for AI-enabled and autonomous weapons systems, arguing that if these systems are properly tested and designed, humans can use them to make the decision to kill, consistent with international humanitarian law. This is misleading: it's difficult to make machine learning's 'black box' nature fully interpretable, or to ensure that AI systems perform as expected after deployment. These systems learn from their environment, and the real world is never as simple as the laboratory. The NSCAI argues that the United States should seek commitments from Russia and China against autonomous nuclear weapons, even as it argues against treaties regulating other autonomous and AI weapons. Instead, the United States should negotiate decreases in nuclear arsenals and establish standards to keep humans in meaningful control. The NSCAI is too dismissive by discounting cooperation. The Chemical Weapons Convention, the Biological Weapons Convention, the UN Sustainable Development Goals and the 1987 Montreal Protocol are examples of accountability on which all the major powers worked together. The United States and Russia established the International Space Station by cooperating closely. Most nations want governance that controls the use of AI in war. In June 2020, the Global Partnership on Artificial Intelligence was created by the Group of Seven industrialized countries (G7) and called for human-centric development and use of AI. The partnership brings scientific and research communities together with industry and government to facilitate international cooperation. This is the path that the United States should take -- with scientists, researchers and industry alike. The relentless pursuit of militarization does not protect us. It diverts resources and attention from nearer existential threats, such as extreme weather events. With the world reeling from COVID-19 -- the shock of the century -- now is not the moment to hasten towards worldwide confrontation. In 2019 alone, climate disasters displaced almost one million people in the United States. China, too, is extremely vulnerable to global warming. This common ground could pave the way to cooperation, including stopping the emerging AI cold war. This is no time to embark on an exorbitant and ineffective race.

### Reject “Arms Race” Rhetoric

#### Rejecting the Arms Race framing is essential to prevent the Abuse of the technology

Roff 2019 – Fellow in the Brookings Foreign Policy Program [Heather, 4/26/19, <https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/00963402.2019.1604836>, “The Frame problem: The Ai ‘Arms Race’ isn’t one’, 6/18/22, LND]

There needs to be a change in thinking about AI. Those dealing with AI must insist on greater clarity about its definition. If policy makers and other leaders are not clear about what the term means and entails, they cannot possibly formulate best practices and governance mechanisms. It would help matters if artificial intelligence discussions were framed in an “AI +” framework, because in many cases, AI is merely a tool included in a system involving other functions or capabilities. The news media should stop framing the global artificial intelligence competition as an “arms race.” This misrepresents the competition going on among countries. The policy community needs a clear-eyed appraisal of AI’s capabilities and limitations. Without that orientation, those who hope to steer research and development in positive directions will create more problems than they solve. Last year, colleagues and I in the “AI Safety” community published a paper[1](https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/00963402.2019.1604836) on the potential misuses of artificial intelligence and potential interventions to lessen the likelihood and impact of misuse. Because of the dual-use nature of artificial intelligence, its almost ubiquitous nature, and its potential to create threats – either new or in combination with previously existing threats – we recommended that policy makers and others involved in AI efforts begin to respond immediately. In many cases, my coauthors have done just that. For example, OpenAI – an organization that hopes to ensure artificial general intelligence benefits humanity and that includes some of my co-authors – recently refused[2](https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/00963402.2019.1604836) to publish the full model of its GPT-2 algorithm, which can generate synthetic natural language text – that is, articles, answers to reading comprehension questions, and other types of writing – of “unprecedented quality.” OpenAI realized that GPT-2 could help malicious actors generate disinformation and abusive content, increasing the likelihood of fraud based on impersonation. So they released only a smaller version of the GPT-2 code, dataset, and related information. That decision has been met with both derision[3](https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/00963402.2019.1604836) and applause.[4](https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/00963402.2019.1604836) Others[5](https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/00963402.2019.1604836) have also begun serious[6](https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/00963402.2019.1604836) and rigorous[7](https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/00963402.2019.1604836) work at the policy and technology nexus, pushing for responsible and ethical[8](https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/00963402.2019.1604836) design, development, and deployment of AI technologies. But there are other development and policy changes that could help head off potential abuses of AI. First, there needs to be a change in thinking about, and framing of, AI. In particular, those dealing with AI must insist on greater clarity about its definition. If policy makers and other leaders are not clear about what the term means and entails, they cannot possibly formulate best practices and governance mechanisms. Also, it would help matters if artificial intelligence discussions were framed in an “AI +” framework, because in many cases, AI is merely a tool included in a system involving other functions or capabilities – for example, artificial intelligence might be a part of a driverless vehicle that involves many technologies. Finally, I certainly hope that scholars, practitioners, policy makers, and especially the news media stop framing the global artificial intelligence competition as an “arms race.” This framing misrepresents the competition going on among countries. To address the risks of AI, the policy community needs a clear-eyed appraisal of its capabilities and limitations. Without that orientation, those who hope to steer AI research and development in positive directions will create more problems than they solve.

## Brittle AI Advantage

### Brittle – Fast Deployment causes Brittle AI

#### Rapidly developed military AI is brittle – it is unreliable because it is immature and untested in real world situations

Horowitz and Scharre, 2021 - Director of the Emerging Capabilities Policy Office in the Office of the Under Secretary of Defense for Policy and  Vice President and Director of Studies at CNAS [Michael, Paul, January 12 2021, “AI and International Stability: Risks and Confidence- Building Measures”, <https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures>, Acc 6/18/22. M.A.]

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](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn18) 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](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn19) 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](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn20) 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](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn21) 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](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn22) 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.

#### Untested AI increases the risk of accidents – friendly fire incidents empirically prove.

Atherton 2022 – Military Technology Journalist [Kelsey, 5/6/22, “Understanding the errors introduced by military AI applications”, <https://www.brookings.edu/techstream/understanding-the-errors-introduced-by-military-ai-applications/>, 6/18/22, LND]

On March 22, 2003, two days into the U.S.-led invasion of Iraq, American troops fired a Patriot interceptor missile at what they assumed was an Iraqi anti-radiation missile designed to destroy air-defense systems. Acting on that recommendation of their computer-powered weapon, the Americans fired in self-defense, thinking they were shooting down a missile coming to destroy their outpost. What the Patriot missile system had identified as an incoming missile, was in fact a UK Tornado fighter jet, and when the Patriot struck the aircraft, it killed two crew on board instantly. The deaths were the first losses suffered by the Royal Air Force in the war and the tragic result of friendly fire. A subsequent RAF Board of Inquiry [investigation](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/82817/maas03_02_tornado_zg710_22mar03.pdf) concluded that the shoot-down was the result of a combination of factors: how the Patriot missile classified targets, rules for firing the missiles, autonomous operation of Patriot missile batteries, and several other technical and procedural factors, like the Tornado not broadcasting its “friend or foe” identifier at the time of the friendly fire. The destruction of Tornado ZG710, the report concluded, represented a tragic error enabled by the missile’s computer routines. The shoot-down of the Tornado happened nearly 20 years ago, but it offers an insight into how AI-enabled systems or automated tools on the battlefield will affect the kinds of errors that happen in war. Today, human decisionmaking is shifting toward machines. With this shift comes the potential to reduce human error, but also to introduce new and novel types of mistakes. Where humans might have once misidentified a civilian as a combatant, computers are expected to step in and provide more accurate judgment. Across a range of military functions, from the movement of autonomous planes and cars to identifying tanks on a battlefield, computers are expected to provide quick, accurate decisions. But the embrace of AI in military applications also comes with immense risk. New systems introduce the possibility of [new types of error](https://www.popsci.com/technology/autonomous-machines-mistakes-un-institute-report/), and understanding how autonomous machines [will fail](https://www.nytimes.com/2018/11/15/magazine/autonomous-robots-weapons.html) is important when [crafting policy](https://www.brookings.edu/blog/techtank/2019/05/10/its-time-to-start-thinking-about-governance-of-autonomous-weapons/) for [buying and overseeing](https://www.brookings.edu/techstream/applying-arms-control-frameworks-to-autonomous-weapons/) this new generation of autonomous weapons.

#### An Arms Race frame for AI increases the risk of accidental wars – the speed of decision making and the vulnerability to hacking make miscalculation and escalation inevitable

**Vestner, 2021 - Head of Security and Law Programme at Geneva Centre for Security Policy** [Tobias, July 8 “Warfare and Artificial Intelligence” in Robin Geiß and Henning Lahmann (eds), Research Handbook on Warfare and Artificial Intelligence -forthcoming https://www.gcsp.ch/publications/military-operations-and-artificial-intelligenceGCSP Acc 5/27/22 TA]

Most importantly for military strategy, AI applications may assist decision-makers to monitor the battlefield and develop scenarios. Indeed, AI could be developed to predict the behaviour and reactions of foreign countries or generate simulations of the progression of ongoing conflicts.32 AI may also be useful to assess threats, provide risk analyses, and suggest courses of action, ultimately guiding decision-makers on the best response to take.33 In addition, AI may support the alignment of the armed forces’ ways and means with the given political and strategic objectives - a major function of military strategy. A consequence of such developments would be an increased speed and quality of military processes. While this would provide significant advantages to those states with the most performant AI,34 this may also pressure armed forces to increasingly delegate the orchestration of military operations to AI systems.35 Indeed, the use of AI for military strategy may also lead to challenges. Reliable AI systems would need to be trained with vast data sets.36 Furthermore, it has been warned that AI may exacerbate threats, transform their nature and characteristics, and introduce new security threats.37 A tabletop exercise on the integration of AI into nuclear C2 systems showed that such systems were ‘vulnerable to malicious manipulation that can severely degrade strategic stability’,38 for instance. Such vulnerabilities would derive mostly from the risk posed by third actors using techniques to deceive, disrupt or impair C2 systems,39 which indicates the importance of system safety for AI to be used for military strategy. Another significant challenge is that AI may accelerate the speed of warfare to the extent that humans cease to be able to follow the developments, ultimately leading humans to lose control. 40 This phenomenon has been termed battlefield ‘singularity’ or ‘hyperwar’.41 This may lead to strategic errors and accidents, including involuntary conflict escalation. Even if such risks can be alleviated, the increased reliance on AI would reduce the human element of military strategy, in particular psychology and human judgment. It has been argued that this could lead to a ‘gap between how the AI solves a problem framed by humans, and how those humans would solve it if they possessed the AI’s speed, precision, and brainpower’.42 Yet it has also been argued that strategy development would require the understanding of values, the balance of costs, and the understanding of the complex social system in which war operates, thereby significantly limiting AI’s use for military strategy.43 Yet it is also possible that when enemies possess high levels of rational prediction power provided by AI systems, the decisive factor in warfare will not be the AI systems’ capabilities but the human judgment, in particular concerning critical and difficult choices.44 This, however, presumes a certain level of meaningful human involvement. In sum, AI may enhance military strategy development and strategic decision-making, notably if able to process more data and make sense of complexity with more precision and at a higher speed than humans and simple computing. A likely result is an acceleration of military operations, which may increase pressure on armed forces to integrate AI and may marginalize human judgment. States’ recent adoption of defence strategies on and related to AI indicate that states increasingly intend to develop, acquire, and operationalize AI for military purposes. As such, the possession and use of AI is a strategic objective itself. In light of secrecy around the development of new technologies, states’ investment in military AI can become a strategic liability as it may increase the risk of destabilizing arms races, misperceptions, and miscalculations.

### Brittle - AI Accidents Inevitable

#### Failures in AI inevitable – inflexibility in unpredictable situations causes mistakes

Scharre, 2016 -- Vice President and Director of Studies at CNAS [Paul, Feb 2016, Center for New American Security, “Autonomous weapons and operational risk”, https://s3.us-east-1.amazonaws.com/files.cnas.org/documents/CNAS\_Autonomous-weapons-operational-risk.pdf?mtime=20160906080515&focal=none, Acc 6/21/22, M. A.]

Autonomous systems lack the flexibility humans have to adapt to novel circumstances. This may mean that in unexpected situations, autonomous systems make mistakes that humans would not have. As militaries think through how they incorporate autonomy into future weapons, a clear-eyed assessment of the risks associated with autonomous weapons is required. Accidents with autonomous weapons could lead to civilian casualties, fratricide, or unintended escalation in a crisis. Understanding these risks is important for policymakers and acquisition professionals weighing whether to build autonomous weapons, as well as for military commanders who are responsible for the weapons they deploy on the battlefield. Of course, humans are far from perfect in war. They make mistakes that lead to fratricide and civilian casualties as well, and humans also commit deliberate acts of atrocity.5 There are important qualitative differences between the risks associated with semi-autonomous (human in the loop) weapons and autonomous weapons, however: • Without a human in the loop to act as a fail-safe, the consequences of failure with an autonomous weapon could be far more severe than an equivalent semi-autonomous weapon. • For extremely simple autonomous weapons, these risks may be manageable if human operators employ them only in limited, controlled contexts. As autonomous weapons increase in complexity, however, it may be more difficult for human operators to fully understand the boundaries of their behavior and accurately predict under what conditions failures might occur, even if they are unlikely. • While improved test and evaluation can mitigate these risks somewhat, they cannot be eliminated entirely. Failures in complex systems may not be likely, but over a long enough time horizon they are inevitable.

**Military AI systems are prone to technical failures which cause unpredictable escalation risks.**

**International Panel on the Regulation of Autonomous Weapons, 2021** [(iPRAW) coordinated by: German Institute for International and Security Affairs, July “Building Blocks for a Regulation on LAWS and Human Control Updated Recommendations to the GGE on LAWS” https://www.readkong.com/page/building-blocks-for-a-regulation-on-laws-and-human-control-8617434 Acc 2/27/22 TA]

Autonomous functions in weapon systems are based on various enabling technologies including sensors, processors, and software. Most prominent are data-driven techniques, like AI and machine learning. Even though a regulation of LAWS should focus on the human role, understanding the options and limitations of enabling technologies is crucial. By integrating, processing, and analyzing large amounts of data quickly, AI-enabled technologies can offer useful decision support and might furthermore allow for new operational options. At the same time, however, due to their brittleness and opaque nature, they could increase unpredictability and escalation risks and perpetuate biases. In some instances, the application of computational methods in the targeting process can lead to better outcomes than human performance alone would. Nevertheless, our discussions and scenario-based workshops showed the limitations of computational methods as enabling technologies in the military domain and highlighted that they most likely cannot replace the unique judgment of human decision-makers, understood as “the ability to evaluate and combine numerous contextual sources of information” . Furthermore, any complex computational system consists of modular subsystems, each of which has inherent limitations and points of failure. Applying multiple computational systems across each step of the targeting cycle may result in cumulative failures that can be hard to anticipate and lead to hazardous and undesired outcomes. Any system that executes sequential processes, such as selecting and engaging targets, can be subject to path dependencies where errors or decisions, in any step, can propagate and reverberate throughout the rest of the targeting process.

**Autonomous swarm technology is excessively prone to malfunction due to complexity and hacking**

**Sauer, 2021 - Senior Research Fellow at Bundeswehr University** [Frank serves on the International Panel on the Regulation of Autonomous Weapons IRRC No. 913 March “Stepping back from the brink: Why multilateral regulation of autonomy in weapons systems is difficult, yet imperative and feasible” https://international-review.icrc.org/articles/stepping-back-from-brink-regulation-of-autonomous-weapons-systems-913 Acc 4/5/22 TA]

New operational vulnerabilities The flip side of the force multiplication effect that militaries hope for with this diffusion-prone technology is scalability, creating the potential for weaker parties to change the power dynamics between themselves and their adversaries. The weaponization of simple, commercially available drones by the so-called Islamic State group and the attack against the Saudi Aramco oil facility give non-autonomous foretastes of what is to come and demonstrate that, advanced aerial defence capabilities notwithstanding, new vulnerabilities are on the rise. Particularly from the point of view of US ground forces, having to face serious threats from above after decades of air dominance represents a paradigm shift.47 The United States is thus already being forced to rethink its air defence capabilities by intensifying the development of lasers and microwaves. Conventional solutions, such as Stinger missiles, are not only unsuitable for defence against the swarms of small, cheap, disposable drones that autonomy now renders possible, they are also not cost-effective. Whether the new defensive systems can remedy this situation is still open for debate.48 Suffice it to say that the combination of cheap unmanned systems, autonomy and swarm behaviour creates new risks in general, for troops on the battlefield, for command and control infrastructure and for senior leaders in so-called decapitation scenarios.49 As argued above, the possible elimination of the remote control link is a key incentive for having more autonomy in a weapons system – but handing control over to the machine opens up new attack vectors as well. Feeding the system spoofed GPS data is one example; in 2011, Iran was seemingly able to hijack an autonomously navigating US drone in this manner.50 What is more, systems relying on machine learning that makes use of deep neural networks,51 which currently represent the state of the art in the field of computer vision, are also particularly susceptible to manipulation. Some reflective tape on a stop sign, for example, can fool a self-driving car's image recognition system. This susceptibility to error is a tricky but eventually solvable problem in a civilian application such as self-driving cars. Training data is plentiful and easily available, and self-driving cars are designed to operate cooperatively in a tightly regulated environment. The battlefield presents itself very differently – it is characterized by a paucity of data and much greater degrees of unpredictability and vulnerability.52 After all, an adversary will of course always try to deceive and tamper with your systems. Research on adversarial examples53 suggests that computer vision will leave autonomous weapons systems open to manipulation by tampering with the environment that the machines perceive54 or even by retraining them if they continue learning during their deployment.55 Facial recognition for targeting purposes would be quite easy to fool and defeat, too, as the rapid development of countermeasures against domestic surveillance demonstrates.56 As the complexity of the software driving a weapons system increases, so does the number of bugs it contains. Such programming errors can have critical effects, including friendly fire.57 Normal accidents theory58 suggests that mistakes are basically inevitable. They occur even in domains with extremely high safety and security standards, such as nuclear power plants or manned space travel.59 The software industry can currently reduce the number of bugs to 0.1–0.5 errors per 1,000 lines of code, which means that complex military systems with several million lines of code, such as the software for the F-35 fighter jet, come with thousands of software errors.60 The unavoidable reality of regularly having to update the systems complicates this issue further;61 it is a potential source of new bugs and new errors arising from interactions between newer and older software. Machine learning systems generate specific difficulties because they present themselves as “black boxes” which cannot be debugged the way conventional software can, meaning that they cannot be selectively cleared of specific errors.62 Finally, weapon autonomy evokes a new proneness to errors in regard to any remaining interactions with human operators. Here, automation bias comes into play – that is, the uncritical, unfounded trust in the functioning of a system.63 Put simply, autonomous systems might operate incorrectly over periods of time without anyone noticing.64 A human making a mistake can understand the situation and correct for it, but unmonitored LAWS will not be able to understand and critically reflect in real time the way humans do.65 This gives rise to risks of unintended military escalation.

#### Adoption of AI in the military increases the risks of mistakes and accidents – Safety and reliability are still issues

Nurkin and Konaev, 2022 - senior fellows at the Center for Strategy and Security at the Atlantic Council [Tate and Margarita, 5-25-2022, “Eye to eye in AI: Developing artificial intelligence for national security and defense” Atlantic Council <https://www.atlanticcouncil.org/in-depth-research-reports/report/eye-to-eye-in-ai/> ARD]

Demonstrating that the age of deployed AI is indeed here, in September 2021 Secretary of the Air Force Frank Kendall announced that the Air Force had “deployed AI algorithms for the first time to a live operational kill chain.”52 According to Kendall, the objective of incorporating AI into the targeting process is to “significantly reduce the manpower intensive tasks of manually identifying targets— shortening the kill chain and accelerating the speed of decision-making.”53 The successful use of AI to support targeting constitutes a milestone for AI development, though there remain ethical, safety, and technical challenges to more complete adoption of AI in this role. For example, a 2021 DoD test highlighted the problem of brittle AI. According to reporting from Defense One, the AI-enabled targeting used in the test was accurate only about 25 percent of the time in environments in which the AI had to decipher data from different angles—though it believed it was accurate 90 percent of the time—revealing a lack of ability “to adapt to conditions outside of a narrow set of assumptions.”54 These results illustrate the limitations of today’s AI technology in security-critical settings, and reinforce the need for aggressive and extensive real-world and digital-world testing and evaluation of AI under a range of conditions. The ethics and safety of AI targeting could also constitute a challenge to further adoption, especially as confidence in AI algorithms grows. The Air Force operation involved automated target recognition in a supporting role, assisting “intelligence professionals”—i.e., human decision-makers.55 Of course, DoD has a rigorous targeting procedure in place, of which AI-enabled targeting algorithms would be a part, and that, thinking further ahead, autonomous systems would have to go through. Still, even as they are part of this process and designed to support human decisions, a high error rate combined with a high level of confidence in AI outputs could potentially lead to undesirable or grave outcomes.

#### Autonomous systems face inherent failure risks from malfunctions and bugs

Scharre, 2016 -- Vice President and Director of Studies at CNAS [Paul, Feb 2016, Center for New American Security, “Autonomous weapons and operational risk”, https://s3.us-east-1.amazonaws.com/files.cnas.org/documents/CNAS\_Autonomous-weapons-operational-risk.pdf?mtime=20160906080515&focal=none, Acc 6/21/22, M. A.]

The risk in employing an autonomous system is that the system might not perform the task in a manner that the human operator intended. There are a number of reasons why an autonomous system might begin performing inappropriately, from simple malfunctions and software bugs to more complex system failures, changing environmental conditions, hacking, and human error. When these failures can be anticipated in advance, human operators can account for these limitations. When failures are unanticipated, however, the result can be autonomous systems that slip out of control. Understanding the likelihood and consequences of a loss of control is essential to assessing the risk in employing autonomous systems.

#### Neural Networks for AI increase mistakes – the “Black Box” makes it impossible for researchers to examine AI decisions.

Scharre, 2016 -- Vice President and Director of Studies at CNAS [Paul, Feb 2016, Center for New American Security, “Autonomous weapons and operational risk”, https://s3.us-east-1.amazonaws.com/files.cnas.org/documents/CNAS\_Autonomous-weapons-operational-risk.pdf?mtime=20160906080515&focal=none, Acc 6/21/22, M. A.]

Neural networks and the black box The challenge of complexity, while problematic for complex rule-based systems, is even more difficult for cutting-edge artificial intelligence (AI) systems that employ neural networks. Neural networks do not perform rule-based calculations like most computers. Instead, they learn by exposure to large data sets. As a result, the internal structure of the network that generates output can be opaque to the designers—a “black box.” Even more unsettling, for reasons that may not be entirely clear to AI researchers, the neural network sometimes can yield odd, counterintuitive results. A study of visual classification AIs using neural networks found that while the AIs were able to generally identify objects as well as humans, in some cases the AIs made confident identifications of objects that were not only incorrect, but that looked vastly different from the purported object to human eyes. The AIs interpreted images that to the human eye looked like static or abstract wavy lines as animals or other objects, and asserted greater than 99.6% confidence in their estimation.

### Brittle - Automation Bias

**Fast development of AI causes catastrophic accidents – humans will assume that AI is working even when it makes mistakes.**

**Horowitz and Kahn, 2022 - Senior and Research Fellows for Defense Technology and Innovation at the Council on Foreign Relations** [Michael and Lauren, with Laura Samotin, May/June Foreign Affairs “A Force for the Future A High-Reward, Low-Risk Approach to AI Military Innovation” [https://www.foreignaffairs.com/articles/united-states/2022-04-19/force-future Acc 5/28/22](https://www.foreignaffairs.com/articles/united-states/2022-04-19/force-future%20Acc%205/28/22) TA]

RISKY BUSINESS Given the stakes, the defense establishment is right to worry about Washington’s torpid pace of defense innovation. But outside the government, many analysts have the opposite fear: if the military moves too quickly as it develops AI weaponry, the world could experience deadly—and perhaps even catastrophic—accidents. It doesn’t take an expert to see the risks of AI: killer robots have been a staple of pop culture for decades. But science fiction isn’t the best indicator of the actual dangers. Fully autonomous, Terminator-style weapons systems would require high-level machine intelligence, which even optimistic forecasts suggest is more than half a century away. One group of analysts made a movie about “Slaughterbots,” swarms of autonomous systems that could kill on a mass scale. But any government or nonstate actor looking to wreak that level of havoc could accomplish the same task more reliably, and cheaply, using traditional weapons. Instead, the danger of AI stems from deploying algorithmic systems, both on and off the battlefield, in a manner that can lead to accidents, malfunctions, or even unintended escalation. Algorithms are designed to be fast and decisive, which can cause mistakes in situations that call for careful (if quick) consideration. For example, in 2003, an MIM-104 Patriot surface-to-air missile’s automated system misidentified a friendly aircraft as an adversary, and human operators did not correct it, leading to the death by friendly fire of a U.S. F-18 pilot. Research shows that the more cognitively demanding and stressful a situation is, the more likely people are to defer to AI judgments. That means that in a battlefield environment where many military systems are automated, these kinds of accidents could multiply.

#### Automation Bias – AI causes humans to depend Too Much on automation, undermining effectiveness military.

Horowitz, 2019 - Professor of Political Science, University of Pennsylvania [Michael C. May 2“When Speed Kills: Autonomous Weapon Systems, Deterrence, and Stability” https://ssrn.com/abstract=3348356]

Autonomous Weapon System Deployment The history of cruise and ballistic missile development – and specifically the desire for faster and more accurate missiles – illustrates why countries might seek to accelerate the development and deployment of LAWS. In the Cold War, for example, increased accuracy was mostly viewed as a good in itself, or a natural development.27 The promise of being able to operate faster than adversaries on the battlefield, or get inside their decision loops, are also constant goals of planners, as is the ability to use force more precisely. In combination, at the outset, these autonomous characteristics of weapon systems should make their development and deployment more likely. A key factor that will influence the deployment of LAWS is trust in the system. As referenced above, the automatic mode on the Phalanx and related CIWS systems provides an approximation of a lethal autonomous weapon system. Moreover, Iron Dome (Israel) and the PAC- 3 Patriot (US and others) integrate autonomy into targeting and firing decisions. A track record of effectiveness can lead operators to trust them too much through a process called automation bias, as they outsource judgment to algorithms.28 For example, in 2003, the Patriot missile system led to two fratricides in the US-led invasion of Iraq. In both cases, the autonomous characteristics of the system proved too brittle for the real ambiguities of combat, and operators trusted the sensors too much.29

### Brittle – Escalation Impacts

**Malfunctions in AI weapons systems pose a high risk of miscommunication and misperception which causes war and escalation.**

**Egel, 2021 - Nuclear Security Visiting Fellow at the Truman Center for National Policy** [Naomi, PhD candidate at Cornell, January, Geneva Centre for Security Policy Strategic Security Analysis” Issue 16, “Reducing Military Risks through OSCE Instruments: The Untapped Potential in the European Arms Control Framework” https://www.gcsp.ch/publications/reducing-military-risks-through-osce-instruments-the-untapped-potential-in-the-European-arms-control-framework Acc 2/27/22 TA]

Risks posed by AWS AWS are weapons that can perform some or all of their functions (including target detection, selection, and engagement) without requiring intervention by a human operator. This definition includes weapons for which a human operator can intervene to override an autonomous function, as well as ‘semi-autonomous’ weapons systems in which some functional decisions are made by human operators and some by the weapons system. Fully autonomous weapons that lack any form of human operator intervention have not yet been deployed. Examples of AWS include air defence systems, armed drones, and active protection systems for armoured vehicles. Although a degree of autonomy in weapons systems is not new (e.g. landmines could be considered partially autonomous in that they are not activated by an operator), the heightened levels of autonomy in recently developed AWS pose significant military risks. AWS are increasingly important in strategic planning and military operations. Although some policymakers argue that AWS can help to reduce risks to both combatants and civilians (e.g. the 2018 US Department of Defense Artificial Intelligence Strategy), the increasing level of autonomy that is being built into weapons systems poses considerable military risks, most notably those of miscommunication, misunderstanding and the inadvertent escalation of a conflict. These risks posed by AWS undermine the stability of the balance of military power among nations. For example, one major purported advantage of AWS is the speed with which they can react and respond to changes in their environment. However, the increased speed at which AWS operate also hampers humans’ ability to correct mistakes made by AWS, which increases the risk of the inadvertent escalation. More broadly, by reducing human involvement in decision-making processes, AWS increase the risk of signals and actions being misinterpreted by other parties, both when autonomy is a feature of a particular weapon and when it is involved in the decision-making cycle. Although AWS do pose military risks on their own (e.g. they could malfunction), the greater risks they pose involve how other actors could (mis)interpret and react to an actor’s use of these weapons systems. Different types of autonomy have varied implications for military risk and the stability of the international system. For example, AWS that rely on unsupervised machine learning to improve their functioning are inherently unpredictable, since the user does not specify the nature or directions of the improvements made and does not know what elements the programme controlling an AWS uses to sort and categorise information. This increases the risk of the inadvertent escalation of a conflict if a system ‘learns’ to respond in ways that appear escalatory or unclear to others. In contrast, AWS that are pre-programmed and do not ‘learn’ on their own are more predictable and thus pose a lower risk of miscommunication and inadvertent escalation. However, the algorithms used in AWS are generally designed to react to a set of information with specific parameters. But even if these algorithms can accurately classify information and produce reliable results in training scenarios, the ‘fog of war’ endemic to military engagement makes it extremely difficult to design AWS for real-world military engagement. Even taking civilian autonomous systems only slightly out of the context in which they have been programmed to operate has produced unpredictable – and in some cases disastrous – results (e.g. self-driving cars crashing into other cars because of what appears to humans to be a minor change in the context in which they operate). Although miscalculation, miscommunication and inadvertent escalation are certainly not new military risks, the increasing use of AWS in military operations heightens these risks in new ways.

#### The AI Arms Race will cause miscalculation and accidents – machine speed removes checks on escalation

Sharre, 2018 - director of the technology and national security program at the Center for a New American Security [Paul, 12 September, “Ultrafast computing is critical to modern warfare. But it also ensures a lot could go very wrong, very quickly.” <https://foreignpolicy.com/2018/09/12/a-million-mistakes-a-second-future-of-war/> ST]

Militaries around the globe are racing to build ever more autonomous drones, missiles, and cyberweapons. Greater autonomy allows for faster reactions on the battlefield, an advantage that is as powerful today as it was 2,500 years ago when Sun Tzu wrote, “Speed is the essence of war.” Today’s intelligent machines can react at superhuman speeds. Modern Chinese military academics have speculated about a coming “battlefield singularity,” in which the pace of combat eclipses human decision-making. The consequences of humans ceding effective control over what happens in war would be profound and the effects potentially catastrophic. While the competitive advantages to be gained from letting machines run the battlefield are clear, the risks would be grave: Accidents could cause conflicts to spiral out of control. Consider what has already happened with stock markets, where computers use algorithms to make decisions so quickly that microseconds make a difference of millions of dollars. Such trading has made brokers huge amounts of money—but has also produced extreme flash crashes that can send markets tumbling in minutes. Regulators have managed these risks by installing circuit breakers that can take a stock offline if the price moves too quickly, but battlefields lack these fail-safes. Flash crashes are bad enough; a flash war would be downright disastrous. Humans have already ceded control to machines in certain military domains. At least 30 countries—with Israel, Russia, and the United States leading the pack—employ human-supervised autonomous weapons to defend bases, vehicles, and ships. These weapons systems, such as the ship-based Aegis combat system, can detect incoming rockets and missiles and, if human supervisors do nothing, respond on their own by firing to eliminate the threat. Such automated responses allow the systems to defend against what are known as saturation attacks, in which salvos of missiles or rockets are launched at a target with such little notice that they could overwhelm human operators. For the time being, autonomous weapons such as these are used purely to protect human-occupied installations or vehicles. Humans supervise the weapons’ operation in real time and can intervene if necessary. Future autonomous weapons could lack these safeguards, however. A number of advanced militaries—including those of China, France, Israel, Russia, the United Kingdom, and the United States—are currently developing stealth combat drones intended to penetrate an adversary’s airspace. Once deep behind enemy lines, these drones might find their communications jammed, so they’re being designed to ensure they can continue to operate on their own. Most countries have not explained how their drones will operate under such circumstances and what rules of engagement they will follow. Countries could require their drones to get human authorization before launching any attacks. Doing so would allow the drones to bomb preapproved fixed targets but would require them to report back and get permission before attacking any newly discovered quarries. Such an approach sounds good in theory, but the problem is that these days many high-priority targets, such as air defense systems and ballistic missile launchers, are highly mobile. This mobility will increasingly tempt military planners to delegate lethal decision-making authority to machines, since doing so could give them an edge in reaction time. No battlefield is static, and the ability to rapidly react to a dynamic environment is critical to mission success—whether in the air, on the ground, or in cyberspace. Air combat strategists call this the OODA (Observe, Orient, Decide, and Act) loop in dogfighting. In the OODA loop paradigm of combat, pilots win dogfights not simply because they enjoy the best hardware but because they assess and react to their situations faster than their opponents, although better sensors and maneuverability might help shorten reaction times. Since machines can react faster than humans, automation will offer tremendous advantages in this competition. That means that the same competitive pressures that led to the creation of systems such as the Aegis could soon be introduced on a wider scale. A military that fully integrated its autonomous systems could always stay one step ahead of its enemy in combat and present a constantly shifting threat. As Gen. Paul Selva, the vice chairman of the U.S. Joint Chiefs of Staff, told the Senate Armed Services Committee in July 2017, “It is very compelling when one looks at the capabilities that artificial intelligence can bring to the speed and accuracy of command and control and the capabilities that advanced robotics might bring to a complex battlespace, particularly machine-to-machine interaction in space and cyberspace, where speed is of the essence.” Yet speed is not an unadulterated good. Forces that react to the enemy so quickly that their own commander does not understand what is happening could risk a breakdown in command and control, a problem that military leaders have struggled with for millennia. Today, email and chat messaging have replaced horses and flags, but the fundamental problem persists. Militaries counter this inherent friction between orders from above and the reality on the ground with a concept known as commander’s intent: succinct goal-oriented statements issued to subordinates that explain the desired goal of a particular mission, and thus ensure that they stick to the general plan, but also allow them the flexibility to adapt to events on the ground. Such statements prevent forces from becoming too predictable and give subordinates the freedom to overcome obstacles in novel ways.

#### Brittle AI causes accidents and miscalculation – Machine learning breaks down when the real world does not act like their simulations

**Flournoy, Haines et al 2020 - former Under Secretary of Defense for Policy and Director of National Intelligence** [Michèle and Avril, Gabrielle Cheftiz Special Assistant to the Under Secretary of Defense for Policy October, “Building Trust through Testing Adapting DOD’s Test & Evaluation, Validation & Verification (TEVV) Enterprise for Machine Learning Systems, including Deep Learning Systems” https://cset. georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf Acc 6/23/22 JZ]

ML/DL systems are not robust and it is difficult to characterize system performance. A fundamental challenge of ML/DL is its brittleness; it has trouble functioning correctly if the inputs or environmental conditions change. However, testing these systems in all possible scenarios and with all ranges of inputs is simply not feasible. It is nearly impossible to predict all the ways a system could break, or an adversary could manipulate or spoof it, which is partly why these systems may be especially vulnerable to adversarial attacks. As these systems become more sophisticated, as is the case with deep learning, their output becomes even less transparent, making it harder to determine the conditions under which they might fail and what steps could correct system behavior. Even when operating under the best conditions (within the same distribution of inputs or environmental conditions present during training), DL models generally don’t work to the reliability standards needed for safety-critical systems,10 and therefore may not be appropriate for these applications, at least given the current state of the technology. ML/DL systems are also particularly vulnerable to operational edge cases (both unintended and intended), which are cases that occur beyond the bounds of a system’s operational envelope or normal operating parameters. Because it is very difficult to characterize the actual performance envelop of these systems, it will be important to prioritize stress testing system performance with boundary conditions. Further, interactions within and between systems (including foreign autonomous systems) can induce unintended consequences and are even more complex to predict or understand. The potential for unintended engagement or escalation is even greater when U.S. and/or adversary systems have the sorts of advanced autonomy features that deep learning can enable, and their interaction cannot be studied or fully tested in advance of deployment.

#### Military AI causes miscalculation and escalation due to acceleration of combat and lack of human control

Horowitz and Scharre, 2021 - Director of the Emerging Capabilities Policy Office in the Office of the Under Secretary of Defense for Policy and  Vice President and Director of Studies at CNAS [Michael, Paul, January 12 2021, “AI and International Stability: Risks and Confidence- Building Measures”, <https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures>, Acc 6/18/22. M.A.]

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](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn5) 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](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn6) 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](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn7) while some Americans have coined the term “hyperwar” to refer to a similar idea.[8](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn8) 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.

#### Autonomous Weapons systems inevitably cause catastrophic accidents and escalation due to the speed of AI decision making

Sharre, 2018 - director of the technology and national security program at the Center for a New American Security [Paul, 12 September, “Ultrafast computing is critical to modern warfare. But it also ensures a lot could go very wrong, very quickly.” <https://foreignpolicy.com/2018/09/12/a-million-mistakes-a-second-future-of-war/> ST]

Real-world accidents with existing highly automated weapons point to these dangers. During the initial invasion of Iraq in 2003, the U.S. Army’s Patriot air defense system accidentally shot down two friendly aircraft, killing three allied service members. During the initial invasion of Iraq in 2003, the U.S. Army’s Patriot air defense system accidentally shot down two friendly aircraft, killing three allied service members. The first fratricide was due to a confluence of factors: a known flaw that caused the radar to mischaracterize a descending plane as a missile, outdated equipment, and human error. The second blue-on-blue incident was due to a situation that had never arisen before. In the hectic march to Baghdad, Patriot operators deployed their radars in a nonstandard configuration likely resulting in electromagnetic interference between the radars that caused a “ghost track”—a signal on the radars of a missile that wasn’t there. The missile battery was in automatic mode and fired on the ghost track, and no one overruled it. A U.S. Navy F-18 fighter jet just happened to be in the wrong place at the wrong time. Both incidents were flukes caused by unique circumstances—but also statistically inevitable ones. Coalition aircraft flew 41,000 sorties in the initial phases of the Iraq War, and with more than 60 allied Patriot batteries in the area, there were millions of possible interactions, seriously raising the risk for even low-probability accidents. Richard Danzig, a former U.S. secretary of the Navy, has argued that bureaucracies actually systematically underestimate the risk of accidents posed by their own weapons. It’s also a problem that it’s nearly impossible to fully test a system’s actual performance outside of war. In the Iraq invasion, these accidents had tragic consequences but did not alter the course of the war. Accidents with fully autonomous weapons where humans cannot intervene could have much worse results, causing large-scale fratricide, civilian casualties, or even unintended attacks on adversaries.Attempts at arms control go back to antiquity, from the Bible’s prohibition on wanton environmental destruction in Deuteronomy to the Indian Laws of Manu that forbade barbed, poisoned, or concealed weapons. In the intervening centuries, some efforts to ban or regulate certain weapons have succeeded, such as chemical or biological weapons, blinding lasers, land mines, cluster munitions, using the environment as a weapon, placing weapons in space, or certain delivery mechanisms or deployment postures of nuclear weapons. Many other attempts at arms control have failed, from the papal decrees denouncing the use of the crossbow in the Middle Ages to 20th-century attempts to ban aerial attacks on cities, regulate submarine warfare, or eliminate nuclear weapons. The United Nations began a series of meetings in 2014 to discuss the perils of autonomous weapons. But so far the progress has been far slower than the pace of technological advances. Despite that lack of success, a growing number of voices have begun calling for a ban on autonomous weapons. Since 2013, 76 nongovernmental organizations across 32 countries have joined a global Campaign to Stop Killer Robots. To date, nearly 4,000 artificial intelligence and robotics researchers have signed an open letter calling for a ban. More than 25 national governments have said they endorse a ban, although none of them are major military powers or robotics developers. But such measures only tend to succeed when the weapons in question are of marginal value, are widely seen as especially horrific or destabilizing, are possessed by only a few actors, are clearly distinguished from other weapons, and can be easily inspected to verify disarmament. None of these conditions applies to autonomous weapons. Even if all countries agreed on the need to restrain this class of arms, the fear of what others might be doing and the inability to verify disarmament could still spark an arms race. Less ambitious regulations could fare better, such as a narrow ban on anti-personnel autonomous weapons, a set of rules for interactions between autonomous weapons, or a broad principle of human involvement in lethal force. While such modest efforts might mitigate some risks, however, they would leave countries free to develop many types of autonomous weapons that could still lead to widespread harm. Humanity stands at the threshold of a new era in war, in which machines will make life-or-death decisions at speeds too fast for human comprehension. The risks of such a world are real and profound. Autonomous weapons could lead to accidental death and destruction at catastrophic scales in an instant. The unrestrained pursuit of fully autonomous weapons could lead to a future where humans cede control over what happens on the battlefield, but the critical decisions about how this technology is used still rest in human hands.

### Brittle - Black Box Scenario

#### AI can disrupt command structure if commanders cannot explain why AI has made specific decisions. AI must include “explainability” into their systems.

Fields et al, 2016 – Chair of the Defense Science Board [Craig, with Dr. Ruth David Dr. Paul Nielsen – co chairs, June “Defense Science Board Summer Study on Autonomy” https://www.hsdl.org/?view&did=794641, LMSi]

Autonomy in support of command and control One of the most contentious applications of autonomy is for command and control in military operations or warfighting, but the potential benefits are real. The time for concepts of operations (CONOPs) development, target selection, and mission assignments can be significantly reduced, and, during combat operations, commanders could be better equipped to respond to changing situations and redirect forces. While commanders understand they could benefit from better, organized, more current, and more accurate information enabled by application of autonomy to warfighting, they also voice significant concerns. Implementation of autonomous capabilities will require significant changes in command and control concepts. A previous DSB study acknowledged the importance of addressing command and control of autonomous systems, but found that it was an unsolved problem and did not address it further.21 Whether mediated by man or machine, all acts, but especially acts related to warfighting, must be executed in accordance with policy and so, in some sense, there is no completely autonomous behavior. Any use of autonomy must conform to a substantive command and control regime laying out objectives, methods and express limitations to ensure that autonomous behavior meets mission objectives while conforming to policy.22 In fact, most autonomous combat systems will and should act under the guidance and instructions of a field commander who will exercise direct oversight. Initial use of autonomous systems for combat will likely assist commanders and their staffs with developing situational awareness and planning missions. The volume and velocity of the data used by the underlying system will change the pace of operations. For example, current air operations planning often involves a several-day process, beginning with identification of objectives; moving to general target selection; to intelligence support identifying particular targets; to determination of final plans by the Commander balancing risks and potential value in achieving objectives; coordination of air, land, and sea assets; and, finally, mission execution and battle damage assessment. Given human limitations, each stage results in static point in time, and planning is generally a linear and timeconsuming process. Because planning often needs to respond to new information, autonomous systems will greatly accelerate the pace of information update and can suggest significant plan changes far more quickly. Commanders will not only need to develop models to calibrate understanding of machine generated information, but will need better automated tools to adjust to the pace of update. This requires careful design of autonomous systems so they can explain and justify recommendations in principled terms that build trust between commanders, staff, and machine-generated output as well as develop a concrete understanding of mission outcomes that depend critically on the data and methodology employed by the autonomous systems. 21 Commanders and their staff need to both exercise oversight of fielded autonomous systems, and have access to efficient mechanisms to develop control datasets that teach autonomous systems so they can react in well-understood ways to unexpected and subtle changes. This requires new approaches to human factors that are informed by warfighting practice and tradition. It also requires development of commander and staff training.

#### Military AI will degrade decision making because the anarchic nature of war denies access to data necessary for predictions.

Goldfarb and Lindsay, 2022 – Chair in AI and health care at the Univ of Toronto, Prof of Cybersecurity at the Georgia Tech [Avi, Jon, 2/25/22, <https://doi.org/10.1162/isec_a_00425>, “Prediction and Judgement: Why Artificial Intelligence Increases the Importance of Humans in War” 6/18/22, LND]

Numerous empirical studies of civilian AI systems have validated the basic finding that AI depends on quality data and clear judgment.[31](javascript:;) To the extent that decision-making is indeed a universal activity, it is reasonable to expect models of it to apply to militaries. But we caution against generalizing from the business world to military affairs (and vice versa). Commercial and military organizations perform dissimilar tasks in different contexts. Militaries are only infrequently “in business” because wars are rare events.[32](javascript:;) Objectives such as “victory” or “security” are harder to define than “shareholder value” or “profit.” combatants attempt to physically destroy their competitors, and the consequences of failure are potentially existential. One underappreciated reason why AI has been applied successfully in many commercial situations is because the enabling conditions of quality data and clear judgment are often present. Peaceful commerce generally takes place in institutionalized circumstances. Laws, property rights, contract enforcement mechanisms, diversified markets, common expectations, and shared behavioral norms all benefit buyers and sellers. These institutional features make transactions more consistent and efficient.[33](javascript:;) Consistency, in turn, provides the essential scaffolding for full automation. We expect AI to be more successful in more institutionalized circumstances and for more structured tasks. War, by contrast, occurs in a more anarchic environment. In the international system, according to the intellectual tradition of realism, there are no legitimate overarching institutions to adjudicate disputes, enforce international agreements, or constrain behavior.[34](javascript:;) Actors must be prepared to defend themselves or ally with others for protection. Allies and adversaries alike have incentives to misrepresent their capabilities and interests, and for the same reasons to suspect deception by others.[35](javascript:;) Militarized crises and conflicts abound in secrecy and uncertainty. War aims are controversial, almost by definition, and they mobilize the passions of the nation, for better or worse. We expect the absence of constraining institutions in war to undermine the AI-enabling conditions of quality data and clear judgment. One exception that proves the rule is that a military bureaucracy may be able to provide scaffolding for some military tasks. Robust organizational institutions, in other words, might substitute for weak international institutions. Yet, there are limits to what organizations can accomplish in the inherently uncertain and contested environment of war. The specific context of data and judgment will determine the viability of automation for any given task.

#### AI undermines command structure because machines cannot exercise leadership when commanders disagree with AI recommendations.

Goldfarb and Lindsay, 2022 – Chair in AI and health care at the Univ of Toronto, Prof of Cybersecurity at the Georgia Tech [Avi, Jon, 2/25/22, <https://doi.org/10.1162/isec_a_00425>, “Prediction and Judgement: Why Artificial Intelligence Increases the Importance of Humans in War” 6/18/22, LND]

The second characteristic of a well-specified AI utility function is that all stakeholders should agree on what goals to pursue. When it is difficult for people to agree on what to optimize, transparent institutional processes for evaluating or aggregating different preferences may help to validate or legitimate decisions that guide AI systems. Unfortunately, consensus becomes elusive as “genius” becomes more geographically distributed, socially collaborative, and technically exacting.[54](javascript:;) In an ethnography of divisional command in Afghanistan, Anthony King writes that “a general must define a mission, manage the tasks of which it is comprised and motivate the troops.”[55](javascript:;) The first of these three factors—specifying positive objectives and negative limitations— is the consummate function of judgment; AI offers little help here. AI might provide some support for the second factor, oversight and administration, which involves a mixture of judgment and prediction. The third factor is leadership, which is fundamentally a judgment problem insofar as leaders attempt to socialize common purposes, values, and interpretations throughout an organization. Again, AI is of little use for issues of leadership, which become more important as organizations become geographically and functionally distributed: “The bureaucratic expertise of the staff has been improved and their cohesiveness has been condensed so that they are now bound in dense solidarity, even when they are not co-present.”[56](javascript:;) Indeed, “decision-making has proliferated” in all three areas—strategy, management, leadership— because a “commander can no longer direct operations alone.”[57](javascript:;) According to King, the commanding general is now less of a central controller and more of a social focal point for coordinating the complex interactions of “the command collective.” Collective command, however, is a collective action problem. In some cases, standard operating procedures and socialization rituals can simplify judgment tasks. King finds that “command teams, command boards, principal planning groups and deputies have appeared to assist and to support the commander and to manage discrete decision cycles to which the commander cannot attend.”[58](javascript:;) Yet, in other cases, personnel from different services, branches, or units may disagree over how to interpret even basic tactics, techniques, and procedures.[59](javascript:;) Disagreement may turn into controversy when mission assignments fall outside the scope of what professionals deem right or appropriate, as when armies are tasked with counterinsurgency, air forces are tasked with close air support, or cultural preferences clash.[60](javascript:;) More serious disagreements about war aims and military methods can emerge within the civil-military chain of command or among coalition partners.[61](javascript:;) Just as data availability and bias vary for any given decision task, we also expect variability in the clarity and consensus of judgment. Any factors that exacerbate confusion or disagreement in military institutions should be expected to make judgment more difficult for AI automation.

## Nuclear AI Advantage

### Nuclear – Early Warning

**Malfunctions in Early Warning AI systems can launch nuclear responses**

**Horowitz and Kahn, 2022 - Senior and Research Fellows for Defense Technology and Innovation at the Council on Foreign Relations** [Michael and Lauren, with Laura Samotin, May/June Foreign Affairs “A Force for the Future A High-Reward, Low-Risk Approach to AI Military Innovation” [https://www.foreignaffairs.com/articles/united-states/2022-04-19/force-future Acc 5/28/22](https://www.foreignaffairs.com/articles/united-states/2022-04-19/force-future%20Acc%205/28/22) TA]

This has frightening implications. AI-enabled machines are unlikely to ever be given the power to actually launch nuclear attacks, but algorithms could eventually make recommendations to policymakers about whether to launch a weapon in response to an alert from an early warning air defense system. If AI gave the green light, the soldiers supervising and double-checking these machines might not be able to adequately examine their outputs and monitor the machines for potential errors in the input data, especially if the situation was moving extremely quickly. The result could be the inverse of an infamous 1983 incident in which a Soviet air force lieutenant arguably saved the world when, correctly suspecting a false alarm, he decided to override a nuclear launch directive from an automated warning system. That system had mistaken light reflecting off of clouds for an inbound ballistic missile.

#### Nuclear AI programs can fail due to programming failures, early warning mistakes or biased data.

**Fitzpatrick, 2019 - former the IISS Non-Proliferation and Nuclear Policy Programme** [Mark, 21 May, “Artificial Intelligence and Nuclear Command and Control,”  [https://www.iiss.org/blogs/survival-blog/2019/04/artificial-intelligence-nuclear-strategic-stability 6/18/22](https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/00396338.2019.1614782%206/18/22) MD]

Skewed early-warning assessments Deep fakes by third parties can be magnified by AI capabilities that fall into the wrong hands and are used to generate false positives. In our scenario, three months before the crisis and start of play, US Cyber Command’s ‘Unified Platform’ for managing and coordinating integrated cyber-, electronic- and information-warfare operations was providing statistically anomalous outputs regarding situational-awareness assessments related to the early warning of selected advanced persistent threat datasets linked to Russian cyber actors. The questionable reports appeared to be skewed by mathematical coefficients derived from large-scale metadata analysis during the beta-testing training phase of the Unified Platform’s software. A workshop background paper noted that AI programmes are driven by the data that they receive – the digital equivalent of the adages ‘you are what you eat’ and ‘garbage in, garbage out’. This is also true of the process by which such programmes initially formulate their pattern-recognition models and evaluative procedures for use in future contexts. The malign manipulation of input data can not only pervert the output of AI functions in specific instances, but also undermine the reliability of an entire algorithm if accomplished during the ‘training’ phase for such programmes. Our scenario showed vividly how this could play out. Non-malign, human-design factors can also corrupt AI assessments. Because human beings define an AI system’s algorithms and curate its training data, they can unintentionally insert their own biases into the system. This can cause the system to behave in unintended ways that may be undetectable to its operators due to the feedback loop created by those very biases. In addition, nuclear command- and-control personnel also faced a black-box problem in determining how or why a system came to a certain conclusion. While system inputs and outputs can be observed, the speed and scale of system processes make it difficult for personnel to isolate the logic behind any particular prediction. Once the operation of AI systems is triggered, humans are unable to monitor the systems’ decision calculus in real time. In the exercise, the US team was aware of this range of potential problems with respect to AI-driven assessments provided by Cyber Command’s Unified Platform. They worried that Russian strategic-warning systems might be similarly skewed. In particular, the Americans realised that the false positives produced by Russia’s strategic-warning system need not have been caused by a malign actor, as they could also have resulted from a problem intrinsic to the algorithm. To preclude Russia from perceiving any US effort to investigate as an attempt to manipulate Russia’s early-warning and command-and-control systems, the US team decided not to conduct any reconnaissance. In turn, the scenario in round two had Kremlin officials learning of a prior intelligence operation by the Main Directorate of the General Staff of the Russian Armed Forces (GRU), Russia’s military-intelligence agency, injecting skewed training data into AI-driven situational-awareness algorithms of the US military during their developmental phase. The objective was to prevent American AI systems from being able to recognise Russia as an aggressor. The background papers for the exercise had noted how data-poisoning attacks introducing skewed inputs into the training dataset for an AI machine-learning process can distort an AI system’s output by degrading its ability to distinguish between good data and bad data. These types of operations are usually carried out through techniques known as content generation, feedback weaponisation, perturbation injection and man-in-the-middle attacks. However, an attacker usually will not have direct access to the actual training data. To overcome this obstacle, the attacker will target the actors or methods used to collect and store machine-learning training data, such as cloud graphics-processing units, web-based repositories, and contractors or third-party service providers. Understanding that early-warning assessments from its own AI-driven software could be skewed, the Russian team chose largely to ignore the AI algorithms that were producing results so disparate from the reality they otherwise perceived, while also deciding not to reveal any analytic problems they were encountering to their adversaries. Unaware that the Russians had, in effect, pulled the plug on the AI-driven early-warning system, the US team spent considerable time trying to persuade their counterparts to investigate its flaws. The US team also considered offering to work with Russia to neutralise any malicious third-party-introduced software problems. They recognised, however, that any such effort might itself be seen as an attempt to compromise the Russian command-and-control system. Indeed, the Russians reacted defensively. With team members playing to stereotype, trilateral meetings and a session of the United Nations Security Council ended inconclusively amid mutual recriminations.

### Nuclear – Escalation Impacts

**Military AI undermines nuclear deterrence – the speed of reactions gives and incentive for preemption**

**Shah, 2019 - Research Assistant at the Center for International Strategic Studies** [Syed Sadam CISS Insight Vol.VII, No.2 “The Perils of AI for Nuclear Deterrence” https://journal.ciss.org.pk/index.php/ciss-insight/article/download/10/9 Acc 5/25/22 TA]

AI based conventional weapons will have an advantage over human conventional operations. During the simulation exercise at the US air force research laboratory (AFRL), AI platform alpha (a software used to operate autonomous flying jets) defeated all other opponents including expert US pilots by using fuzzy tree methodology.18 This happened even when in several circumstances alpha was disadvantaged with respect to air technology, payloads, and aircraft model.19 The faster and more accurate delivery platforms like hypersonic missiles, combined with advanced AI algorithms will generate a greater sense of confidence in a state’s ability to strike first. For instance, hypersonic cruise missiles coupled with the AI guidance system will take only a few minutes to reach their targets. This will not only add complexity to the decision-making, but states with lesser or no AI based capabilities may resort to the first strike in order to avoid the threat of decapitation or to offset the disadvantage of technological asymmetry. In any case, the use of conventional AI would question the state of mind and perceptions on both sides, and would further deteriorate the nuclear deterrence and strategic stability.

**Military AI increases the risk of nuclear escalation because autonomous weapons can Entangle nuclear and conventional forces.**

**Sauer, 2021 - Senior Research Fellow at Bundeswehr University** [Frank serves on the International Panel on the Regulation of Autonomous Weapons IRRC No. 913 March “Stepping back from the brink: Why multilateral regulation of autonomy in weapons systems is difficult, yet imperative and feasible” https://international-review.icrc.org/articles/stepping-back-from-brink-regulation-of-autonomous-weapons-systems-913 Acc 4/5/22 TA]

Strategic instability It has recently been suggested that AI as a decision-making aid to humans might help improve the performance of nuclear early-warning and command and control systems, thus reducing the risk of false alarms and inadvertent nuclear use.72 That said, calls for complete automation in the nuclear realm – that is, for handing over the decision to use nuclear weapons from humans to machines – are practically non-existent.73 But even with the proverbial push of the button not yet delegated to algorithms, the rush to increase autonomy in military applications and to automatize military processes increases the risk of nuclear stability.74 For instance, the increasing capacities of conventional weapons systems – including weapon autonomy – are beginning to affect the strategic level. This development has been described as the increasing “entanglement” of the nuclear and the conventional realm resulting, for example, from “non-nuclear threats to nuclear weapons and their associated command, control, communication, and information (C3I) systems”.75 Simply put, advanced conventional capabilities increasingly allow for nuclear assets to be put at risk. Autonomy in conventional weapons systems is one such advanced capability, thus feeding into this increasing entanglement and, in turn, deteriorating strategic stability. One specific illustration of this dynamic is the deployment of stealthy unmanned aerial vehicles and the use of “swarming”. Perdix is a swarming test program pursued by the US Air Force. In the future, drone swarms of this type might facilitate the search for dispersed mobile missile launchers. Another example is the use of maritime autonomous systems for hunting nuclear-powered ballistic missile submarines, known as SSBNs. The DARPA-funded Anti-Submarine Warfare Continuous Trail Unmanned Vessel is a program that resulted in the development of an autonomous trimaran called Sea Hunter, which is currently being tested by the US Navy. Its ability to detect and pursue SSBNs could potentially limit the second-strike capabilities of other nuclear powers. These capabilities are just emerging, and neither Perdix nor Sea Hunter, nor their successors, will single-handedly destabilize the global nuclear order. Also, the hypothesis that systems such as Sea Hunter would render the oceans “transparent”,76 virtually nullifying the utility of sea-launched nuclear weapons as a reliable second-strike asset, is hotly debated. Nevertheless, the mere perception of nuclear capabilities becoming susceptible to new risks from the conventional realm is bound to sow distrust between nuclear-armed adversaries. Furthermore, a system like Sea Hunter demonstrates how autonomous weapon technologies are expediting the completion of the targeting cycle, thus putting the adversary under additional pressure and potentially creating “use-them-or-lose-them” scenarios with regard to executing a nuclear second strike. The entanglement problem, which weapon autonomy is feeding into, is further aggravated by an increasing political willingness to use nuclear means to retaliate against non-nuclear attacks on early-warning and control systems or the weapons themselves. The Trump administration's nuclear posture review77 signals that the United States may, from now on, respond with nuclear means to significant, non-nuclear strategic attacks (moving away from a “single-purpose” nuclear deterrence framing for nuclear weapons). Russia has already held this position for some time due to the United States’ advantage in conventional weapons technology. This does not bode well for stability between the two largest nuclear powers. To sum up this section, weapon autonomy not only promises military benefits but also creates new vulnerabilities and, more importantly, contributes to an overall accumulation of strategic risk and instability. Increasing operational speed beyond the capability of human cognition removes humans as a valuable fail-safe against unwanted escalation.

## Cyber Attacks Advantage

### Cyber – Arms Race

#### The AI arms race includes cyber space – the major powers are all rapidly developing AI cyber weapons

Taddeo and Floridi, 2018 - Senior research fellow at the Oxford Internet Institute and Professor of Philosophy and Ethics of Information at the University of Oxford[Mariarosaria, Luciano, Nature.com, “Regulate artificial intelligence to avert cyber arms race,” <https://www.nature.com/articles/d41586-018-04602-6> 6/18/22 MD]

Meanwhile, research on AI for cyber defence is progressing fast. The US is in the lead, technologically. It aims to incorporate AI into its cyber-defence systems by 2019 [4]. The Department of Defense (DoD) has earmarked $150 million for research. The US Defense Advanced Research Projects Agency (DARPA) is developing techniques to do this. Steps have already been taken. In its 2016 Cyber Grand Challenge competition, seven AI systems, developed by academics and private companies from Greece, Russia, Switzerland, and the US played against each other, identifying and targeting opponents’ weaknesses, while finding and patching their own. The DoD will issue the first US report on AI strategies for national defence in May. There is no indication of what its approach will be. Previous documents, like the 2015 US Cyber Strategy or the 2016 National Cyber Incident Response Plan, did not cover autonomous systems, machine learning, or AI. The 2012 DoD Directive on ‘Autonomy in Weapon Systems’ [9] focused on internal procedures for deploying AI, but was silent on when the US would do so in the international arena. AI is a priority for China, which aims to become a world leader in machine learning technologies. In July 2017, the Chinese government issued its AI Development Plan [5]. Military implementation of AI, on the battlefield as well as in cyberspace, is a crucial part of the strategy. But it is unclear to what degree China plans to deploy AI actively in cyber defence. Russia has not released any public documents about its strategies for AI in defence. However, in a video message released in 2017, President Putin expressly referred to AI and stated “whoever becomes the leader in this sphere will become the ruler of the world”. Experts agree that Russia is focusing on developing AI-enhanced tools for its conventional forces. However, since 2014, the Russian National Defense Control Center (NDCC) has also used machine learning algorithms to detect online threats. Allegedly, Russia has pioneered the use of AI to spread disinformation and intervene in the public debate of other nations, like in the case of the US Presidential election and the UK European Union membership referendum, in 2016. Although these operations are not part of national defence strategies, they speak to the level of AI capabilities developed by Russia over the past two years. North Korea has a history of cyberspace aggression. WannaCry in 2016 and the cyber-attack against Sony Picture in 2014 are two good example of North Korea aggressive attitude in cyberspace. The country lacks technical expertise in AI but is likely to want to catch up with its adversaries.

### Cyber – AI Vulnerable to Attacks

#### AI makes systems vulnerable to mistakes or cyber attacks because they are brittle and lack operator trust.

Nurkin and Konaev, 2022 - senior fellows at the Center for Strategy and Security at the Atlantic Council [Tate and Margarita, 5-25-2022, “Eye to eye in AI: Developing artificial intelligence for national security and defense” Atlantic Council <https://www.atlanticcouncil.org/in-depth-research-reports/report/eye-to-eye-in-ai/> ARD]

There are, however, important limitations and drawbacks to AI systems—particularly in operational environments—in large part, because of their brittleness. These systems perform well in stable simulation and training settings, but they can struggle to function reliably or correctly if the data inputs change, or if they encounter uncertain or novel situations. ML systems are also particularly vulnerable to adversarial attacks aimed at the algorithms or data upon which the system relies. Even small changes to data sets or algorithms can cause the system to malfunction, reach wrong conclusions, or fail in other unpredictable ways.20 Another challenge is that AI/ML systems do not typically have the capacity to explain their own reasoning, or the processes by which they reach certain conclusions, provide recommendations, and take action, in a way that is evident or understandable to humans. Explainability—or what some have referred to as interpretability—is critical for building trust in human-AI teams, and is especially important as advances in AI enable greater autonomy in weapons, which raises serious ethical and legal concerns about human control, responsibility, and accountability for decisions related to the use of lethal force. A related set of challenges includes transparency, traceability, and integrity of the data sources, as well as the prevention or detection of adversary attacks on the algorithms of AI-based systems. Having visibility into who trains these systems, what data are used in training, and what goes into an algorithm’s recommendations can mitigate unwanted bias and ensure these systems are used appropriately, responsibly, and ethically. All these challenges are inherently linked to the question of trust explored later in the report.

#### AI creates many vulnerabilities to collapse – there are many methods to exploit data and computing systems

**Artificial Intelligence/Machine Learning Risk & Security Working Group, 2020** [Wharton, ‘Artificial Intelligence Risk & Governance’, <https://ai.wharton.upenn.edu/artificial-intelligence-risk-governance/>]

Key potential risks of AI relate to potential harms that may affect organizations, consumers, or create broader detrimental effects on society. Such potential risks may arise in whole or in part from sources including the data used to train the AI system; potential risks arising from the AI system itself; potential risks arising from the usage of the AI system; and potential risks arising from poor overall governance of the AI system. 2.1 Risk Categorization Various research papers, articles, and discussions have covered the topics of risks associated with AI. It is up to individual financial services firms to categorize potential risks of using AI: however, we have included a suggested approach to AI risk categorization in the financial services industry below. The areas of data related risks, AI/ML attacks, testing and trust, as well as people risk constitute potential areas of risk, which could be subcategorized as illustrated in Figure 1. We address these sub-categories in further detail below. It’s important to note that the applicability and relevance of risks illustrated in Figure 1 are dependent on an individual organization’s risk profile, appetite, and existing controls, and it is up to each firm to determine whether their existing controls are sufficient. 2.1.1 Inadequate Governance Learning Limitations Unlike humans, AI systems lack the judgment and context for many of the environments in which they are deployed. An AI/ML system is generally as effective as the data used to train it and the various scenarios considered while training the system. In most cases, it is not possible to train the AI system on all possible scenarios and data. Lack of context, judgment, and overall learning limitations may play a key role in informing risk-based reviews, and strategic deployment discussions. Data Quality The risk of poor data quality is not unique to AI, but for AI/ML systems, poor data quality could not only limit the learning capability of the system, but could also potentially negatively impact how it makes inferences and decisions in the future. Poor data quality could include incomplete data, erroneous or unsuitable data, stale data, or data used in the wrong context. Such deficiencies may give rise to potentially erroneous or poor predictions, or potentially result in a failure to achieve the intended objectives. 2.1.2 Potential AI/ML Attacks The proliferation of research papers on AI/ML has increased significantly over the last decade, with many of these devoted to potential security weaknesses in AI/ML systems. Most of the known potential attacks against AI/ML systems could be grouped into one of the following categories: data privacy, data poisoning, and model extraction. The likelihood and impact of various potential attacks are specific to each organization’s risk posture and controls. It is possible that some potential attacks illustrated below may not be relevant to a particular organization and may be mitigated by customary security controls. Data Privacy Attacks In data privacy attacks, an attacker is potentially able to infer the data set used to train the model, thereby potentially compromising the privacy of the data. An adversary could potentially infer sensitive information from the training data set by analyzing the parameters or querying the model. Two major attack types in data privacy include membership inference and model inversion attacks. In a membership inference attack, an attacker could potentially determine if a particular record (or set of records) exists in a training data set. Generally, if the attack is successful, an attacker could determine, to a certain degree of probability, whether a particular record was part of the training data set used to train the AI system. In model inversion attacks, an attacker could potentially extract the training data used to train the model directly. Training Data Poisoning Data poisoning is the contamination of data used to train the AI/ML system, negatively affecting its learning process or output. Data poisoning could potentially be used to increase the error rate of the AI/ML system or to potentially influence the retraining process. Some of the attacks in this category are known as “label-flipping” and “frog-boil” attacks. Adversarial Inputs AI systems that use input from external system(s)/ user(s), interpret the input and perform an action, like classifying the input data. An adversary could potentially use a malicious input or a payload explicitly designed to bypass AI systems classifier. Such malicious inputs are known as adversarial inputs. Model Extraction In this attack, an adversary tries to steal the model itself. Model extraction attacks are potentially the most impactful types of AI/ML attacks, as the stolen model could be used as a ‘tool’ to create additional risks. Research into such attacks indicates that, given unlimited ability to query the model, extraction could occur without requiring high levels of technical sophistication and could be accomplished at high speeds.

#### Adapting the military for AI only creates new vulnerabilities – Adversaries will adapt and target weaknesses that AI creates

Goldfarb and Lindsay, 2022 – Chair in AI and health care at the Univ of Toronto, Prof of Cybersecurity at the Georgia Tech [Avi, Jon, 2/25/22, <https://doi.org/10.1162/isec_a_00425>, “Prediction and Judgement: Why Artificial Intelligence Increases the Importance of Humans in War” 6/18/22, LND]

If a military organization can figure out how to recruit, train, and retain highly talented personnel, and to thoroughly reorganize and decentralize its C2 institutions, such reforms may help to inculcate and coordinate judgment. Doing so would enable the military to make the most of human-machine teaming in war. If judgment is a source of military strength, however, then it may also be a political vulnerability. As organizational and political judgment becomes the preeminent source of strength for AI-enabled military forces, we expect that judgment will also become the most attractive target for adversaries. If AI relies on federated data and command structures, then adversaries will pursue wedge strategies to break up military coalitions.[122](javascript:;) If the consensus about war aims depends on robust political support, adversaries will conduct disinformation and influence campaigns to generate controversy and undermine popular support.[123](javascript:;) If automated systems operate under tightly controlled rules of engagement, adversaries will attempt to manipulate normative frameworks that legitimize the use of force.[124](javascript:;) If AI enables more efficient targeting, the enemy will present more controversial and morally fraught targets to test political resolve.[125](javascript:;) As prediction machines make some aspects of military operations more certain, we argue that the entire military enterprise will become less certain.

### Cyber – Data Attacks

#### Military AI requires immense quantities of data about politics, geography and weather – this information will become contested.

Goldfarb and Lindsay, 2022 – Chair in AI and health care at the Univ of Toronto, Prof of Cybersecurity at the Georgia Tech [Avi, Jon, 2/25/22, <https://doi.org/10.1162/isec_a_00425>, “Prediction and Judgement: Why Artificial Intelligence Increases the Importance of Humans in War” 6/18/22, LND]

We adopt standard international relations distinctions between the international system and domestic institutions.[21](javascript:;) The “strategic environment” in [figure 1](javascript:;) refers to the external problems confronting a military organization. To alter or preserve facts on the ground, through conquest or denial, a military needs information about many things, such as the international balance of power, diplomatic alignments and coalitions, geographical terrain and weather, the enemy's operational capabilities and disposition, and interactions with civil society. These external matters constitute threats, targets, opportunities, resources, and constraints for military operations. A military also needs information about internal matters, such as the capabilities and activities of friendly forces, but these are a means to an end.[22](javascript:;) The strategic environment is ultimately what military data are about, and the structure and dynamics of the environment affect the quality of the data. “Institutions and preferences” in [figure 1](javascript:;) refer to the ways in which a military organization solves its strategic problems. This general category encompasses bureaucratic structures and processes, interservice and coalition politics, civil-military relations, interactions with the defense industry, and other domestic politics. Any of these factors might influence the goals and values of a military organization or the way it interprets a given situation. Organizational institutions embody preferences, whatever their source, which in turn affect the quality of judgment.[23](javascript:;) Institutional structures and processes may produce coordination problems, political controversies, or interpretive difficulties that make it hard for a military organization to figure out what matters and why. Furthermore, as discussed below, we expect that the adoption of AI for some military decision tasks will (endogenously) affect the strategic environment and military institutions over time. As data and judgment become more valuable, strategic competitors will have incentives to improve and contest them. We thus expect conflicts over information to become more salient while organizational coordination will become more complex.

#### Integrating AI into the military increases Data Attacks – AI depends on predictive data, which becomes a target by adversaries.

Goldfarb and Lindsay, 2022 – Chair in AI and health care at the Univ of Toronto, Prof of Cybersecurity at the Georgia Tech [Avi, Jon, 2/25/22, <https://doi.org/10.1162/isec_a_00425>, “Prediction and Judgement: Why Artificial Intelligence Increases the Importance of Humans in War” 6/18/22, LND]

Second, and for the same reasons, adversaries have incentives to complicate both data and judgment. In a highly competitive environment, organizational strengths become attractive targets and potential vulnerabilities. Since predictable adversaries will play to AI strengths, intelligent adversaries will behave unpredictably. If AI creates military power in one area, adversaries will create military challenges in another. Facing an AI-empowered force, the enemy will attempt to change the game by either undermining the quality of predictions or making them irrelevant. Therefore, strategies to contest, manipulate, or disrupt data and judgment become more relevant as military competitors adopt AL The informational and organizational dimensions of war will continue to increase in salience and complexity. Again, this leads us to the conclusion that more military AI will make the human aspects of conflict more important. This increased importance of human personnel challenges the emerging wisdom about AI and war. Many analyses either assume that AI will replace warriors for key military tasks or speculate that war will occur at machine speed, which in turn creates first-mover advantages that incentivize aggression and undermine deterrence.[104](javascript:;) The states that are first to substitute machines for warriors, moreover, are assumed to gain significant military advantages that will shift the balance of power toward early adopters. These outcomes are plausible, but they are based on problematic assumptions about AI substitutability. Conflicts based on AI complementarity may exhibit very different dynamics. We argue that it is more useful to consider the militarized contestation of AI complements (i.e., data and judgment) than to conceive of wars between automated military forces. Conflicts in which data and judgment are perennially at stake may be full of friction, controversy, and unintended consequences, and they may drag on in frustrating ways. In short, we expect the growing salience of data and judgment in war to subtly alter strategic incentives. As a result, AI-enabled conflicts are more likely to be decided by the slow erosion of resolve and institutional capacity than set-piece battles between robotic forces.

### Cyber - Solvency

#### Increasing funding for testing and safety is critical to prevent rushing into deploying AI susceptible to cyber attacks.

Scharre, 2019 - Vice President and Director of Studies at CNAS[Paul, May-June, “Killer Apps: The Real Dangers of an AI Arms Race,” https://omnilogos.com/killer-apps-real-dangers-of-ai-arms-race/6/18/22 MD]

Testing AI systems often takes even more time and money than testing traditional military hardware. Their complexity, which makes them more capable, also creates more opportunities for unexpected glitches. Imagine that a government develops an AI system that can hack into its adversaries' computer networks while avoiding detection. The first government to deploy such a system would gain a huge advantage over its competitors. Worried that an adversary was developing a similar tool, the government might feel compelled to cut testing short and deploy the system early. This dynamic has already played out in other industries, such as self-driving cars. The consequences of accidents caused by [national security](https://go-gale-com.proxy.lib.umich.edu/ps/i.do?p=AONE&u=umuser&id=GALE|A585763278&v=2.1&it=r) AI tools could be far worse. AI wouldn't be the first case of governments relying on a powerful but unsafe technology. That's exactly what happened with computers, which play critical roles in everything from trading stocks to guiding missiles even though they suffer from enormous vulnerabilities. In 2018, investigators at the U.S. Government Accountability Office found that U.S. weapons systems were riddled with cybersecurity loopholes that could be exploited with "relatively simple tools and techniques." Even worse, Defense Department program managers didn't know about the problems and dismissed the GAO'S findings, claiming the tests were not realistic. Computer security vulnerabilities aren't limited to government-run systems. Company after company has suffered major data breaches. Digital security is already too often an afterthought. A world of widespread, unprotected AI systems isn't just a possibility; it's the default setting. SAFETY FIRST Urgent threats require urgent responses. One of the most important ways policymakers can deal with the dangers of AI is to boost funding for AI safety research. Companies are spending billions of dollars finding commercial applications for AI, but the U.S. government can play a valuable role in funding basic AI research, as it has since the field's early days. The AI Next initiative, a program run by the Defense Advanced Research Projects Agency that is set to spend $2 billion over the next five years, is aimed at tackling many of the limitations of narrow AI systems. Expanding on this effort, the White House should increase the funding going to AI safety research as part of its new American AI Initiative, and it should ask Congress for additional money for R & D and safety research. When it comes to applying AI to national security, government agencies will have to reconsider their traditional approaches to testing new systems. Verifying that a system meets its design specifications isn't enough. Testers also need to ensure that it will continue to function properly in the real world when an adversary is trying to defeat it. In some cases, they can use computer simulations to tease out bugs, as manufacturers now do for autonomous cars. On top of that, the Departments of Defense and Homeland Security and the intelligence community should create red teams--groups that act as attackers to test a system's defenses--to ferret out vulnerabilities in AI systems so that developers can fix them before the systems go live.

#### Testing and Verification are essential to AI cyber security – transparency is key to effective Deterrence of cyber threats

Taddeo and Floridi, 2018 - Senior research fellow at the Oxford Internet Institute and Professor of Philosophy and Ethics of Information at the University of Oxford[Mariarosaria, Luciano, Nature.com, “Regulate artificial intelligence to avert cyber arms race,” <https://www.nature.com/articles/d41586-018-04602-6> 6/18/22 MD]

Cyber-attacks are becoming more frequent, sophisticated, and destructive. Each day the US suffers more than 4,000 ransomware attacks that encrypt computer files until the owner pays to release them [1]. This is triple the number in 2015. In May 2017, when WannaCry virus crippled hundreds of IT systems across the UK’s National Health Service, 20,000 appointments were cancelled. A month later, the NotPetya ransomware cost pharmaceutical giant Merck, shipping firm Maersk, and logistics company FedEx around $300 million each in immediate damages. Global damages from cyber-attacks totalled $5 billion in 2017 and may reach $6 trillion a year by 2021. States are partly behind this rise. They use cyber-attacks both offensively and defensively. For example, North Korea has been linked to WannaCry, Russia to NotPetya. As the threats escalate, so do defence strategies. Since 2012, the US has employed ‘active’ cyber defence strategies that enable computer experts to neutralise or distract viruses with decoy targets, and to break back into a hacker’s system to delete data or to destroy it completely. In 2016, the UK announced a £1.9bn investment and a 5-year plan to combat cyber threats. NATO also began drafting principles for such operations, to be agreed by 2019. The US and the UK are leading this initiative, with Denmark, Germany, the Netherlands, Norway, and Spain also involved [2]. Artificial intelligence (AI) is poised to revolutionize this activity. Attacks and responses will become faster, more precise, and more disruptive. Threats will be dealt with in hours, not days or weeks. AI solutions are being rolled out that can verify code and identify bugs and vulnerabilities, making the digital systems they protect more robust. For example, in April 2017 software firm DarkTrace launched Antigena, which uses machine learning to spot abnormal behaviour on an IT network, shut off communications to that part of the system, and issue an alert. Implementations of AI in cyber security 2 will become increasingly pervasive. According to some estimates, the value of AI in cyber security will reach $18.2 billion by 2023 (Figure 1). Experts welcome the potential of AI to improve system verification and robustness [6]. However, this potential is crippled by two risks: the escalation of responses and the lack of control over snowballing effects that AI-led counter-attacks may have. Thus, it is crucial that experts ensure transparency about problems, limitations, and possible shortcomings of unregulated uses of AI for defence, while also working with policy-makers and end users to design adequate testing and oversight mechanisms for this technology. By the end of this decade, many states plan to deploy AI for national cyber defence [4,5]. AI makes deterrence possible [3] because algorithms can identify the source and neutralise it without having to identify the actor behind it, so that attacks can be punished. Currently, countries hesitate to push back because they are unsure who is responsible, given that campaigns may be waged through third-party computers and often use common software. States need regulations that establish legitimate uses of AI for national defence, shape the process leading to the deployment of AI, such as testing and auditing mechanisms, and define the chain of responsibilities [7]. Cyberspace is a domain of warfare and AI is a new defence capability, so regulations are necessary for state use of AI, much like there exist regulations for the other domains of warfare – air, see, land, and space – and the corresponding military capabilities. What is long overdue are criteria for determining proportional responses; clear thresholds, so called ‘red lines’, for distinguishing legal and illegal cyber-attacks; and appropriate sanctionsfor illegal acts[8]. In each case, unilateral approaches will be ineffective. Rather, it is crucial to define an international doctrine for state action in cyberspace. Alarmingly, international efforts to regulate cyber conflicts have stalled. We call on regional forums, like NATO and EU, to revive efforts and prepare the ground for an UN-led initiative.

#### Testing and Reliability are essential to prevent cyber attacks

**National Security Commission on Artificial Intelligence 2022** [March 19, “NCSAI Final Report 2021 – Chapter 7: Establishing Justified Confidence in AI Systems”, <https://assets.foleon.com/eu-west-2/uploads-7e3kk3/48187/nscai_ch7_digital_02-26-21.24159789cf2e.pdf>]

1. Robust and Reliable AI. Current AI systems, such as those used for perception and classification, have different kinds of failure––characterized as rates of false positives and false negatives. They are often brittle when operating at the edges of their performance competence, and it is difficult to anticipate their competence boundaries.3 They are also vulnerable to attack, and they can exhibit unwanted bias in operation. For national security missions, these can be serious problems. U.S. government agencies should: Focus more federal R&D investments on advancing AI security and robustness. These investments should also advance the interpretability and explainability of AI systems, so users can better understand whether the systems are operating as intended. Consult interdisciplinary groups of experts to conduct risk assessments, improve documentation practices, and build overall system architectures to limit the consequences of system failure.4 Such architectures should securely monitor component performance and handle errors when anomalies are detected5 ; contain AI components that are self-protecting (validating input data) and self-checking (validating data passed to the rest of the system); and include aggressive stress testing.

#### TEVV is key to recognize new cyber attacks on autonomous AI

Haugh, 2018 - project leader for the Institute for Defense Analysis [Brian A., September “The Status of Test, Evaluation, Verification, and Validation (TEV&V) of Autonomous Systems,” <https://www.ida.org/-/media/feature/publications/t/th/the-status-of-test-evaluation-verification-and-validation-of-autonomous-systems/p-9292.ashx>, 6/23/22 MD]

7. Exploitable vulnerabilities When systems operate themselves, they can be vulnerable to modes of attack— cyber, electronic, or physical—that would not be as much of a concern for a human-operated system. For example, a cyberattack that compromised the ability of an autonomous unmanned aircraft system (UAS) to recognize other aircraft or physical proximity illusions that repeatedly triggered the UAS’s collision avoidance subroutine might be much more effective than they would be against a human-piloted aircraft. AI based on machine learning has its own set of potential vulnerabilities, both during training of the AI and in operation. TEV&V of autonomous systems will need to be aware of this expanded attack surface. **8**. Emergent behavior US DoD Directive 3000.09 specifically warns against the possibility of “unanticipated emergent behavior resulting from the effects of complex operational environments on autonomous or semi-autonomous systems.” Developing T&E methods to analyze the potential for emergent behavior in order to avoid it will be central to providing assured dependability for autonomous systems.

#### Government regulation over AI networks is crucial to avoid Disinformation disruptions

**Fitzpatrick, 2019 - former the IISS Non-Proliferation and Nuclear Policy Programme** [Mark, 21 May, “Artificial Intelligence and Nuclear Command and Control,”  [https://www.iiss.org/blogs/survival-blog/2019/04/artificial-intelligence-nuclear-strategic-stability 6/18/22](https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/00396338.2019.1614782%206/18/22) MD]

Finally, while disinformation is a long-standing intelligence and strategic problem that pre-dates the cyber age, the integrity of AI systems is especially vulnerable to it. The combination of reliance on AI and its appropriation by malevolent actors could seriously amplify the disruptive effects of disinformation campaigns. Governments could play a key role in establishing regulatory frameworks that shape how new technology interfaces with the current suite of disinformation tools, such as social media. Although time limitations precluded further exploration of an AI system’s vulnerability, hints and suspicions of AI-linked espionage permeated the tabletop exercise. They are likely to persist in the real world as well.

### Cyber – NATO Key

#### NATO is key to AI cyber safety – they must establish norms on testing and verifying cyber AI weapons – the EU is not good enough

Taddeo and Floridi, 2018 - Senior research fellow at the Oxford Internet Institute and Professor of Philosophy and Ethics of Information at the University of Oxford[Mariarosaria, Luciano, Nature.com, “Regulate artificial intelligence to avert cyber arms race,” <https://www.nature.com/articles/d41586-018-04602-6> 6/18/22 MD]

The EU is stepping-up too. In 2017, it reassessed cyber security and defence policies and launched the European Centre of Excellence for Countering Hybrid Threats. EU directives present the most comprehensive regulatory framework for state conduct in cyberspace to date. Yet they do not go far enough. The EU treats cyber defence as a case of cyber security, to be improved passively by making member states’ information systems more resilient. It disregards active uses of cyber defence and does not include AI. This is a missed opportunity. The EU could have begun defining ‘red lines’ and proportionate responses. For example, the EU Directive on Security of Network and Information Systems [10] provides criteria for identifying critical national infrastructures, such as health systems or key energy and water supplies that should be protected. The same criteria could be used to define illegitimate targets of state-sponsored cyber-attacks. Regional forums, like NATO and the EU, must take the following three steps to avoid serious imminent attacks on state infrastructures and to maintain international stability. Three steps Define legal boundaries. The international community needs to agree urgently upon ‘red lines’ distinguishing legitimate and illegitimate targets and definitions of proportionate responses for cyber defence strategies. International consensus at the UN level will ultimately be required. Until then guidelines from regional multilateral bodies, like NATO and the EU, must cover these issues and lead by example. Test strategies with allies. ‘Sparring’ exercises should be organised between friendly countries to test AI-based defence tactics. These tests should be mandatory before any system is deployed. They could be in the form of DARPA’s Grand Challenge or the cyber simulation exercises routinely run by NATO and the EU. Because AI learns by experience, these matches will improve the strategies of the alliance, while finding and healing weaknesses. Fatal vulnerabilities of key systems and crucial infrastructures should be shared with allies; policy frameworks should demand disclosure. Agreements and regulations with similar sharing and disclosure requirements include the EU eIDAS (electronic IDentification, Authentication and trust Services) Regulation and the IPA (Industry Partnership Agreement) within NATO. Monitor and enforce rules. The international community needs to agree on how to audit and oversee AI-based state cyber defence operations. Alert and remedy mechanisms are needed to address 5 mistakes and unintended consequences. A third-party authority with teeth, such as the UN Security Council, should rule on whether red lines, proportionality, responsible deployment or disclosure norms have been breached. As in the case with other international norms, economic or political sanctions should be imposed on states violating rules. NATO and the EU should enforce the norms within their remits.

#### NATO is key to AI cyber defense – the UN has already tried and failed to meet a consensus

Taddeo and Floridi, 2018 - Senior research fellow at the Oxford Internet Institute and Professor of Philosophy and Ethics of Information at the University of Oxford[Mariarosaria, Luciano, Nature.com, “Regulate artificial intelligence to avert cyber arms race,” <https://www.nature.com/articles/d41586-018-04602-6> 6/18/22 MD]

No rules Right now the UN process is in deadlock. In 2004, the United Nations set up a Group of Governmental Experts (UN GGE) to agree on a set of voluntary rules for how states should behave in cyberspace. Their 5th meeting in 2017 ended in a stand-off. The Group could not reach consensus on whether existing international laws on self-defence, humanitarian laws, and state responsibility should apply in cyberspace. The US argued cyber-defence regulations should build on these foundations. Other 3 nations, including Cuba, Russia, and China, firmly disagreed, arguing that this would ‘militarize’ cyberspace,sending the wrong message about peaceful conflict resolution. The Group failed to deliver its report. It is unclear whether it will meet again, or what will happen next. International dialogue and action must resume. NATO could pave the way through its forthcoming guidelines. Although it is currently unclear what their scope will be.

### Cyber – DOD Key

#### DOD is key to respond to Cyber threats due to is functional role– AI makes us vulnerable to bot attacks and disinformation

Nurkin and Konaev, 2022 - senior fellows at the Center for Strategy and Security at the Atlantic Council [Tate and Margarita, 5-25-2022, “Eye to eye in AI: Developing artificial intelligence for national security and defense” Atlantic Council <https://www.atlanticcouncil.org/in-depth-research-reports/report/eye-to-eye-in-ai/> ARD]

Use Case 3: The Limits of AI Adoption in the Information Domain Intensifying competition with China and Russia is increasingly playing out in the information and cyber domains with real, enduring, and disruptive 53 Ibid. implications for US security, as well as the US economy, society, and polity. For cyber and information operations, AI technologies and techniques are central to the future of both offensive and defensive operations, highlighting both the peril and promise of AI in the information domain. Concern is growing about the threat of smart bots, synthetic media such as deepfakes—realistic video or audio productions that depict events or statements that did not take place—and large language models that can create convincing prose and text.56 And, these are just the emerging AI-enabled disinformation weapons that can be conceived of today. While disinformation is a challenge that requires a societal and whole-of government response, DoD will undoubtably play a key role in managing and responding to this threat— due to its prominence in US politics and society, the nature of its functional role, and the impact of its ongoing activities. AI is at the forefront of Pentagon and other US government efforts to detect bots and synthetic media. DARPA’s MediaForensics (MediFor) program is using AI algorithms to “automatically quantify the integrity of an image or video,” for example.57 Still, there is concern about the pace at which this detection happens, given the speed of diffusion of synthetic media via social media. As Lt. General Dennis Crall, the Joint Staff’s chief information officer, observed, “the speed at which machines and AI won some of these information campaigns changes the game for us…digital transformation, predictive analytics, ML, AI, they are changing the game…and if we don’t match that speed, we will make it to the right answer and that the right answer will be completely irrelevant.”58

## Operator Trust Advantage

### Operator Trust – Unsafe AI Undermines Trust

#### Current TEVV cannot establish Trust in AI – the tests were built for human pilots not autonomous drones

Costello, 2020 - Chief of Staff for Test Wing Atlantic [Donald, "Certifying an Autonomous System to Complete Tasks Currently Reserved for Qualified Pilots” - ProQuest <https://www.proquest.com/docview/2495035965/fulltextPDF/D926A79B68D249D7PQ/1?accountid=14667> accessed on 6-21-2022 TM]

While some level of trust is required for certification officials to accept risk, there is not a clearly defined method for how we build trust in military systems. The current V&V techniques are designed for systems with a qualified pilot or operator. To achieve a level of trust, V&V techniques used for certifying autonomy must be sufficient that the certification officials will have trust in the actions of the decision engines controlling the vehicle, and trust that the system will not pose an unnecessary risk to mission completion. For a system to operate autonomously, it will need to sense the environment it is currently operating in. Upon sensing the environment, it will then need to properly classify the input its onboard sensors give it. Once it has interpreted the environment, and built its own SA, it will need to perform appropriate actions based on the unpredictable environments it will find itself in. This is a similar process that a fully qualified pilot is expected to perform once leadership has bestowed trust on their actions. One issue with using sensors to build the SA for a UAV 41 is the vast amount of data the needs to be filtered through. Using automation to process onboard imagery is not a new concept. From December 2013 to January 2015 a CubeSat was flown with multiple onboard sensors. An autonomous onboard processing was used to determine what images would be passed back to the ground station for further processing [64]. While trust is required for certification officials to accept risk, there is not a clearly defined method for how we build that trust in autonomous military systems. The V&V techniques used for certifying autonomy must be sufficient that the certification officials will have trust in the actions of the decision engines controlling military aircraft.

#### Confidence in AI reliability is key for Operator Trust – even if no technology is perfect, it is essential to follow best practices.

**National Security Commission on Artificial Intelligence 2022** [March 19, “NCSAI Final Report 2021 – Chapter 7: Establishing Justified Confidence in AI Systems”, <https://assets.foleon.com/eu-west-2/uploads-7e3kk3/48187/nscai_ch7_digital_02-26-21.24159789cf2e.pdf>]

Artificial intelligence (AI) systems must be developed and fielded with justified confidence.1 If AI systems do not work as designed, or are unpredictable in ways that can have significant negative consequences, then leaders will not adopt them, operators will not use them, Congress will not fund them, and the American people will not support them. “ “If AI systems ... are unpredictable in ways that can have significant negative consequences, then leaders will not adopt them, operators will not use them, Congress will not fund them, and the American people will not support them.” Achieving acceptable AI performance often is linked to the decision to accept some level of risk. No technology works perfectly under all conditions. Risk calculus changes with circumstances. The variables and considerations that inform judgments to rely on AI will vary significantly across military, intelligence, homeland security, and law enforcement missions. In a high-threat environment like combat, in some cases it may be reasonable to employ a system offering some immediate military advantage, while recognizing that it might fail; in other cases, however, a reasonable commander might want the highest assurances of AI reliability before fielding when lives are at risk. As departments and agencies rely more heavily on machines, a central guiding principle across national security scenarios is the continued centrality of human judgment. Those charged with utilizing AI need an informed understanding of risks, opportunities, and tradeoffs. They need awareness of the possibilities and limitations in a system’s expected performance. Ultimately, they need to formulate an educated answer to this question: In the given circumstance, how much confidence in the machine is enough confidence? These issues bear on the full lifecycle of an AI system––from acquisition or system development and the thresholds for justified confidence to deploy a specific AI-intensive system to the performance of the system in the field. While there is no absolute assurance of perfection, there are policies and best practices that support making these decisions responsibly. Agencies are broadly aware of the principal challenges in employing AI systems and the necessity of incorporating best practices in the engineering and management of AI systems.

### Operator Trust – Operator Trust essential to Military

#### Operator trust is crucial for military effectiveness – if users do not trust AI systems, they cannot adopt them.

Costello, 2020 - Chief of Staff for Test Wing Atlantic [Donald, "Certifying an Autonomous System to Complete Tasks Currently Reserved for Qualified Pilots” - ProQuest <https://www.proquest.com/docview/2495035965/fulltextPDF/D926A79B68D249D7PQ/1?accountid=14667> accessed on 6-21-2022 TM]

Trust is vital for certification. When a military commander certifies a pilot as fully qualified they are bestowing their trust on that pilot. Following qualification, the pilot is expected to use his judgment to make decisions based on their experiences. When dealing with autonomy, trust is not inherent and certification is not business as usual. For commanders to trust that an autonomous system will perform as a pilot would will require methods and metrics different than what they are accustomed to. 37 The Defense Science Board identified trust as an integral requirement for the use of autonomous systems by the DoD. “The decision for DoD to deploy autonomous system must be based both on trust that they will perform effectively in their intended use and that such use will not result in high-regret, unintended consequences. Without such trust, autonomous systems will not be adopted except in extreme cases such as mission that cannot otherwise be performed. Further, inappropriate calibration of trust assessments – whether over-trust or under-trust during design, development, or operations will lead to misapplication of these systems. It is therefore important for DoD to focus on critical trust issues and the assurance of appropriate levels of trust [91].” Autonomy is a new concept for the DoD. When employing military systems in the field, someone is ultimately responsible for the actions of that system. This may include the actual military member employing the technology or the individual that certified it for use. Having a system that exhibits non-deterministic behavior inherent in autonomy is a new concept for certification officials. Trust needs to be built prior to the use of autonomy within the DoD.

**Operator trust is essential because AI is a black box – operators cannot examine the machine logic for themselves.**

**Flournoy, Haines et al 2020 - former Under Secretary of Defense for Policy and Director of National Intelligence** [Michèle and Avril, Gabrielle Cheftiz Special Assistant to the Under Secretary of Defense for Policy October, “Building Trust through Testing Adapting DOD’s Test & Evaluation, Validation & Verification (TEVV) Enterprise for Machine Learning Systems, including Deep Learning Systems” https://cset. georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf Acc 6/23/22 JZ]

The black box challenge: unique features of DL traceability and interpretability. It is critical that all ML/DL systems are trustworthy, traceable, and transparent to the greatest extent possible. Deep learning presents challenges for each of these features. Unlike most previous types of computer systems, it may not be possible to trace why a deep learning system made the decision it did in a particular scenario. Not being able to determine what led to an error can obviously create significant challenges for TEVV. It can also undermine user confidence in any solution devised to address the problem identified. Challenges with interpretability in real-time will also hamper human-machine teaming—operators are more likely to trust a system and interact with it effectively if they understand roughly why it is taking certain actions or decisions. Further, the opacity of deep learning systems makes it difficult to identify or trace back certain kinds of adversary attacks, such as some forms of data poisoning. Some forms of attacks are not obvious to human intuition and, therefore, difficult to imagine and test against (i.e., a 3-D printed turtle that fooled Google’s image classifier into classifying it as a rifle.)12 The lack of interpretability, traceability, and explainability of DL systems has the potential to undermine trust and exacerbate challenges associated with developing, deploying, and governing ML/DL at scale.

### Operator Trust – TEVV Solves

**Increase TEVV improves Operator confidence in AI systems – it weapons don’t work, users won’t use them.**

**Schmidt and Work, 2021 – Chairs of The National Security Commission on Artificial Intelligence** [Eric and Robert, “National Security Commission on Artificial Intelligence Final Report Executive Summary” <https://www.nscai.gov/2021-final-report/> Acc 6/7/22 TA]

Establish justified confidence in AI systems. If AI systems routinely do not work as designed or are unpredictable in ways that can have significant negative consequences, then leaders will not adopt them, operators will not use them, Congress will not fund them, and the American people will not support them. To establish justified confidence, the government should focus on ensuring that its AI systems are robust and reliable, including through research and development (R&D) investments in AI security and advancing human-AI teaming through a sustained initiative led by the national research labs. It should also enhance DoD’s testing and evaluation capabilities as AI-enabled systems grow in number, scope, and complexity. Senior-level responsible AI leads should be appointed across the government to improve executive leadership and policy oversight. Present a democratic model of AI use for national security. AI tools are critical for U.S. intelligence, homeland security, and law enforcement agencies. Public trust will hinge on justified assurance that government use of AI will respect privacy, civil liberties, and civil rights. The government must earn that trust and ensure that its use of AI tools is effective, legitimate, and lawful. This imperative calls for developing AI tools to enhance oversight and auditing, increasing public transparency about AI use, and building AI systems that advance the goals of privacy preservation and fairness. It also requires ensuring that those impacted by government actions involving AI can seek redress and have due process. The government should strengthen oversight and governance mechanisms and establish a task force to assess evolving concerns about AI and privacy, civil liberties, and civil rights.

#### Improved Verification and Validation would solve operators’ safety concerns about AI, building Trust.

Tarraf, Shelton, Parker 2019 - Senior Information Scientist, Senior Engineer, and physical Scientist at the RAND Corporation [ Danielle C., William, Edward, et al RAND Cooperation, “The Department of Defense Posture for Artificial Intelligence”, [file:///Users/MiraAgarwal/Downloads/RAND\_RR4229%20(2).pdf](file:///C:\Users\MiraAgarwal\Downloads\RAND_RR4229%20(2).pdf), Acc 6/18/22, M.A.]

As we noted in Chapter Three, our interviewees agreed that V&V of AI systems is a critical topic that DoD will need to address, particularly when it seeks to deploy and use safety-critical AI systems. 18 Interviewees both inside and outside DoD emphasized the importance of improving the understanding of human-machine interaction. 52 The Department of Defense Posture for Artificial Intelligence However, like our academic and industry interviewees (see section “Industry: Advancement and Adoption” in Appendix C), the interviewees lacked current V&V solutions and could not predict which technical avenues toward V&V would eventually prove to be effective for AI systems. Concerns about the “valley of death” and the failures in transitioning research to prototypes and products, both broadly and specifically for AI, also came up in our interviews.19 Finally, although our interviewees generally agreed that AI will have a bigger role to play, they also emphasized the need for training and experimentation to build trust and enhance adoption of these technologies, while noting that generational factors might help down the line.20

#### Safety and reliability policies increase a soldier’s Trust in their weapons, which is essential to their effective use. It doesn’t matter how good your AI weapons are if the military won’t use them.

Gilli, 2020 - Senior Researcher at the NATO Defense College [Andrea, NDC Research Paper No.15 – December ““NATO-Mation”: Strategies for Leading in the Age of Artificial Intelligence” https://www.ndc.nato.int/news/news.php?icode=1514 Acc 4/21/22 TA]

This is important because AI is data-intensive, unpredictable and brittle.98 While a system may work well in the context in which it was trained, it may break in an unfamiliar setting. The reliance on BD in the current second wave of ML also creates fallibilities,99 given that the algorithms can scale up harm if the data over- or under-represents certain groups.100 For neural networks in particular, it may be impossible to explain or interpret results. Some refer to this as the “black-box” problem, meaning that the outcomes of these complex AI systems are opaque to humans that either want to reproduce the good, or prevent the bad from recurring. While traditional software can be debugged to solve a performance issue, the lack of linear causality between a programmer’s inputs and the AI system’s outputs means that it is difficult to track bias and reliability. Creating organizational processes to minimize these concerns across the AI lifecycle is critical to responsible use of the technology. Considering AI ethics also means developing trust in systems. Organizations developing AI applications should think deeply about user expectations related to transparency and disclosure. Ultimately, these norms will change the distinction between human, AI-assisted and AI interactions.101 “Calibrating” trust is especially important for AI-assisted decision making,102 a concept whose value will only increase given the emphasis on human-machine teaming in Allied militaries.103 In many cases, the unethical or unacceptable outcomes of AI systems pertain not to moral dilemmas, but to the reliability and robustness of the systems at hand. As such, building trust is fundamentally tied to building safe and secure systems.104 The problem of bias in AI illustrates the overlap between AI safety and AI ethics: bias is morally problematic, because it can unfairly harm or systematically discriminate against specific groups105 – and it can also be seen as a failure mode that reduces the reliability of a given algorithm in a given context.106 For NATO, this relates to the safety of enterprise tools and weapons systems alike. In deployments, these safety measures would be critical to ensure that AI-enabled weapons systems are used in a manner consistent with the principles of international humanitarian law. As explained in greater detail below, this may also feed into the eventual standardization process.

#### TEVV is key to operator Trust in AI – it ensures Explainability – it allows humans to understand AI decisions

Haugh, 2018 - project leader for the Institute for Defense Analysis [Brian A., September “The Status of Test, Evaluation, Verification, and Validation (TEV&V) of Autonomous Systems,” <https://www.ida.org/-/media/feature/publications/t/th/the-status-of-test-evaluation-verification-and-validation-of-autonomous-systems/p-9292.ashx>, 6/23/22 MD]

B. Cognitive Instrumentation and Explainable AI 1. Summary The key distinguishing feature of autonomous systems is that they make decisions, when interpreting their environments and selecting courses of action. Test and evaluation of autonomy will depend critically on the ability to assess the quality of this decision making capability. In general, observed high-level system performance in limited test scenarios will not be sufficient to validate and verify autonomous decision-making in the ways necessary for successful development and deployment, especially for systems designed to team with humans. Instead, novel instrumentation approaches will be required to diagnose and characterize: • Adequacy of architectures and algorithms (including machine learning). • Adequacy and appropriateness of training data. • Effectiveness of operational concepts for HMT. 2. Why Explanations Are Needed Explanations of system behavior support at least four distinct goals for military systems: 1. Diagnosis: Knowing why the system is exhi biting undesired behavior is the first step toward fixing it. 2. Prediction: Being able to forecast how the system will behave in given circumstances is essential to effective employment of the system. 3. Bounding: Understanding the limits of dependable performance allows formulation of tactics/techniques/procedures for how the system is to be used and identification of where run-time monitoring of system state may be the only way to avoid undesirable behaviors. 3-3 4. Trust: If humans are teaming with autonomous machines, the overall performance of the human-machine system may depend on how well the humans feel they understand the reasons behind the machine’s behaviors. For traditional software employing procedural algorithms, the logic of the algorithm traditionally serves as the explanation. Describing the flow of control (e.g., using pseudocode or flowcharts) at various levels of description could “explain” the behavior of the system. As systems have grown in complexity, the ability of humans to understand the logic of the systems in a way that counts as an “explanation” has eroded. Machine learning breaks this paradigm completely by replacing comprehensible procedural logic with a trained ability to generate outputs from inputs in a manner that looks much more like an intuition or hunch than like reasoning. The relationship between input and output is fundamentally opaque; there is no procedural logic to be traced. Producing surrogate procedural descriptions that summarize how the system is reaching its conclusions in ways that can function as explanations for purposes of diagnosis, prediction, bounding, and trust requires additional work and access to the internal states (and possibly the training algorithm) of the machine-learning module. (Note that different explanatory frameworks will generally be needed for each of these goals—the explanation appropriate for diagnosis may be very different from the explanation required for bounding or trust.) The US Defense Advanced Research Projects Agency (DARPA) has recently launched a funded research program in “Explainable Artificial Intelligence” (XAI). The four goals mentioned above (diagnosis, prediction, bounding, and trust) are all represented in the Explainable AI program. “Explainability” in the XAI program involves multiple measures of effectiveness, depending on which of the goals is being pursued. These measures can include human satisfaction with the explanation, how helpful the explanation is in choosing how to interact with the autonomous system, how well the explanation supports diagnosis of unexpected behaviors, or how well the explanation predicts future system behavior. 3. Cognitive Instrumentation The phrase cognitive instrumentation refers to any measurement of internal system states of the software modules that provide the autonomous capabilities of a system. There is an obvious analogy with physical measurement of internal system state (e.g., temperature, pressure, voltage, torque) during development, test, and evaluation of hardware systems. As with those physical analogues, the measurements may be only for TEV&V purposes, or they may be incorporated into the design of the system as run-time monitors on behaviors that cannot be adequately assured through other means. One area of explainability research involves designing autonomous systems to selfreport their reasons for their behaviors. This is a promising approach, but it does not avoid 3-4 the need for instrumenting the internal states of the system. At a minimum, verification and validation of the self-reporting system will involve the same diagnosis, prediction, and trust issues as the mission behavior of the system, with an associated need to look inside the black box and make sense of what is happening there. At some point, all system development requires knowing the truth about what is happening inside the system. In the special case of HMT, where the machines have significant autonomous capabilities, cognitive instrumentation will include measurements of both the machines and the humans, as well as the interface between them. Measurement and characterization of trust by humans is already a rich area of academic research. Measurement and characterization of machine understanding of human intent is also receiving increasing attention. There are also active research programs exploring the design of HMT protocols. Some of these focus on the optimal allocation of work between humans and machines in various contexts. Others are concerned with optimal communications protocols—under what circumstances should humans or machines volunteer information (and which information), make suggestions, or ask for guidance? Without cognitive instrumentation, those optimization programs could easily devolve into guesswork and trial-and-error. Cognitive instrumentation is a necessary condition for Explainable AI; any valid explanation of how the AI is thinking or why it is behaving a certain way must be based on accurate measurements of its internal states. However, the measurements themselves are not the explanation. A complete state description will generally be so complex that it is not understandable, whereas too small a subset of measurements will generally not be sufficient to support explanation. Additional effort will be required to identify a minimal sufficient set of measurements that can be used as inputs to explanatory models that can be understood in human terms for diagnosis, prediction, bounding, and developing appropriate operator trust. There are many tricky steps between “the weights in layer 17 of the neural network are as follows” to “the system seems to be recognizing cats by their ears and coloration.” As noted above, different explanatory models will be needed for the different goals—an explanatory model supporting useful diagnosis for developers may be much more complex (and unwieldy) than a model that produces explanations of mission behavior for operators or commanders.

#### TEVV is crucial for Operator Trust in autonomous systems

Haugh, 2018 - project leader for the Institute for Defense Analysis [Brian A., September “The Status of Test, Evaluation, Verification, and Validation (TEV&V) of Autonomous Systems,” <https://www.ida.org/-/media/feature/publications/t/th/the-status-of-test-evaluation-verification-and-validation-of-autonomous-systems/p-9292.ashx>, 6/23/22 MD]

4. CONOPS and training as design features To date, the paradigm for designing systems has been to make a reasonable guess about how the operator(s) will use that system and what the user interface should look like, and then work out the details of the Concept of Operations (CONOPS), tactics for effective employment, and training of future operators long after the basic design has been decided. For autonomous systems, where the system operates itself to some extent and interacts autonomously with humans, the details of CONOPS and tactics (and corresponding training) are part of the system design, at least on the machine side (and probably on the human side as well), and will have to be identified, verified, and validated much earlier in the development process. This will pose organizational and personnel challenges to T&E, as well as methodological challenges. 5. Human trust In human-machine teaming (HMT) contexts, how the humans behave (and thus how well the team performs) depends in part on the humans’ psychological attitudes toward the autonomous systems. “Trust” is the term generally used to describe those attitudes, though in practice those attitudes are generally more nuanced and multi-dimensional than simply asking “how much do I trust it?” In order to design, debug, and assure performance, TEV&V will need to be able to measure the various dimensions of trust, to support understanding of how trust affects team performance.

### AI Apocalypse

#### An AI Arms race is an existential threat due to a lack of control over runaway technology

Effoduh, 2021 – PhD candidate at Osgoode Hall Law School [Jake Okechukwu June 23 World Economic Forum, “Weapons powered by artificial intelligence pose a frontier risk and need to be regulated” <https://www.weforum.org/agenda/2021/06/the-accelerating-development-of-weapons-powered-by-artificial-risk-is-a-risk-to-humanity/#:~:text=are%20frontier%20risks%3F-,Frontier%20risks%20are%20low%2Dlikelihood%2C%20high%2Dimpact%20threats%20that,world%20as%20we%20know%20it>. BK]

What are frontier risks? Frontier risks are low-likelihood, high-impact threats that could arise as humans explore new realms, whether technological, ecological or territorial. The outcomes cannot be predicted easily (especially by a non-niche group without expert insights), but they have the capacity to alter the world as we know it. These frontier risks are best illustrated by three metaphors: “dragon kings” (rare, high-impact events that are somewhat predictable), “grey rhinos” (highly probable, high-impact events that occur only after a lengthy series of warnings) and “black swans’ (high-impact events that are virtually impossible to predict). The frontier risks that could emerge from the full militarization of autonomous weapons include catastrophic fallout from army raids and a human existential crisis in the age of machine sentience. Why are military systems becoming autonomous? Drones are already used by the military today, but they have vulnerabilities. Most rely on human command through a set of communication channels that can be infiltrated by adversaries. In addition, considerable personnel is required to operate the multiple drones dispatched on the field. By cutting out the need for human control, thousands of L.A.W.S can be dispatched at once, all able to make tactical and timely decisions on their own. This intensifies the speed and expansion of combatants’ reach and reduces the need for proximity to targets. L.A.W.S also exceed the limitations of the human mind and body (for example, they do not risk suffering post-traumatic stress disorders or make erroneous judgment in service of self-preservation). There are, however, some unanswered questions and potential unforeseen consequences. The race by world powers to become leaders in this field have led to what four-star U.S. Army General David Petraeus called “the early stages of a tech Cold War.” Similarly, the entrepreneur Elon Musk warned in 2017 that the scramble for military applications of artificial intelligence could become so intensified that it causes WWIII. Because of the popularity of robot doomsday speculation in pop culture over the decades, the fear of L.A.W.S. is already ingrained in the public imagination. But due to political sovereignty some states do not seem as moved by the fears of the potential risks from L.A.W.S. There are also insufficient global regulations on the use of militarized AI. While the United Nations and the European Parliament have fielded proposals to ban the use of lethal autonomous robots in war, there are still unanswered questions about how the weapons could be regulated and whether regulation would conflict with states’ sovereignty. Will the machines save us or kill us? Machines make mistakes, albeit less often than human soldiers do. But while humans are held responsible for their actions, machines cannot suffer legal consequences. With the complexities of international humanitarian law (IHL) regarding warfare, there could be the risk of setting loose machines that have difficulty differentiating among active combatants, injured soldiers and those surrendering, flouting the 1949 Geneva Conventions. One could argue that the idea of fully autonomous weapons already contravenes IHL. For instance, the Martens Clause prioritizes human dignity and the capacity of soldiers to show compassion toward their fellow humans. But robots powered by lines of code do not have the capacity to make such decisions yet and could end up flouting the rules. Another concern about violations of IHL is underscored by the potential for L.A.W.S. to act of their own accord. An idea that used to seem like fodder for the most imaginative of dystopian fiction is gaining more traction in recent years; Musk and theoretical physicist Professor Stephen Hawking both warned about sentient AIs establishing a takeover and threatening human civilization. The UN Convention on Certain Conventional Weapons and Article 14 of the New Delhi Rules prohibits the use of weapons that could escape from the control of those who employ them, thus endangering the civilian population. While that scenario might still be considered unlikely, this is a kind of frontier risk that could have catastrophic impact on the world.

#### Global cooperation on AI safety is key to avoiding the most dangerous AI technology

Green et al., 2019 - Director of technology ethics at the Markkula Center for Applied Ethics[Brian Patrick et al., Big Data and Cognitive Computing, “Global Solutions vs. Local Solutions for the AI Safety Problem,” https://proxy.lib.umich.edu/login?url=https://www.proquest.com/scholarly-journals/global-solutions-vs-local-ai-safety-problem/docview/2546937568/se-2, 6/18/22 MD]

4.4.2. Ways to Affect a Race to Create the First AI An AI creation race is generally regarded as bad because it encourages the creation of the least-safe AIs first. A war between AIs may also become possible if several AIs are created simultaneously [64,65]. There are many ideas on how to affect an AI race in order to make it safer, that is, to lower the probability of creating dangerous AI. As a race with many participants is a very complex game, there are not obvious ways to predict how it will react to seemingly good interventions, for example, openness. Bostrom has shown that if no one knows the capabilities of others and their own capabilities, it will slow down the race, so openness about capabilities may be dangerous [66]. Actions that may affect an AI race and make it safer may include: Changing the number of participants. Increasing or decreasing information exchange and level of openness. Reducing the level of enmity between organizations and countries, and preventing conventional arms races and military buildups. Increasing the level of cooperation, coordination, and acceptance of the idea of AI safety among AI researchers. Changing the total amount of funding available. Promoting intrinsic motivations for safety. Seth Baum discussed the weakness of monetary incentives for beneficial AI designs, and cautions: “One recurrent finding is that monetary incentives can reduce intrinsic motivation” [67]; when the money is gone, people lose motivation. Baum also noted that the mere fact that a law existed promoted obedience in some situations and that social encouragement can increase intrinsic motivation. Changing social attitudes toward the problem and increasing awareness of the idea of AI safety. Trying to affect the speed of the AI race, either slowing it down or accelerating it in just one place by concentrating research. It is interesting to note that acceleration could be done locally, but slowing it would require global cooperation, and so is less probable. Affecting the idea of the AI race as it is understood by the participants [67]: if everybody thinks that the winner takes everything, the race is more dangerous. A similar framing solution has been suggested in the field of bioweapons, that is, to stop claiming bioweapon creation is easy, as it might become attractive to potential bioterrorists. In fact, bioweapons are not as easy to develop and deploy as is shown in movies, and would probably kill the terrorists first [68].

## Solvency

### Solvency - TEVV

#### TEVV is crucial for AI development – it provides Assurance which is necessary to apply AI across many domains

Batarsheh et al. 2021 - Professor of Electrical and Computer Engineering, Virginia Polytechnic Institute [Feras, Laura Freeman Director, Intelligent Systems Division – National Security Institute Research Associate Professor, Chih Hao Huang, Data Analyst at Turing Research. April 26 “A survey on artificial intelligence assurance” https://link-springer-com.proxy.lib.umich.edu/article/10.1186/s40537-021-00445-7#Sec8]

Artificial Intelligence (AI) algorithms are increasingly providing decision making and operational support across multiple domains. AI includes a wide (and growing) library of algorithms that could be applied for different problems. One important notion for the adoption of AI algorithms into operational decision processes is the concept of assurance. The literature on assurance, unfortunately, conceals its outcomes within a tangled landscape of conflicting approaches, driven by contradicting motivations, assumptions, and intuitions. Accordingly, albeit a rising and novel area, this manuscript provides a systematic review of research works that are relevant to AI assurance, between years 1985 and 2021, and aims to provide a structured alternative to the landscape. A new AI assurance definition is adopted and presented, and assurance methods are contrasted and tabulated. Additionally, a ten-metric scoring system is developed and introduced to evaluate and compare existing methods. Lastly, in this manuscript, we provide foundational insights, discussions, future directions, a roadmap, and applicable recommendations for the development and deployment of AI assurance. Introduction and survey structure The recent rise of big data gave birth to a new promise for AI based in statistical learning, and at this time, contrary to previous AI winters, it seems that statistical learning enabled AI has survived the hype, in that it has been able to surpass human-level performance in certain domains. Similar to any other engineering deployment, building AI systems requires evaluation, which may be called assurance, validation, verification or another name. We address this terminology debate in the next section. Defining the scope of AI assurance is worth studying, AI is currently deployed at multiple domains, it is forecasting revenue, guiding robots in the battlefield, driving cars, recommending policies to government officials, predicting pregnancies, and classifying customers. AI has multiple subareas such as machine learning, computer vision, knowledge-based systems, and many more—therefore, we pose the question: is it possible to provide a generic assurance solution across all subareas and domains? This review sheds light on existing works in AI assurance, provides a comprehensive overview of the state-of-the-science, and discusses patterns in AI assurance publishing. This section sets that stage for the manuscript by presenting the motivation, clear definitions and distinctions, as well as the inclusion/exclusion criteria of reviewed articles. Relevant terminology and definitions All AI systems require assurance; it is important to distinguish between different terms that might have been used interchangeably in literature. We acknowledge the following relevant terms: (1) validation, (2) verification, (3) testing, and (4) assurance. This paper is concerned with all of the mentioned terms. The following definitions are adopted in our manuscript, for the purposes of clarity and to avoid ambiguity in upcoming theoretical discussions:

#### Assurance and evaluation are essential to military AI – it directs research toward successful programs.

Batarsheh et al. 2021 - Professor of Electrical and Computer Engineering, Virginia Polytechnic Institute [Feras, Laura Freeman Director, Intelligent Systems Division – National Security Institute Research Associate Professor, Chih Hao Huang, Data Analyst at Turing Research. April 26 “A survey on artificial intelligence assurance” https://link-springer-com.proxy.lib.umich.edu/article/10.1186/s40537-021-00445-7#Sec8]

Providing a scoring system to evaluate existing methods provides support to scholars in evaluating the field, avoiding future mistakes, and creating a system where AI scientific methods are measured and evaluated by others, a practice that is becoming increasingly rare in scientific arenas. More importantly, practitioners –in most cases– find it difficult to identify the best method for assurance relevant to their domain and subarea. We anticipate that this comprehensive review will help in that regard as well. As part of AI assurance, ethical outcomes should be evaluated, while ethical considerations might differ from one context to another, it is evident that requiring outcomes to be ethical, fair, secure, and safe necessitates the involvement of humans, and in most cases, experts from other domains. That notion qualifies AI assurance as a multidisciplinary area of investigation. In some AI subareas, there are known issues to be tackled by AI assurance, such as deep learning’s sensitivity to adversarial attacks, as well as overfitting and underfitting issues in machine learning. Based on that and on the papers reviewed in this survey, it is evident that AI assurance is a necessary pursuit, but a difficult and multi-faceted area to address. However, previous experiences, successes, and failures can point us to what would work well and what is worth pursuing. Accordingly, we suggest performing and developing AI assurance by (1) domain, by (2) AI sub area, and by (3) AI goal; as a theoretical roadmap, similar to what is shown in Fig. 3. In some cases, such as in unsupervised learning techniques, it is difficult to know what to validate or assure [86]. In such cases, the outcome is not predefined (contrary to supervised learning). Genetic algorithms and reinforcement learning have the same issue, and so in such cases, feature selection, data bias, and other data-relevant validation measures, as well as hypothesis generation and testing become more important. Additionally, different domains require different tradeoffs; trustworthiness for instance is more important when it comes to using AI in healthcare versus when its being used for revenue estimates at a private sector firm; also, AI safety is more critical in defense systems than in systems built for education or energy application.

#### TEVV is key to effective autonomous systems because it will provide an internal model of how AI perceives its environment

Haugh, 2018 - project leader for the Institute for Defense Analysis [Brian A., September “The Status of Test, Evaluation, Verification, and Validation (TEV&V) of Autonomous Systems,” <https://www.ida.org/-/media/feature/publications/t/th/the-status-of-test-evaluation-verification-and-validation-of-autonomous-systems/p-9292.ashx>, 6/23/22 MD]

1. Instrumenting machine thinking To diagnose the causes of incorrect behavior or inadequate performance, it will be necessary determine whether the problem lies in the Perception, Reasoning, or (course of action) Selection functions of the autonomous system—even after it has been established that the problem is not in the sensor hardware or signal processing. It will also be necessary to distinguish coding errors from inadequate algorithms or bad training data. Without the ability to instrument and monitor internal states of the autonomy, diagnosing problems will be slow at best and impossible at worst. 2. Linking system performance to autonomous behaviors In complex collaborative activities, it can be very difficult to determine what enables (or hinders) success. For example, on a soccer or basketball team it can be very difficult to pinpoint which players (and which behaviors) are leading to wins and losses. To design and improve autonomous systems, it will be necessary to understand how the system’s various autonomous capabilities interact to enable (or hinder) mission execution. The requirements specification for autonomous behavior is often a problem here due to incomplete specifications based on an analogy with human behavior. 2. Linking system performance to autonomous behaviors 3. Comparing AI models to reality Autonomous systems represent reality through stylized internal models. Perception provides inputs for these models; Reasoning allows them to be expanded and corrected. The ability of an autonomous system to do its mission will depend on the degree to which the internal modeling of reality supports accurate Perception, valid Reasoning, and effective Selection. This will not generally be a function of how detailed the models are (“high resolution”) or even of how closely the models mirror reality (“high fidelity”)—it will be a function of whether the right kind of information is incorporated into the model and that the resolution and fidelity be enough to support the mission needs. TEV&V will necessarily include prototyping and experimentation to determine 2-2 what kind of internal model, using what kind of representation, is needed to achieve both performance and dependability.

#### Assurance is essential for AI systems – it is necessary for reliability and user trust.

Batarsheh et al. 2021 - Professor of Electrical and Computer Engineering, Virginia Polytechnic Institute [Feras, Laura Freeman Director, Intelligent Systems Division – National Security Institute Research Associate Professor, Chih Hao Huang, Data Analyst at Turing Research. April 26 “A survey on artificial intelligence assurance” https://link-springer-com.proxy.lib.umich.edu/article/10.1186/s40537-021-00445-7#Sec8]

The need for AI assurance The emergence of complex, opaque, and invisible algorithms that learn from data motivated a variety of investigations, including: algorithm awareness, clarity, variance, and bias [94]. Algorithmic bias for instance, whether it occurs in an unintentional or intentional manner, is found to severely limit the performance of an AI model. Given AI systems provide recommendations based on data, users’ faith in that the recommended outcomes are trustworthy, fair, and not biased is another critical challenge for AI assurance.

**Safety and reliability are the best balance of speed and caution**

**Horowitz and Kahn, 2022 - Senior and Research Fellows for Defense Technology and Innovation at the Council on Foreign Relations** [Michael and Lauren, with Laura Samotin, May/June Foreign Affairs “A Force for the Future A High-Reward, Low-Risk Approach to AI Military Innovation” [https://www.foreignaffairs.com/articles/united-states/2022-04-19/force-future Acc 5/28/22](https://www.foreignaffairs.com/articles/united-states/2022-04-19/force-future%20Acc%205/28/22) TA]

FAST, NOT LOOSE The United States, then, faces dueling risks from AI. If it moves too slowly, Washington could be overtaken by its competitors, jeopardizing national security. But if it moves too fast, it may compromise on safety and build AI systems that breed deadly accidents. Although the former is a larger risk than the latter, it is critical that the United States take safety concerns seriously. To be effective, AI must be safe and reliable. So how can Washington find a sort of Goldilocks zone for innovation? It can start by thinking of technological development in terms of three phases: invention, incubation, and implementation. Different speeds are appropriate for each one. There is little harm from moving quickly in the first two phases, and the U.S. military should swiftly develop and experiment with new technologies and operational concepts. But it will need to thoroughly address safety and reliability concerns during implementation.

### Solvency – Accidents

#### Standards for monitoring and oversight help reduce the risk of AI accidents.

**Artificial Intelligence/Machine Learning Risk & Security Working Group, 2020** [Wharton, ‘Artificial Intelligence Risk & Governance’, <https://ai.wharton.upenn.edu/artificial-intelligence-risk-governance/>]

Key Takeaways AIRS believes there are significant potential benefits of AI and that its adoption within financial services presents opportunities to improve both business and societal outcomes when risks are managed responsibly. This paper explores the potential risks of AI and provides a standardized practical categorization of these risks: Data Related Risks, AI/ML Attacks, Testing and Trust, and Compliance AI governance frameworks could help organizations learn, govern, monitor, and mature AI adoption. Four core components of AI governance are: definitions, inventory, policy/standards, and a governance framework, including controls. AI, in certain use cases, could lead to privacy issues, and/or potentially discriminatory or unfair outcomes, if not implemented with appropriate care. We explore, in detail, the subject of interpretability and discrimination in using AI for certain use cases. While there is no one-size-fits-all approach, practices institutions might consider adopting to mitigate AI risk include oversight and monitoring, enhancing explainability and interpretability, as well as exploring the use of evolving risk-mitigating techniques like differential privacy, and watermarking, among others.

### Solvency - Cooperation

**NATO cooperation is key to developing standards for safety and reliability of AI weapons**

**Gilli, 2020 - Senior Researcher at the NATO Defense College** [Andrea, NDC Research Paper No.15 – December ““NATO-Mation”: Strategies for Leading in the Age of Artificial Intelligence” https://www.ndc.nato.int/news/news.php?icode=1514 Acc 4/21/22 TA]

Approach to standards. AI does raise new challenges, but this does not mean that all new challenges require new solutions. Some standards, for instance, can be imported horizontally from nearby fields, like that of cyber security. This issue highlights the importance of the approach to standards, which NATO Allies may want to discuss: should new standards be pursued when it comes to AI, or should countries agree to use existing ones? Should they opt for stable standards, or try to update them regularly? These are important questions. For instance, the U.S. National Institute for Standards and Technology provides a list of features that the US should pursue for standardization – e.g., innovation-oriented, applicable across sectors and applications, human-centred.324 The Atlantic Alliance would benefit from a similar discussion as well as, eventually, trying to coordinate each Ally’s preferences, position and approaches towards standard multinational standard-setting forums, like Standard Development Organizations. Standards and effectiveness. One of the main challenges of the interaction between algorithms and BD is their brittleness.325 Algorithms can be (easily) fooled. In the civilian economy, this may lead to controversial or even unacceptable outcomes, such as discrimination or bias.326 For instance, HR algorithms may discard some candidates because historical underrepresentation of their sex or ethnicity disadvantages them in the data on which the recommender algorithm is trained. In the military domain, this may lead to errors, like mistaking a military target for a civilian facility or the other way around, and overall combat ineffectiveness, or even defeat. The use of generative adversarial networks which intentionally exploit these vulnerabilities to undermine the opponent’s capabilities makes matters worse.327 To address these problems, AI and related technologies must comply with a set of principles that standards can translate and consolidate, including safety and reliability as well as accepted degrees of accuracy. Non-technical standards can also play an important role in defining procedures, processes and approaches in developing and fielding those systems, including about issues related to human control.328 However, standards should be also complemented by tools to enable the development of effective, reliable, robust and trustworthy AI technologies, in areas such as accountability and auditing, annotations and documentation, benchmarks, metrics, AI testbeds, methodologies and datasets in standardized formats. Standards and R&D. Standardization is, however, not just about agreeing on technical and non-technical parameters, processes and procedures: it is first and foremost about conducting R&D to identify, understand and address technical and non-technical issues, problems and risks. On the one hand, this is a government effort at national level. On the other hand, the private sector, with the bulk of R&D funds for AI, has a role to play, and thus a solid partnership between governments and industry is important. From a NATO perspective, there is a strong rationale for trying to coordinate some of these national-level R&D investments into standards, although the issue is not simple, given that some are going to accrue more economic value than others.329

#### Cooperation with allies is essential to lead responsible use of AI due to our shared democratic values.

Schmidt 2021 – Chairman of the National Security Commission on Artificial Intelligence [Eric, March 12, “House Armed Services Subcommittee on Cyber, Innovative Technologies and Information Systems and House Oversight and Reform Subcommittee on National Security Hold Joint Hearing on AI and the National Security Commission”, https://congressional-proquest-com.proxy.lib.umich.edu/congressional/result/congressional/congdocumentview?accountid=14667&groupid=95663&parmId=180DE199B78&rsId=180DE197DA6]

We need to build coalitions with like-minded nations, the technology democracies, the techno democracies in my own ver--verbiage to advance the development and use of AI in emerging technologies that support our values, which is critical. We spent a lot of time on our report talking about values. And the second one, consisting with the values is responsible use. In the face of dig--digital authoritarianism, we need, we the U.S. need to present a democratic model of responsible use of AI for national security. You can imagine the opponents and how they might use or misuse these things. The trust of our nation, the trust of our citizens will hinge on justified assurance that the government's use of AI will respect privacy, cybil--civil liberties and civil rights. We have a set of recommendation along those lines. I really thank you all for giving just this opportunity. It has been a true privilege for me to be part of this, and to help lead it. Thank you very much.

#### International cooperation is key to AI safety – empirical examples prove.

Scharre, 2019 - Vice President and Director of Studies at CNAS[Paul, May-June, “Killer Apps: The Real Dangers of an AI Arms Race,” https://omnilogos.com/killer-apps-real-dangers-of-ai-arms-race/6/18/22 MD]

Finally, the United States should look for ways to work with other countries, even hostile ones, to ensure AI safety. International cooperation on new technologies has a mixed record, but countries have sometimes succeeded in working together to avoid mutual harm. During the Cold War, the United States and the [Soviet Union](https://go-gale-com.proxy.lib.umich.edu/ps/i.do?p=AONE&u=umuser&id=GALE|A585763278&v=2.1&it=r) worked together to limit certain types of delivery systems for nuclear warheads that both sides agreed were particularly destabilizing. The United States also encouraged other countries to adopt safety measures to prevent the unauthorized use of [nuclear weapons](https://go-gale-com.proxy.lib.umich.edu/ps/i.do?p=AONE&u=umuser&id=GALE|A585763278&v=2.1&it=r). Today, the United States should work with both allies and adversaries to boost international funding on AI safety. It should also begin discussions with China and [Russia](https://go-gale-com.proxy.lib.umich.edu/ps/i.do?p=AONE&u=umuser&id=GALE|A585763278&v=2.1&it=r) over whether some applications of AI pose unacceptable risks of escalation or loss of control and what countries can do jointly to improve safety. The biggest danger for the United States in an AI race is not losing but creating a world in which no one wins.

### Solvency – Global Norms

#### Investing in Effective TEVV solidifies US leadership on AI – it sets a norm for other nations and private companies

**Flournoy, Haines et al 2020 - former Under Secretary of Defense for Policy and Director of National Intelligence** [Michèle and Avril, Gabrielle Cheftiz Special Assistant to the Under Secretary of Defense for Policy October, “Building Trust through Testing Adapting DOD’s Test & Evaluation, Validation & Verification (TEVV) Enterprise for Machine Learning Systems, including Deep Learning Systems” https://cset. georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf Acc 6/23/22 JZ]

Third, with an effective TEVV system, the United States can reduce barriers to innovation and facilitate U.S. leadership in ML/DL technologies. As most of the innovation in ML/DL will come from the private sector, unless the U.S. government is able to effectively draw on private sector work in this arena, it will not be able to leverage the best cutting-edge technology. Research on new TEVV methods and organizational reforms to adapt the current system is simply not keeping pace with private sector development. Without urgent reforms and prioritized investment in new research and infrastructure, the Defense Department will lose its chance to shape industry’s approach to ML/DL development in a manner consistent with DOD standards for safety, reliability, and accountability. It will lose the opportunity to take advantage of new private sector developments, while allowing other nations without such standards to adopt the latest innovations. It is critical that the U.S. government not only shape its own U.S. industry standards but also promote compatible global standards and norms.

#### DOD TEVV standards will establish a global model – this promotes interoperability and reduces our vulnerabilities

**Flournoy, Haines et al 2020 - former Under Secretary of Defense for Policy and Director of National Intelligence** [Michèle and Avril, Gabrielle Cheftiz Special Assistant to the Under Secretary of Defense for Policy October, “Building Trust through Testing Adapting DOD’s Test & Evaluation, Validation & Verification (TEVV) Enterprise for Machine Learning Systems, including Deep Learning Systems” https://cset. georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf Acc 6/23/22 JZ]

7. Develop industry/U.S. government TEVV standards and promote them internationally. DoD, working with the National Institute of Standards and Technology (NIST) and industry, and building on DoD requirements and processes developed for industry, should develop standards for ML/DL testing for the private sector that can be publicly promoted and help inform private sector development of ML/DL systems. Such standards would focus on a range of issues, including robustness, interpretability, performance metrics, fail-safe design, traceability for data collection and management, privacy, and testability. These should be broad standards that serve as guidance for both government and commercial developers to develop operationally specific design requirements and testing metrics. DoD entities—including the JAIC and R&E—and the IC are already playing a role in the development of U.S. government AI standards, led by NIST and the Office of Science and Technology Policy. NIST’s 2019 plan on U.S. leadership on AI provides an important foundation and calls for the development of standards and metrics for trustworthiness (e.g., accuracy, explainability, resiliency, safety, reliability, objectivity, and security), complexity, domain-specific and context-dependent risk, and uncertainty.35 NIST and the State Department should play a leading role in promoting U.S. government standards for AI and ML development and testing domestically and throughout international standards-setting organizations, such as the International Standards Organization and Institute for Electrical and Electronics Engineering, and multilateral institutions, such as the OECD AI Policy Observatory. The promotion of U.S. standards globally will help bolster U.S. economic competitiveness, create a level playing field for U.S. industries who are collaborating with DoD and consequently subject to these standards, incorporate U.S. values and ethical principles into AI and ML development, and ensure that the United States and its allies are interoperable. As we’ve seen in other critical technology areas, the United States must ensure that competitors do not set standards, which make it harder to manage vulnerabilities and hinder U.S. efforts to establish the highest degree of ethics, safety, and risk management.

#### The plan establishes a norm for military AI – it sends a strong signal of the importance of safety and reliability – following through with actions is crucial to uphold our leadership

Horowitz and Scharre 2021 - Director of the Emerging Capabilities Policy Office in the Office of the Under Secretary of Defense for Policy and  Vice President and Director of Studies at CNAS [Michael, Paul, January 12 2021, “AI and International Stability: Risks and Confidence- Building Measures”, <https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures>, Acc 6/18/22. M.A.]

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](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn51) (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](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn52) A focus of the strategy was “leading in military ethics and AI safety.”[53](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn53) 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.

### Solvency – CBMs

**US leadership on AI TEVV is a Confidence Building Measure – the US is in a unique position to rally support. US leadership is necessary to make the CBM effective**

**Horowitz and Kahn, 2021 – Senior and Research Fellows at the Council on Foreign Relations** [Michael and Lauren, The Washington Quarterly 3-19-2021 “Leading in Artificial Intelligence through Confidence Building Measures”  [https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/0163660X.2021.2018794 acc on 6-21-2022](%20https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/0163660X.2021.2018794%20acc%20on%206-21-2022%20) TM]

US Leadership Can Decrease the Risk of Unintentional Conflict While these CBMs provide possible routes to reducing the risks to international stability from unintentional conflict related to military uses of AI, to be truly successful, they require not just US participation, but US leadership. The intersection of the strength of the American military, the network of American defense partnerships, and America’s role in AI research and development in the public and private sectors means the United States has unique convening power when it comes to international dialogue and agreements surrounding military uses of AI. The American public’s confidence in US leadership in AI also could render leadership in this arena not merely palatable to the general public, but highly supported.36 Avoiding inadvertent escalation and accidents caused by AI is also of universal interest. In particular, being able to dictate and determine the mechanisms, frameworks, and constraints put in place to do so would be useful for US national security moving forward. Building norms through standards of behavior for military AI use will increase the probability that America’s commitment to international humanitarian law and strong TEVV standards will shape how the rest of the world also uses military AI. Even if all nations do not follow America’s lead or join relevant CBM regimes, the more countries that join, the more they will lower the risk of accidents or miscommunication related to AI. Leadership in standard-setting and confidence-building measures will also enhance US military capability, rather than constraining the US military, assisting the United States in strategic competition. Unsafe AI systems do not just risk unintentional conflict and inadvertent escalation—they are less likely to be effective systems. Committing to lead the world in AI safety could create a ripple effect in the US defense enterprise and the private sector about putting a premium on safe and ethical AI, in turn making it more likely that military (and civilian) uses of algorithms are reliable, improving their utility for the military. Such signaling will also likely improve the reputation of the US military with Silicon Valley and AI/ML researchers, whose concerns about military uses of AI have loomed large since a protest halted Google’s renewal of its Project Maven contract in 2018. While concerns about Silicon Valley’s opposition to the Department of Defense are overstated according to some survey research,37 other surveys of AI/ML professionals show that increasing the emphasis on safety is a high priority.38 A public commitment to safety will thus help the Department of Defense improve at attracting top STEM talent, including AI/ ML talent, increasing its ability to keep the American military ahead. One might argue that the United States should let others lead on AI, focusing instead on developing AI-enabled capabilities and not concerning itself with how other countries behave. But there is no substitute for American leadership and its ability to rally countries around the world to support shared standards. If promoting norms of responsible behavior with AI encourages other states to use military applications of AI in more responsible ways, it will create a more ethical and predictable security environment, likely benefiting the United States. Additionally, current international dialogue about military uses of AI focuses almost exclusively on lethal autonomous weapon systems (LAWS), the subject of a Group of Governmental Experts in the Convention on Certain Conventional Weapons.39 Currently, the international conversation has been largely been driven by NGOs such as the Campaign to Stop Killer Robots.40 While such conversations help bring attention to some of these issues, they oversimplify the risks and fixate on worst-case scenarios that are more likely outcomes of artificial general intelligence or human level machine intelligence rather than technology today.

#### CBMs prevent global conflicts from autonomous systems – AI expansion is inevitable. Only US leadership can promote Safe expansion

Horowitz, 2021 - Professor of Political Science at the University of Pennsylvania [Michael Horowitz, 1-26-2021, , Bulletin of the Atomic Scientists, "How Joe Biden can use confidence-building measures for military uses of AI” https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/00963402.2020.1860331 accessed on 6-18-2022]

Autonomous naval ships, algorithms to interpret drone video footage – the US military and other armed forces around the world are thinking about how to incorporate applications of artificial intelligence systems into their operations. The Biden administration has an opportunity to foster international cooperation on military AI to reduce the risk of inadvertent conflict while still pursuing US military leadership. Perhaps no technology holds global imagination concerning the future of military power more than artificial intelligence (AI), in both positive and negative ways. Anticipating the impact of AI systems on the security environment, in 2018, the United States Department of Defense launched the Joint Artificial Intelligence Center (JAIC). The JAIC is one of several US military initiatives designed “to seize upon the transformative potential of Artificial Intelligence technology” (DOD 2019) and assist the American military in preparing for future wars filled with autonomous systems and algorithms. While it is still true that most military uses of AI are closer to concept than reality, many countries, including Russia and China – competitors seeking to challenge US conventional military superiority – are seeking to incorporate ever-greater levels of autonomy into their military operations. It will be a busy four years on the military AI front. President-elect Joe Biden’s administration will have to make critical choices about what strategy to pursue regarding the international governance of military AI. These choices will be key to ensuring that, even as the US military invests heavily in AI, these investments promote, rather than undermine, global stability. Competition for AI leadership is inevitable, and it is important not to exaggerate the potential for cooperation, but countries should have shared interests in preventing accidents or inadvertent war emerging from uses of AI. Confidence-building measures, best known from the Cold War, such as an AI equivalent to the Incidents at Sea Agreement, could decrease the likelihood that military uses of AI increase the risk of conflict.

**US dialogue with allies about AI safety is a confidence building measure that addresses the risk of accidental conflicts and mistaken escalation**

**Horowitz and Kahn, 2021 – Senior and Research Fellows at the Council on Foreign Relations** [Michael and Lauren, The Washington Quarterly 3-19-2021 “Leading in Artificial Intelligence through Confidence Building Measures”  [https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/0163660X.2021.2018794 acc on 6-21-2022](%20https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/0163660X.2021.2018794%20acc%20on%206-21-2022%20) TM]

The role of artificial intelligence (AI) in military use has been the subject of intense debates in the national security community in recent years— not only the potential for AI to reshape capabilities, but also the potential for unintentional conflict and escalation. For many analysts, fear that military applications of AI would lead to increased risk of accidents and inadvertent escalation looms large, regardless of the potential benefits. Those who are concerned can cite a plethora of potential ways things can go awry with algorithms: brittleness, biased or poisoned training data, hacks by adversaries, or just increased speed of decision-making leading to fear-based escalation. Yet, given its importance for the future of military power, it is imperative that the United States moves forward with responsible speed in designing, integrating, and deploying relevant military applications of AI. How should the United States simultaneously pursue AI swiftly while reducing the risk of unintentional conflict or escalation in the United States or elsewhere? The answer may lie in US leadership to promote responsible norms and standards of behavior for AI as part of a series of confidence-building measures (CBMs) tailored to reduce the likelihood of these scenarios. Four CBMs in particular could be fertile grounds for cooperation. First, the United States could promote dialogues with Russia and China, integrating discussions of AI into ongoing strategic stability talks with both nations. Second, the United States could announce norms of responsible behavior for military applications of AI and pursue a multilateral code of conduct surrounding those norms. Even if signatories are primarily democratic allies and partners, it could nevertheless make a powerful statement about where the United States stands and help to isolate those nations that opt-out. Third, the United States could declare that it will always keep a human-in-the-loop (HITL) when it comes to the use of nuclear weapons and invite other nuclear powers to commit to the same. Finally, the United States could lead the negotiation of an Autonomous Incidents Agreement—similar to the Incidents-at-Sea Agreement during the Cold War—designed to decrease the risk that the deployment of AI-enabled autonomous systems could create uncertainty or surprise that leads to unintentional conflict or inadvertent escalation. Given the interests that all nations have in avoiding unintentional conflict, CBMs focused in this area—where the United States and its competitors have shared interests—could prove effective. Moreover, CBMs can serve as building blocks for cooperation by enhancing mutual understanding and therefore trust.1 By focusing on CBMs that commit the United States to actions that it could pursue independently regardless, the United States can further ensure that, even if cooperation does not materialize, it is not constrained in its efforts to be a world leader in AI. US leadership internationally could therefore help the United States create a safer military AI future without undermining American efforts to leverage AI for military purposes.

#### CBMs reduce the danger of unintended conflict from untested AI.

Horowitz and Scharre, 2021 - Director of the Emerging Capabilities Policy Office in the Office of the Under Secretary of Defense for Policy and  Vice President and Director of Studies at CNAS [Michael, Paul, January 12 2021, “AI and International Stability: Risks and Confidence- Building Measures”, <https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures>, Acc 6/18/22. M.A.]

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.

#### CBMs can reduce the risk of AI instability through transparency and monitoring.

Horowitz and Scharre 2021 - Director of the Emerging Capabilities Policy Office in the Office of the Under Secretary of Defense for Policy and  Vice President and Director of Studies at CNAS [Michael, Paul, January 12 2021, “AI and International Stability: Risks and Confidence- Building Measures”, <https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures>, Acc 6/18/22. M.A.]

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](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn39) 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](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn40) 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](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn41) 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 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](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn42)

### Solvency – Arms Race

#### We must act now to prevent an escalating AI arms race – Nations are not fully autonomous yet, but will be soon.

Sharre, 2018 - director of the technology and national security program at the Center for a New American Security [Paul, 12 September, “Ultrafast computing is critical to modern warfare. But it also ensures a lot could go very wrong, very quickly.” <https://foreignpolicy.com/2018/09/12/a-million-mistakes-a-second-future-of-war/> ST]

Militaries around the globe are racing to build ever more autonomous drones, missiles, and cyberweapons. Greater autonomy allows for faster reactions on the battlefield, an advantage that is as powerful today as it was 2,500 years ago when Sun Tzu wrote, “Speed is the essence of war.” Today’s intelligent machines can react at superhuman speeds. Modern Chinese military academics have speculated about a coming “battlefield singularity,” in which the pace of combat eclipses human decision-making. The consequences of humans ceding effective control over what happens in war would be profound and the effects potentially catastrophic. While the competitive advantages to be gained from letting machines run the battlefield are clear, the risks would be grave: Accidents could cause conflicts to spiral out of control. Consider what has already happened with stock markets, where computers use algorithms to make decisions so quickly that microseconds make a difference of millions of dollars. Such trading has made brokers huge amounts of money—but has also produced extreme flash crashes that can send markets tumbling in minutes. Regulators have managed these risks by installing circuit breakers that can take a stock offline if the price moves too quickly, but battlefields lack these fail-safes. Flash crashes are bad enough; a flash war would be downright disastrous. Humans have already ceded control to machines in certain military domains. At least 30 countries—with Israel, Russia, and the United States leading the pack—employ human-supervised autonomous weapons to defend bases, vehicles, and ships. These weapons systems, such as the ship-based Aegis combat system, can detect incoming rockets and missiles and, if human supervisors do nothing, respond on their own by firing to eliminate the threat. Such automated responses allow the systems to defend against what are known as saturation attacks, in which salvos of missiles or rockets are launched at a target with such little notice that they could overwhelm human operators. For the time being, autonomous weapons such as these are used purely to protect human-occupied installations or vehicles. Humans supervise the weapons’ operation in real time and can intervene if necessary. Future autonomous weapons could lack these safeguards, however. A number of advanced militaries—including those of China, France, Israel, Russia, the United Kingdom, and the United States—are currently developing stealth combat drones intended to penetrate an adversary’s airspace. Once deep behind enemy lines, these drones might find their communications jammed, so they’re being designed to ensure they can continue to operate on their own. Most countries have not explained how their drones will operate under such circumstances and what rules of engagement they will follow. Countries could require their drones to get human authorization before launching any attacks. Doing so would allow the drones to bomb preapproved fixed targets but would require them to report back and get permission before attacking any newly discovered quarries. Such an approach sounds good in theory, but the problem is that these days many high-priority targets, such as air defense systems and ballistic missile launchers, are highly mobile. This mobility will increasingly tempt military planners to delegate lethal decision-making authority to machines, since doing so could give them an edge in reaction time. No battlefield is static, and the ability to rapidly react to a dynamic environment is critical to mission success—whether in the air, on the ground, or in cyberspace. Air combat strategists call this the OODA (Observe, Orient, Decide, and Act) loop in dogfighting. In the OODA loop paradigm of combat, pilots win dogfights not simply because they enjoy the best hardware but because they assess and react to their situations faster than their opponents, although better sensors and maneuverability might help shorten reaction times. Since machines can react faster than humans, automation will offer tremendous advantages in this competition. That means that the same competitive pressures that led to the creation of systems such as the Aegis could soon be introduced on a wider scale.

### Solvency – Life Cycle Testing

#### Expanding TEVV is necessary to ensure confidence in AI systems. Continuous testing is better than current efforts.

**National Security Commission on Artificial Intelligence 2022** [March 19, “NCSAI Final Report 2021 – Chapter 7: Establishing Justified Confidence in AI Systems”, <https://assets.foleon.com/eu-west-2/uploads-7e3kk3/48187/nscai_ch7_digital_02-26-21.24159789cf2e.pdf>]

3. Testing and Evaluation, Verification and Validation (TEVV). Having justified confidence in AI systems requires assurances that they will perform as intended, including when interacting with humans and other systems. The TEVV of traditional legacy systems is not sufficient at providing these assurances. As a result, agencies lack common metrics to assess trustworthiness that AI systems will perform as intended. To minimize performance problems and unanticipated outcomes, an entirely new type of TEVV will be needed. This is a priority task, and a challenging one. The federal government will need to increase R&D investments to improve our understanding of how to conduct AI and software-related TEVV. Toward this end: DoD should tailor and develop TEVV policies and capabilities to meet the changes needed for AI as AI-enabled systems grow in number, scope, and complexity in the Department. This should include establishing a TEVV framework and culture that integrates continuous testing; making TEVV tools and capabilities more readily available across DoD; updating or creating live, virtual, and constructive test ranges for AI-enabled systems; and restructuring the processes that underlie requirements for system design, development, and testing.6 National Institute of Standards and Technology (NIST) should provide and regularly refresh a set of standards, performance metrics, and tools for qualified confidence in AI models, data, and training environments, and predicted outcomes. NIST should lead the AI community in establishing these resources, closely engaging with experts and users from industry, academia, and government to ensure their efficacy.

**The DOD must improve TEVV by extending testing for the life cycle of the system – this is essential for AI which evolves over time.**

**Flournoy, Haines et al 2020 - former Under Secretary of Defense for Policy and Director of National Intelligence** [Michèle and Avril, Gabrielle Cheftiz Special Assistant to the Under Secretary of Defense for Policy October, “Building Trust through Testing Adapting DOD’s Test & Evaluation, Validation & Verification (TEVV) Enterprise for Machine Learning Systems, including Deep Learning Systems” https://cset. georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf Acc 6/23/22 JZ]

5. Bridge the gap between development and testing. ML/DL systems will require testing and verification across their entire life cycle, which will require stronger links between program managers and testers as well as methods to capture lessons learned throughout deployment. There are already good models for how this could be done. For example, developers for Project Maven need to submit to T&E in every sprint cycle, or they cannot move forward to the next stage of development. We recommend replicating and scaling this approach in other programs. DoD could also look to the Joint Improvised-Threat Defeat Organization, established in 2006, to rapidly field new technologies to counter IEDs. The organization’s approach to fielding prototypes and testing them in the field provides a useful model for the rapid deployment of ML/DL. Another way to bridge this gap is to leverage the testing framework and requirements language to inform the acquisition process. The JAIC is already using its acquisition process to impact development of AI technologies by using the AI ethical principles as “applicable standards.” One could envision a similar process with AI testing and governance standards. To do so, DoD will need to maintain a robust dialogue with industry on what is feasible to help inform the process.

#### Continuous TEVV is necessary for AI because initial tests will still have bugs in the testing programs themselves

Haugh, 2018 - project leader for the Institute for Defense Analysis [Brian A., September “The Status of Test, Evaluation, Verification, and Validation (TEV&V) of Autonomous Systems,” <https://www.ida.org/-/media/feature/publications/t/th/the-status-of-test-evaluation-verification-and-validation-of-autonomous-systems/p-9292.ashx>, 6/23/22 MD]

6. Elevated safety concerns and asymmetric hazard Traditionally, TEV&V personnel have relied on the training and common sense of equipment operators to provide many kinds of safety assurance, both in the field and on the test range. Autonomous systems potentially take many of the decisions underlying routine safety out of the hands (and minds) of operators, and depend instead on complex software that allows the system to “operate” itself. During Developmental Test and Evaluation and into Operational Test and Evaluation, it is likely that the software will still contain major bugsand that the algorithms and training data being used might not be the final choices. This creates a potential for various kinds of mischief—especially for weapon systems, highly-mobile systems, or other systems that could be dangerous in the hands of an unreliable operator. Similarly, the most striking successes of AI, machine learning, and autonomous systems to date have occurred in contexts where the cost of error is low. For autonomous military systems, this will not generally be the case—designating incorrect targets, responding to spurious cyber attacks, misidentifying 2-3 individuals or objects, crashing unmanned vehicles into obstacles, and other potential failure modes of autonomous systems are all potentially very costly. Apart from driverless car efforts, there are few active research programs today for applications with highly asymmetric hazard functions.

### NATO Key

**NATO standardization is key to TEVV reform – NATO smooths the testing transition**

**Gilli, 2020 - Senior Researcher at the NATO Defense College** [Andrea, NDC Research Paper No.15 – December ““NATO-Mation”: Strategies for Leading in the Age of Artificial Intelligence” https://www.ndc.nato.int/news/news.php?icode=1514 Acc 4/21/22 TA]

The above discussion highlights how, despite their technical nature, standards have an important political, economic and ethical dimension. Standards, however, also have an important military dimension, as the case of NATO highlights. Historically, NATO has played a critical role in standardization. As Paul Beckley notes, “with nearly 6,000 documents in its database, NATO is the most prolific international military standardization organization in the world”.318 Furthermore, standards are the backbone of interoperability – which, in turn, is the lifeblood of a military alliance. NATO collective capacity is, however, more than the simple sum of its parts, or “having the same missiles or tanks”.319 The emergence of AI, ML and BD highlights this point. As this new set of technologies are integrated into the economy as well as in countries’ armed forces, standards will provide a fundamental contribution. NATO has a key role to play here, both in ensuring a swift and risk-free technological transition through standards and in infusing its ethical and moral commitments into the new technologies.320 At least four main issues are important for the Alliance. Standards and ethics. NATO cannot lead and shape ethical discussions about AI if is unable to translate its ethical commitments into technology. The propagation channel to achieve this goal is constituted by technical and non-technical standards. However, in recent years, as Paul Beckley warns, NATO’s once robust standardization programme has lost vitality: since the early 2000s, NATO has developed few new standards and, since 2010, most have been doctrinal rather than technical.321 Additionally, while AI technologies evolve quickly, the standardization process may often be lengthy, and given its consensus-based nature, it can require significant time in large organizations. Key ethical values such as explainability, traceability or transparency need to be translated into technical standards, in order to have real meaning and effects.322 Following previous discussions, given that ML blurs the line between research and development, a robust system of T&E/V&V is necessary, both at national and at Alliance level.323 Working together on these standards is crucial to ensure that Allies can plug together their capabilities and that the same parameters are adopted, both for combat effectiveness and for economic efficiency.

**NATO is Key to safety and reliability because they can coordinate acquisitions and NATO has the experts**

**Stanley-Lockman and Trabucco, 2022 – prof of Defense and Strategic Studies, Nanyang Technological University and prof of Political Science, University of Copenhagen** [Zoe and Lena, The Oxford Handbook of AI Governance, March, “NATO’s Role in Responsible AI Governance in Military Affairs” https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780197579329.001.0001/oxfordhb-9780197579329-e-69 Acc 4/16/22 TA]

NATO has an important role to play in military standardization and Allied policy planning for safe, secure, and interoperable AI. This includes the coordinating role of the Conference of National Armaments Directors and the Command, Control and Consultation Board to implement complementary acquisition processes that fuse AI adoption measures with safety responsibilities. Furthermore, entities including STO and NSO have a significant role setting the technical baseline and promulgating materiel standards that provide the technical framework for safety and security. Although their staffs are themselves small, they both convene hundreds, if not thousands, of subject matter experts in working groups. As such they both offer unique technical networks to help shape safety and security in a way that minimize risk in operations. NATO’s resources and leadership are vital to using standards and coalition policy to instill safe and secure technological development, a necessary condition to interoperable and successful future operations.

#### Only NATO can solve – it is essential for cooperation and interoperability.

**Stanley-Lockman and Christie, 2021 - Innovation Officer in the Emerging Security Challenges Division in NATO and former Deputy Head of Innovation in NATO’s International Staff**  [Zoe and Edward, NATO Review, October “An Artificial Intelligence Strategy for NATO” <https://www.nato.int/docu/review/articles/2021/10/25/an-artificial-intelligence-strategy-for-nato/index.html>, BK]

One does not have to look far to see how Artificial Intelligence (AI) – the ability of machines to perform tasks that typically require human intelligence – is transforming the international security environment in which NATO operates. Due to its cross-cutting nature, AI will pose a broad set of international security challenges, affecting both traditional military capabilities and the realm of hybrid threats, and will likewise provide new opportunities to respond to them. AI will have an impact on all of NATO’s core tasks of collective defence, crisis management, and cooperative security. With new opportunities, risks, and threats to prosperity and security at stake, the promise and peril associated with this foundational technology are too vast for any single actor to manage alone. As a result, cooperation is inherently needed to equally mitigate international security risks, as well as to capitalise on the technology’s potential to transform enterprise functions, mission support, and operations. The continued ability of the Alliance to deter and defend against any potential adversary and to respond effectively to emerging crises will hinge on its ability to maintain its technological edge. Militarily, futureproofing the comparative advantage of Allied forces will depend on a common policy basis and digital backbone to ensure interoperability and accordance with international law. With the fusion of human, information, and physical elements increasingly determining decisive advantage in the battlespace, interoperability becomes all the more essential. Further, as competitors and potential adversaries invest in AI for military purposes, ensuring that Allies develop common responses to ensure their collective security will only become more urgent. With the formal adoption of the NATO AI Strategy, Allies have committed to the necessary cooperation and collaboration to meet these very challenges in both defense and security, naming NATO as the primary transatlantic forum. The aim of NATO’s AI Strategy is to accelerate AI adoption by enhancing key AI enablers and adapting policy, including by adopting Principles of Responsible Use for AI and by safeguarding against threats from malicious use of AI by state and non-state actors. By acting collectively through NATO, Allied governments also ensure a continued focus on interoperability and the development of common standards. Overall, with innovation ecosystems implicating different actors and faster technology lifecycles than typically included in traditional capability development systems, the NATO AI Strategy is also a recognition that exploitation of AI will require new efforts to foster and leverage the Alliance’s innovation potential, including through new partnerships and mechanisms. Taken together, these efforts will in turn strengthen the Alliance’s ability to pursue cooperative security efforts and to engage with international partners and other international organizations on matters of international security.

#### NATO adopting AI Responsible Use standards increase global modelling which draws an AI workforce and reduces proliferation. Only NATO has the expertise.

**Stanley-Lockman and Christie, 2021 - Innovation Officer in the Emerging Security Challenges Division in NATO and former Deputy Head of Innovation in NATO’s International Staff**  [Zoe and Edward, NATO Review, October “An Artificial Intelligence Strategy for NATO” <https://www.nato.int/docu/review/articles/2021/10/25/an-artificial-intelligence-strategy-for-nato/index.html>, BK]

Through the adoption of principles of responsible use, NATO and Allies are sending a deliberately public message to their domestic populations, to Allied forces, and to other states, reiterating the Alliance’s enduring values and commitments under international law. More than just an obligation, this democratic commitment is also a pre-condition for common policy bases among Allies – and for partnership with non-traditional innovators across the Alliance. Accelerating principled and interoperable adoption With the ethical aspects of adoption that the principles underscore, NATO has the chance to signal – and follow through on – responsibility at the core of its outreach efforts. This includes engagement with start-ups, innovative small and medium enterprises, and academic researchers that either have not considered working on defence and security solutions, or simply find the adoption pathways too slow or restrictive for their business models. In contrast to the development of traditional military platforms, AI integration entails fast refresh cycles and requires constant upgrading. This requires a change of mind-set for iterative, adaptive capability development, in contrast to sequential development cycles that take years to deliver small numbers of highly sophisticated platforms. With hostile state and non-state actors increasing their investments in Emerging and Disruptive Technologies including AI, this more flexible approach to adoption is all the more urgent. In this context, with its focus on TEVV and collaborative activities, the AI Strategy sets the framework for technological enablers to out-adapt competitors and adversaries. With more of a focus on agility and adaptation, NATO can make defense and security a more attractive sector for civilian innovators to partner with, while also allowing them to maintain other commercial opportunities. In doing so, efforts to bolster the transatlantic innovation ecosystem can also serve as a bulwark against undesirable foreign investment and technology transfers. NATO’s experience not only in operations, but also in trials, exercises, and experimentation provide several avenues in which Allies and NATO can test principles against intended use cases. This is further reinforced by NATO’s scientific and technical communities, which have worked on issues such as trust, human-machine and machine-machine interactions, and human-systems integration, among many others. This work requires coordination across the NATO Enterprise. Indeed, several stakeholders across the NATO Enterprise are already involved in the development of AI-related use cases, concepts, and programmes. With the AI Strategy, these activities can gain coherence to ensure the proper connections exist between all innovation stakeholders, including operational end-users.

#### NATO cooperation on AI spills over to other Emerging Disruptive Technologies due to their Data Exploitation Framework Policy

**Stanley-Lockman and Christie, 2021 - Innovation Officer in the Emerging Security Challenges Division in NATO and former Deputy Head of Innovation in NATO’s International Staff**  [Zoe and Edward, NATO Review, October “An Artificial Intelligence Strategy for NATO” <https://www.nato.int/docu/review/articles/2021/10/25/an-artificial-intelligence-strategy-for-nato/index.html>, BK]

To be sure, the implementation of accelerated, principled, and interoperable AI adoption depends not just on technology, but equally on the talented and empowered people who drive the technological state-of-the-art and integration forward. NATO has also dedicated attention to other AI inputs, notably through the development of a NATO Data Exploitation Framework Policy. With actions to treat data as a strategic asset, develop analytical tools, and store and manage data in the appropriate infrastructure, the Data Exploitation Framework Policy sets the conditions for the AI Strategy’s success. In addition to the interrelationships between data and AI, ensuring coherence between NATO’s efforts on AI and other Emerging and Disruptive Technologies such as autonomy, biotechnology, and quantum computing will be vital. As Allies and NATO seek to fulfil the aim of this AI Strategy, the linkages between responsible use, accelerated adoption, interoperability, and safeguarding against threats are critical. Indeed, these linkages will also apply to NATO’s follow-on work on other Emerging and Disruptive Technologies, including the development of principles of responsible use. More broadly, this entails further coherence between the work strands on these technologies, understanding that NATO’s future technological edge – and threats the Alliance will face – may depend on their convergence. As such, not only does the NATO AI Strategy apply to this foundational technology, but it also sets the stage for NATO’s and Allies’ ambitions with regards to other Emerging and Disruptive Technologies. For each of them, the future strategic advantage that comes with NATO innovation efforts will derive from the connections between ethical leadership, iterative adoption, and integration that prizes flexibility, interoperability, and trust.

### US Key

**US leadership on TEVV is key to implementing effective AI – it builds links to businesses and international organizations**

**Flournoy, Haines et al 2020 - former Under Secretary of Defense for Policy and Director of National Intelligence** [Michèle and Avril, Gabrielle Cheftiz Special Assistant to the Under Secretary of Defense for Policy October, “Building Trust through Testing Adapting DOD’s Test & Evaluation, Validation & Verification (TEVV) Enterprise for Machine Learning Systems, including Deep Learning Systems” https://cset. georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf Acc 6/23/22 JZ]

The future of U.S. leadership on ML/DL and DoD’s ability to harness these critical technologies depends on DoD investing in the science of ML/DL TEVV to develop new approaches and metrics, as well as standing up the coordination and governance mechanisms to accelerate progress and scale solutions. It will require developing the testing frameworks, requirements, and standards to bridge the gap between industry and government and shape a more iterative development and testing approach; shifting culture and practice toward the testing and certification of human-machine teams; and securing the talent, infrastructure, and resources to implement this new approach. Finally, DoD will need to deepen partnerships with the private sector, academia, non-governmental organizations, international organizations, and international partners to realize a multi-stakeholder approach to ML/DL development, testing, and deployment. Adapting the TEVV enterprise for ML/DL is critical to increasing trust in and, consequently, accelerating the deployment of these systems on a timeline consistent with the rate of innovation, operational need, and U.S. ethics and principles. The steps DoD and the broader U.S. government take now to adapt the ML/DL testing ecosystem will determine the long-term safety, reliability, and relevance of these systems in the coming decades.

#### US leadership on AI safety standards reduces the risk of unintended conflicts – The US is key to confidence building measures on Safety

**Horowitz and Kahn, 2021 – Senior and Research Fellows at the Council on Foreign Relations** [Michael and Lauren, The Washington Quarterly 3-19-2021 “Leading in Artificial Intelligence through Confidence Building Measures”  [https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/0163660X.2021.2018794 acc on 6-21-2022](%20https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/0163660X.2021.2018794%20acc%20on%206-21-2022%20) TM]

Leveraging US Leadership Advances in artificial intelligence, driven by machine learning methods and related approaches, are already reshaping international politics. Economics, societies, and now militaries are adapting, with various degrees of speed. As it is early in the age of AI, there is still significant uncertainty about the specific ways that AI and machine learning will impact military behavior and the future of war. One significant concern involves the potential for AI-enabled military applications to increase the risk of accidents, unintentional conflict, and inadvertent escalation. Poorly programmed, trained, or deployed algorithms could be subject to accidents and their uses could be misinterpreted by adversaries. Even if algorithms work as intended and give militaries an advantage, the increase in the speed of warfare from their use could create pressure for escalation in a crisis or early in a conflict. In the midst of competition with China and Russia, the United States can simultaneously benefit and be at risk from military applications of AI. Given its economic, military, and scientific leadership, the United States has a unique opportunity to shape the global AI landscape through the promotion of norms and CBMs that could decrease the risk of unintentional conflict and escalation. Only the United States has the convening power to bring allies and adversaries to the table, whether bilaterally or multilaterally, for dialogue around areas of shared interests—the areas most likely to be building blocks for cooperation. Key areas for potential cooperation include AI safety standards, dialogue on AI and strategic stability, commitments to keep humans in the loop for the use of nuclear weapons, and an Autonomous Incidents Agreement. All require further conversation. Through proposals that involve shared commitments to standards and policies that the United States would be willing to pursue unilaterally, the United States can increase global AI safety without revealing information that would compromise US capabilities or undermine US military adoption of AI. The United States can therefore leverage AI to ensure future military superiority and simultaneously decrease the risk that military uses of AI will have disastrous unintended consequences.

#### US AI leadership is essential to promote responsible use to avoid accidents.

Middendorf, 2021 - Former Secretary of the Navy [J William “Opinion: Artificial Intelligence’s Military Risks, Potential” <https://www.govtech.com/news/opinion-artificial-intelligences-military-risks-potential.html>, BK]

AI technologies are now widely used in tactical warfare situations, such as target acquisition for missiles launched from drones. But the actual command to fire the missile is reserved for the human operator. What might happen if the **decision time is reduced from minutes to seconds, removing the human operator from the process entirely?** And might that scenario be adopted by one of our less responsible adversaries? Or might it just happen accidentally — an accident far more plausible than ever before. A global leader in AI will emerge in the near future, achieving enormous international clout and the power to dictate the rules governing AI. The world will be safer and more peaceful with strong U.S. leadership in AI.

#### The US must lead International Cooperation on AI arms control – only maintaining human control can prevent instability

Garcia, 2021 - Vice-chair of the International Committee for Robot Arms Control[Denise, May 13, 2021, Nature.com, “Stop the emerging AI cold war,” <https://www.nature.com/articles/d41586-021-01244-z#:~:text=Proliferating%20military%20artificial%20intelligence%20will,on%20ethics%20and%20global%20cooperation.&text=Denise%20Garcia%20is%20a%20professor,Committee%20for%20Robot%20Arms%20Control>., 6/18/22 MD]

A race to militarize artificial intelligence is gearing up. Two years ago, the US Congress created the [National Security](https://go-gale-com.proxy.lib.umich.edu/ps/i.do?p=ITOF&u=umuser&id=GALE|A661476451&v=2.1&it=r) Commission on Artificial Intelligence (NSCAI). This March, it recommended that the [United States](https://go-gale-com.proxy.lib.umich.edu/ps/i.do?p=ITOF&u=umuser&id=GALE|A661476451&v=2.1&it=r) must accelerate artificial-intelligence (AI) technologies to preserve national security and remain competitive with China and Russia. This will undermine the United States' ability to lead emerging global norms on AI. In April, the European Commission published the first international legal framework for making AI secure and ethical; in January, the European Parliament issued guidelines stating that military AI should not replace human decisions and oversight. By contrast, the NSCAI recommendations advocate "the integration of AI-enabled technologies into every facet of war-fighting". Enhancing AI war-fighting capacity will decrease security in a world where the biggest threats are instability -- political, social, economic and planetary. The NSCAI should heed the research community. Some 4,500 AI and [robotics](https://go-gale-com.proxy.lib.umich.edu/ps/i.do?p=ITOF&u=umuser&id=GALE|A661476451&v=2.1&it=r) researchers have declared that AI should not make the decision to take a human life -- aligning with the European Parliament guidelines and the European Union regulation. The NSCAI resurrected disastrous ideas from the cold war and framed its report in terms of winning a competition for AI-enabled warfare. During the cold war, the drive to stay ahead in the technological race led to the accumulation of 70,000 nuclear weapons and today's global arsenal of 13,100 warheads. This brought extortionate costs: US$70 billion is spent annually to maintain nuclear weapons globally. Other threats demand similar investments: in 2019, climate-induced natural disasters displaced 25 million people, and decentralized conflicts forced 8.6 million to move. Still more threats affect infrastructure, such as the ransomware attack on 8 May that shut down a 8,850-kilometre US fuel pipeline. The NSCAI does not prioritize international cooperation to create new regulations. Indeed, it speaks against a global ban on autonomous weapons, saying that other countries cannot be trusted to comply. But an AI-militarization race would be profoundly destabilizing. Unlike nuclear arms, AI is already ubiquitous in civilian spheres, so the dual-use risks of, say, flying drones or computer night [vision](https://go-gale-com.proxy.lib.umich.edu/ps/i.do?p=ITOF&u=umuser&id=GALE|A661476451&v=2.1&it=r) are much higher. Since 2014, I have been an observer and adviser at United Nations meetings, and I testified in 2017 as part of the International Panel on the Regulation of Autonomous Weapons. In my view, rather than focusing on counting weapons or on particular weapons systems, policies should specify human intention and human-machine interaction, obligating countries to maintain human control over military force. Other agreements could mitigate malicious uses of AI, such as using facial recognition to oppress citizens or biased data to guide decisions about employment or incarceration. The world's people need protection from cyberattacks to infrastructure -- such as those on US hospitals in 2020 or those that hit national electrical grids. The NSCAI report calls for international standards for AI-enabled and autonomous weapons systems, arguing that if these systems are properly tested and designed, humans can use them to make the decision to kill, consistent with international humanitarian law. This is misleading: it's difficult to make machine learning's 'black box' nature fully interpretable, or to ensure that AI systems perform as expected after deployment. These systems learn from their environment, and the real world is never as simple as the laboratory. The NSCAI argues that the United States should seek commitments from Russia and China against autonomous nuclear weapons, even as it argues against treaties regulating other autonomous and AI weapons. Instead, the United States should negotiate decreases in nuclear arsenals and establish standards to keep humans in meaningful control. The NSCAI is too dismissive by discounting cooperation. The Chemical Weapons Convention, the Biological Weapons Convention, the UN Sustainable Development Goals and the 1987 Montreal Protocol are examples of accountability on which all the major powers worked together. The United States and Russia established the International Space Station by cooperating closely. Most nations want governance that controls the use of AI in war. In June 2020, the Global Partnership on Artificial Intelligence was created by the Group of Seven industrialized countries (G7) and called for human-centric development and use of AI. The partnership brings scientific and research communities together with industry and government to facilitate international cooperation. This is the path that the United States should take -- with scientists, researchers and industry alike. The relentless pursuit of militarization does not protect us. It diverts resources and attention from nearer existential threats, such as extreme weather events. With the world reeling from COVID-19 -- the shock of the century -- now is not the moment to hasten towards worldwide confrontation. In 2019 alone, climate disasters displaced almost one million people in the United States. China, too, is extremely vulnerable to global warming. This common ground could pave the way to cooperation, including stopping the emerging AI cold war. This is no time to embark on an exorbitant and ineffective race.

### DOD Key

#### The DOD is key to solvency – the DOD is essential to coordinate AI programs from research through long term implementation.

Tarraf, Shelton, and Parker 2019 - Senior Information Scientist, Senior Engineer, and physical Scientist at the RAND Corporation [ Danielle C., William, Edward, et al RAND Cooperation, “The Department of Defense Posture for Artificial Intelligence”, [file:///Users/MiraAgarwal/Downloads/RAND\_RR4229%20(2).pdf](file:///C:\Users\MiraAgarwal\Downloads\RAND_RR4229%20(2).pdf), Acc 6/18/22, M.A.]

The DIB’s Technologies and Capabilities Recommendation  5, cited earlier in this chapter, proposed establishing a centralized, focused, and well-resourced organization to propel applied research in AI and ML forward. The insights gathered from our industry interviews (see section “Industry: Organization” in Appendix C) lead us to believe there is indeed value in, if not strict necessity for, a centralized organization. This organization would have a mandate that goes beyond applied research and would be supported at the highest levels with long-term funding commitments to institute organizational change and scale AI across DoD. One of our industry interviewees noted that centralization at onset was key to their organization’s success, and premature decentralization of effort likely would have been detrimental (see section “Industry: Organization” in Appendix C).7 Based on the premises that (1) the JAIC is the focal point of DoD AI activity; (2) it fulfills the need for a centralized, well-resourced organization to scale AI and its impact across DoD (see above); (3) it will continue in that role for several years because of the expected timeline for AI deployment at scale across enterprise, mission-support, and operational AI; and (4) it needs to be able carry out all the roles it has been tasked with in the current strategy and establishing memo, we identified friction points that we discuss in the following paragraphs. 7 We note that this comment, and those of other industry interviewees, imply that if the effort to scale AI across an organization is successful, its natural ending point is the sunset of the centralized entity that drove the transformation as AI capabilities are diffused across the organization. We therefore expect the JAIC’s role to evolve, though we expect that to happen along a longer timeline (ten or more years) based on our assessment of the state of technology in Chapter Three. DoD Posture for Artificial Intelligence 47

#### Only the DOD has the expertise to lead AI efforts

**National Security Commission on Artificial Intelligence 2022** [March 19, “NCSAI Final Report 2021 – Chapter 7: Establishing Justified Confidence in AI Systems”, <https://assets.foleon.com/eu-west-2/uploads-7e3kk3/48187/nscai_ch7_digital_02-26-21.24159789cf2e.pdf>]

4. Leadership. Responsible development and fielding of AI requires end users and senior leaders to be aware of system capabilities and limitations so that they are not misused. It also requires subject-matter experts to support training, acquisition, risk assessment, and adoption of best practices as they evolve. Today, only the DoD has a dedicated lead for Responsible AI; employees in national security agencies taking on these roles typically do so on a voluntary, part-time basis. Without full-time dedicated staff, agencies will not succeed in fully adopting and implementing these recommended practices. The government should: Appoint a full-time, senior-level Responsible AI lead in each department or agency critical to national security and each branch of the armed services. Such an official should drive Responsible AI training, provide expertise on Responsible AI policies and practices, lead interagency coordination, and shape procurement policies.

**DOD leadership on AI safety is essential to engage our allies and maintain an AI advantage – this is critical during the Ukraine conflict**

**Nurkin and Konaev, 2022 - senior fellows at the Center for Strategy and Security at the Atlantic Council** [Tate and Margarita, 5-25-2022, “Eye to eye in AI: Developing artificial intelligence for national security and defense” Atlantic Council <https://www.atlanticcouncil.org/in-depth-research-reports/report/eye-to-eye-in-ai/> ARD]

The infamous Pentagon bureaucracy, an antiquated acquisition and contracting system, and a risk-averse organizational culture continue to inhibit the DoD’s ability to bring in external innovation and move more rapidly toward widespread AI integration and adoption. Solving systemic problems of this caliber is a tall order. But, important changes are already under way to facilitate DoD engagement with the commercial technology sector and innovative startups, and there seems to be a shared sense of urgency to solidify these public-private partnerships in order to ensure sustained US technological and military advantage. Still, much remains to be done in aligning the DoD’s and its industry partners’ perspectives about the most impactful areas for AI development, as well as articulating and implementing common technical standards and testing mechanisms for trustworthy and responsible AI. Key Takeaways and Recommendations The DoD must move quickly to transition from a broad recognition of AI’s importance to the creation of pathways, processes, practices, and principles that will accelerate adoption of the capabilities enabled by AI technologies. Without intentional, coordinated, and immediate action, the United States risks falling behind competitors in the ability to harness game-winning technologies that will dominate the kinetic and non-kinetic battlefield of the future. This report identifies three courses of action for the DoD that can help ensure the US military retains its global leadership in AI by catalyzing the internal changes necessary for more rapid AI adoption and capitalizing on the vibrant and diverse US innovation ecosystem, including ● prioritizing safe, secure, trusted, and responsible AI development and deployment; ● aligning key priorities for AI development and strengthening coordination between the DoD and industry partners to help close AI capability gaps; and ● promoting coordination between leading defense-technology companies and nontraditional vendors to accelerate DoD AI adoption. This report is published at a time that is both opportune and uncertain in terms of the future trajectory of the DoD’s AI adoption efforts and global geopolitics. The ongoing conflict in Ukraine has placed in stark relief the importance of constraining authoritarian impulses to control territory, populations, standards, and narratives, and the role that alliances committed to maintaining long-standing norms of international behavior can play in this effort. As a result, the authors urge the DoD to engage and integrate the United States’ allies and trusted partners at governmental and, where possible, industry levels to better implement the three main recommendations of this paper.

#### The DOD is key– pushing programs and funding through the JAIC sends the strongest signal of commitment

Tarraf et. al 2019 - Senior Information Scientist at the RAND Corporation [Danielle with William Shelton, Edward Parker, Brien Alkire, Diana Gehlhaus, Justin Grana, Alexis Levedahl, Jasmin Léveillé, Jared Mondschein, James Ryseff, “The Department of Defense Posture for Artificial Intelligence: Assessment and Recommendations”, LMSi]

Recommendation S-1: DoD should adapt AI governance structures that align authorities and resources with the mission of scaling AI. 66 The Department of Defense Posture for Artificial Intelligence As we noted in Chapter Four, DoD’s vision for AI and the scale, urgency, and unity of efforts conveyed in that vision are at odds with the visibility, authorities, and resources it has provided to the JAIC— the focal point of DoD AI. DoD needs to develop governance and organizational structures that align authorities and resources with its vision of scaling AI across DoD. The insights gathered from our industry interviews and the supporting change management literature (see section “Industry: Organization” in Appendix C) lead us to believe there is value in, if not strict necessity of, a centralized effort supported at the highest levels with long-term funding commitments to institute organizational change and scale AI across DoD. Indeed, one of our industry interviewees even noted that centralization at onset was key to their organization’s success, and premature decentralization of effort would have likely been detrimental (see section “Industry: Organization” in Appendix C).1 Starting from that premise, we highlight two possible options for organizational and governance structures, and the rationale for them, although noting that other options might be viable as well, subject to further study. The first option would likely require congressional support to execute, while the second can be executed without, as it aligns with current DoD procedures and organizational structures. Option 1: Enhance the visibility and authorities of the JAIC to enable it to carry out its mission of scaling AI and its impact across DoD, including budgetary and workforce authorities over the military services. Option 2: Take a two-pronged organizational approach as follows: • Establish a JAIC council chaired by the JAIC director and consisting of one AI leadership representative from each service.2 1 We note that this comment, and those of other industry interviewees, imply that if the effort to scale AI across an organization is successful, its natural ending point might be the sunset of the centralized entity that drove the transformation as AI capabilities are diffused across the organization. We therefore expect the JAIC’s role to evolve, though we expect that to happen along a longer timeline (ten or more years), based on our assessment of the state of AI technologies in Chapter Three. 2 Regardless of which governance structure is instituted, if any, establishment of a JAIC council, as described, could facilitate coordination between the JAIC and the services. Recommendations 67 • Establish or reinforce a centralized AI coordination and investment organization within each of the services, with appropriate visibility and authorities, to facilitate scaling AI and its impact across the service, and to promote mandated coordination with the JAIC. In either option: • The DSD should provide the JAIC director with opportunities, at least annually, to present and be heard at the Deputy’s Management Action Group (DMAG) forum (or whichever Deputy Secretary–level forum performs the functions of the DMAG). 3 The rationale for Option 1 is as follows: There is evidence to support that DoD has taken the right approach in establishing the JAIC as a centralized focal point for DoD’s AI strategy (see Chapter 4, Organization). The evidence also suggests that the JAIC, pending initial success, will need to continue in that role for several years because of the expected timeline for AI deployment across enterprise, missionsupport, and operational AI (Chapter Three). What DoD needs to do now is continue on that path by providing the requisite high-level support, visibility, and authorities (including directive and budget authorities) to enable the JAIC to enact change. Doing so would ensure that the JAIC has a chance of succeeding at its mandate of scaling AI and its impact across DoD. It would also ensure that DoD’s intent, messaging, and actions are all consistent.

### AT China and Russia

#### Expanding TEVV helps reduce the risk of Chinese and Russian AI – confidence building measure reduce pressures that cause crisis instability. Even if they do not model, US action alone reduces the risk.

Scharre 2021 – Director of the Technology and National Security Program at the Center for a New American Security [Paul, 6/28/21,https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/, “Debunking the AI Arms Race Theory”, 6/18/22, LND]

The National Security Commission on AI has demonstrated what such an approach might look like in practice. The commission’s 700-plus page report issued sweeping recommendations for enhancing U.S. competitiveness in AI and military adoption, but also dedicated an entire chapter to “Autonomous Weapon Systems and Risks Associated with AI-Enabled Warfare.”[70](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn70) With regards to safety concerns, the report acknowledged: Russia and China are likely to field AI-enabled systems that have undergone less rigorous TEVV than comparable U.S. systems and may be unsafe or unreliable … The United States should … highlight how deploying unsafe systems could risk inadvertent conflict escalation [and] emphasize the need to conduct rigorous TEVV.[71](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn71) The commission issued a number of recommendations to mitigate the risks of AI competition, including improving Defense Department TEVV processes and working with allies to develop “international standards of practice for the development, testing, and use of AI-enabled and autonomous weapon systems” to reduce the risk of accidents.[72](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn72) Recently, the Defense Department has taken positive steps toward emphasizing AI safety and improving TEVV processes. In May 2021, Deputy Secretary of Defense Kathleen Hicks issued a memorandum on implementing “Responsible AI.” The memo launched a series of internal bureaucratic structures, including “establishing a test and evaluation and verification and validation framework that integrates real-time monitoring, algorithm confidence metrics, and user feedback to ensure trusted and trustworthy AI capabilities.”[73](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn73) Additionally, in public comments Hicks emphasized the importance of AI “safety.”[74](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn74) These are important and valuable steps toward establishing the necessary bureaucratic processes to ensure U.S. military AI systems are robust and reliable as well as setting a constructive tone publicly. The United States has been active in promulgating norms about the responsible use of military AI. A deliberate approach toward acknowledging and mitigating the risks of AI competition need not come at the expense of adopting AI to improve military effectiveness. Ideally, a frank assessment of the risks of AI competition and U.S. transparency about measures it is taking to mitigate these risks would open the door to cooperative measures among competitors. There may be a variety of confidence-building measures that states could adopt to reduce the risks of AI competition.[75](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn75) “Track II” dialogues among academic experts to better understand these risks and potential cooperative measures are already underway. Future direct government-to-government dialogues could explore whether there is opportunity for common ground. Cooperative measures to reduce risk will depend on other states such as Russia and China engaging in good faith. However, there is no guarantee that they will do so. What the United States can do is improve its own internal processes for AI TEVV and for ensuring human responsibility. The United States should also publicly articulate why it would be in other states’ best interests to cooperate to avoid some of these mutual risks. Even as the United States adopts AI to improve its national defense, it should take measures — and incentivize others to do so as well — to ensure military AI systems are safe and that warfare remains under effective human control.

#### China and Russia would have good incentives to follow the US lead – if they did not, their AI work force would choose the US

**Horowitz and Kahn, 2021 – Senior and Research Fellows at the Council on Foreign Relations** [Michael and Lauren, The Washington Quarterly 3-19-2021 “Leading in Artificial Intelligence through Confidence Building Measures”  [https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/0163660X.2021.2018794 acc on 6-21-2022](%20https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/0163660X.2021.2018794%20acc%20on%206-21-2022%20) TM]

China and Russia might also have good reasons to sign onto standards that would commit to enhanced military AI safety. Neither country would like to be perceived as not following international standards on military AI safety, because it could undermine support within their AI research communities and ability to keep those researchers. Moreover, if China and Russia did not sign, it would help the United States build international credibility as a responsible AI actor, increasing its attractiveness in the global competition for AI talent.24 Thus, whether China or Russia sign or not, the United States would benefit. Overall, standardization would help to promote best practices concerning safety and ethics in the development and adoption of these technologies, and would help to alleviate some pressure and remove some sources of error during use.

### AT Testing Fails

#### No one TEVV system covers all of AI – there are different Tests and Verifications that apply to each sub area of AI

Batarsheh et al. 2021 - Professor of Electrical and Computer Engineering, Virginia Polytechnic Institute [Feras, Laura Freeman Director, Intelligent Systems Division – National Security Institute Research Associate Professor, Chih Hao Huang, Data Analyst at Turing Research. April 26 “A survey on artificial intelligence assurance” https://link-springer-com.proxy.lib.umich.edu/article/10.1186/s40537-021-00445-7#Sec8]

The state of AI assurance This section introduces some milestone methods and discussion in AI assurance. Many of the discussed works rely on standard software validation and verification methods. Such methods are inadequate for AI systems, because they have a dimension of intelligence, learning, and re-learning, as well as adaptability to certain contexts. Therefore, errors in AI system “may manifest themselves because of autonomous changes” [211], and among other scenarios would require extensive assurance. For instance, in expert systems, the inference engine component creates rules and new logic based on forward and backward propagation [20]. Such processes require extensive assurance of the process as well as the outcome rules. Alternatively, for other AI areas such as neural networks, while propagation is used, taxonomic evaluations and adversarial targeting are more critical to their assurance [145]. For other subareas such as machine learning, the structure of data, data collection decisions, and other data-relevant properties need step-wise assurance to evaluate the resulted predictions and forecasts. For instance, several types of bias can occur in any phase of the data science lifecycle or while extracting outcomes. Bias can begin during data collection, data wrangling, modeling, or any other phase. Biases and variances which arise in the data are independent of the sample size or statistical significance, and they can directly affect the context or the results or the model. Other issues such as incompleteness, data skewness, or lack of structure have a negative influence on the quality of outcomes of any AI model and require data assurance [117]. While the historic majority of methods for knowledge-based systems and expert systems (as well as neural networks) aimed at finding generic solutions for their assurance [21, 218], and [166], other “more recent” methods were focused on one AI subarea and one domain. For instance, in Mason et al. [142, 144], assurance was applied to reinforcement learning methods for safety–critical systems. Prentzas et al. [174] presented an assurance method for machine learning as its applied to stroke predictions, similar to Pawar’s et al.’s [167] XAI for healthcare framework. Pepe et al. [169], and Chittajallu et al.’s [42] developed a method for surgery video detection methods. Moreover, domains such as law and society would generally benefit from AI subareas such as natural language processing for analyzing legal contracts [135], but also require assurance.

### AT Will Shift to Semi Autonomous

#### Semi Autonomous weapons have a Much Lower risk than autonomous ones – there is always a failsafe.

Scharre, 2016 -- Vice President and Director of Studies at CNAS [Paul, Feb 2016, Center for New American Security, “Autonomous weapons and operational risk”, https://s3.us-east-1.amazonaws.com/files.cnas.org/documents/CNAS\_Autonomous-weapons-operational-risk.pdf?mtime=20160906080515&focal=none, Acc 6/21/22, M. A.]

Autonomous weapon systems—potential future weapons that would select and engage targets on their own—raise a host of legal, ethical, and moral questions. They also raise critically important considerations regarding safety and risk. Autonomous weapons have a qualitatively different degree of risk than equivalent semi-autonomous weapons that would retain a human in the loop. The consequences of a failure that causes the weapon to engage an inappropriate target could be far greater with an autonomous weapon. The result could be fratricide, civilian casualties, or unintended escalation in a crisis. Humans are not immune from errors, and semi-autonomous weapons can also fail. However for semi-autonomous weapons, requiring a human in the loop to authorize each engagement creates a natural fail-safe. If the weapon system begins to fail, the human controller can modify the weapon system’s operation or halt the engagement before further damage is done. With an autonomous weapon, however, the damage potential before a human controller is able to intervene could be far greater. In the most extreme case, an autonomous weapon could continue engaging inappropriate targets until it exhausts its magazine, potentially over a wide area. If the failure mode is replicated in other autonomous weapons of the same type, a military could face the disturbing prospect of large numbers of autonomous weapons failing simultaneously, with potentially catastrophic consequences.

### AT No Definition of AI

#### There are clear definitions of TEVV – field experts provide their own definitions

Batarsheh et al. 2021 - Professor of Electrical and Computer Engineering, Virginia Polytechnic Institute [Feras, Laura Freeman Director, Intelligent Systems Division – National Security Institute Research Associate Professor, Chih Hao Huang, Data Analyst at Turing Research. April 26 “A survey on artificial intelligence assurance” https://link-springer-com.proxy.lib.umich.edu/article/10.1186/s40537-021-00445-7#Sec8]

Verification: “The process of evaluating a system or component to determine whether the products of a given development phase satisfy the conditions imposed at the start of that phase”. Validation: “The process of evaluating a system or component during or at the end of the development process to determine whether it satisfies specified requirements” (Gonzalez and Barr, 2020). Another definition for V&V is from the Department of Defense, as they applied testing practices to simulation systems, it states the following: Verification is the “process of determining that a model implementation accurately represents the developer’s conceptual descriptions and specifications”, and Validation is the process of “determining the degree to which a model is an accurate representation” [60]. Testing: according to the American Software testing Qualification Board, testing is “the process consisting of all lifecycle activities, both static and dynamic, concerned with planning, preparation and evaluation of software products and related work products to determine that they satisfy specified requirements, to demonstrate that they are fit for purpose and to detect defects”. Based on that (and other reviewed definitions), testing includes both validation and verification. Assurance: this term has been rarely applied to conventional software engineering; rather, it is used in the context of AI and learning algorithms. In this manuscript, based on prior definitions and recent AI challenges, we propose the following definition for AI assurance: A process that is applied at all stages of the AI engineering lifecycle ensuring that any intelligent system is producing outcomes that are valid, verified, data-driven, trustworthy and explainable to a layman, ethical in the context of its deployment, unbiased in its learning, and fair to its users. Our definition is by design generic and therefore applicable to all AI domains and subareas. Additionally, based on our review of a wide variety of existing definitions of assurance, it is evident that the two main AI components of interest are the data and the algorithm; accordingly, those are the two main pillars of our definition. Additionally, we highlight that the outcomes the AI enable system (intelligent system) are evaluated at the system level, where the decision or action is being taken.

#### The DOD definition of autonomous weapons is the best – it focuses on functionality, best reflects the concerns about AI, and gives urgency to the debate.

Amoroso and Tamburrini, 2021 - Prof of International Law at the University of Cagliari and Prof of Philosophy of Science and Technology at the University of Naples Federico [Daniele and Guglielmo, Feb Italian Journal of International Affairs “In Search of the ‘Human Element’: International Debates on Regulating Autonomous Weapons Systems” <https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/03932729.2020.1864995> TM]

Despite heated diplomatic debates, agreement is still lacking among states about what an AWS is. Nevertheless, discussion about this fundamental issue is not fragmented into a myriad of competing notions of ‘autonomy’. Rather, it is polarised around two basic construals, epitomised by those advanced by the UK Ministry of Defence (MoD) and the US Department of Defense (DoD). The UK MoD sets a very demanding requirement for a weapons system to be autonomous. Indeed, the MoD’s Joint Doctrine Publications devoted to Unmanned Aircraft Systems describe AWS as systems “capable of understanding higher level intent and direction”, thereby being able “to take appropriate action to bring about the desired state” (UK MoD 2011, 14; 2017, 13). This is a tall order. Neither existing nor foreseeable weapons systems genuinely comprehend the meaning of higher-level goals and intentions. And there are presently no educated guesses as to the prospects for their development. Thus, weapons systems that are genuinely autonomous according to the UK MoD are projected into an undetermined technological future. If one endorses this approach, then ethical and legal discussions about AWS are simply premature in that they concern neither present nor foreseeable technological developments. This conclusion, however, conflicts with the intuitions underlying current academic and diplomatic debates, according to which normative analysis and deliberation on AWS are an urgent issue (Amoroso and Tamburrini 2017, 3-4). Remarkably absent in the US DoD Directive on ‘Autonomy in Weapons System’ is any attempt to account for autonomy in terms of high-level intent comprehension. Instead, this Directive introduces a functional, task-oriented condition for weapons systems to be considered autonomous: a weapon system must be able, after activation, to “select and engage targets without further intervention by a human operator” (US DoD 2012). It is of the utmost relevance for the international debate that the US DoD requirement on autonomy was embraced by the International Committee of the Red Cross (ICRC 2016, 1) and the Campaign (Campaign 2013, 1). Therefore, the chief stakeholders – the technologically most advanced military power, the main guardian of international humanitarian law (IHL) and the coalition of NGOs advocating a ban on AWS – converged on this formulation of necessary properties of autonomy.2 Such an international convergence is a good reason to endorse the DoD’s functional condition of autonomy as a starting point for AWS legal and ethical debates. There is, however, another reason for preferring the DoD over the UK MoD’s approach: it supports the intuition that normative analysis and deliberation about AWS is an urgent and impending issue since a variety of existing and technologically foreseeable systems satisfy the DoD functional condition.

#### AI should be defined in the context it is used – a broad definition is necessary for this latitude.

**Artificial Intelligence/Machine Learning Risk & Security Working Group, 2020** [Wharton, ‘Artificial Intelligence Risk & Governance’, <https://ai.wharton.upenn.edu/artificial-intelligence-risk-governance/>]

1.3 Definitions & Assumptions We use several terms throughout this document specific to AI/ML, some of which are subject to vigorous discussions and debates in the research community. Because this document attempts to form a starting point for broader AI/ML governance and risk management efforts, we purposely leverage and encourage readers to refer to various papers for the definition of AI. As such, we note that specific definitions should be (and generally are) tailored to each organization depending on the scope, risk appetite, internal structure, culture, and implementation details of AI/ML efforts. While there is no universally accepted definition of AI, it is generally understood to refer to “a branch of computer science dealing with the simulation of intelligent behavior in computers, or the capability of a machine to imitate intelligent human behavior.” Generally, machine learning is referred to as, “a field of computer science that uses algorithms to process large amounts of data and learn from it.”[2] The term AI is broadly used and typically includes aspects of machine learning and natural language processing. For purposes of this paper, the focus is largely on the use of and potential risks related to machine learning, though the overarching discussion applies more broadly to the abovementioned areas.

### AT Private Circumvention

#### DOD standards for TEVV promotes effective deployment of AI – it gives motivation and guidance to private industry and government bureaucracy

**Flournoy, Haines et al 2020 - former Under Secretary of Defense for Policy and Director of National Intelligence** [Michèle and Avril, Gabrielle Cheftiz Special Assistant to the Under Secretary of Defense for Policy October, “Building Trust through Testing Adapting DOD’s Test & Evaluation, Validation & Verification (TEVV) Enterprise for Machine Learning Systems, including Deep Learning Systems” https://cset. georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf Acc 6/23/22 JZ]

As a general matter, investments in the science of ML/DL TEVV are a critical prerequisite to developing the most effective and efficient standards, tools, and methodologies needed to assure system performance. As these technologies and their applications evolve, new areas of research will undoubtedly arise that warrant investment. 3. Develop a tailored, risk-based framework for ML/DL testing and safety. The AI/ML TEVV CFT should lead on the development of a framework that establishes architecture and testing standards for TEVV. DoD cannot have a one-sizefits- all approach; it needs a flexible testing framework that is mission and use-case dependent. A DoD-wide testing framework for AI/ML will help shorten the testing cycle and make test results interpretable and comparable across the Department. This framework should also incorporate DoD’s legal and ethical requirements, serving as implementation guidance for the AI ethics principles and 3000.09. The DIB AI ethics principles call on the JAIC to create a taxonomy of DoD use cases of AI, based on their ethical, safety, and legal risk considerations. The CFT should leverage this taxonomy and develop corresponding testing criteria and safety standards. Testing standards, for levels of interpretability or assurance, for example, should be determined based on several dimensions of risk and performance, including: the likelihood of error detection, the consequence of the error given the complexity of the operating context, the potential for unintended escalation, the size of the attack surface, system performance relative to that of a human operator, impact on human decision-making, and the risk associated with not adopting a system (e.g., the risk may be a 10 percent error for DL, but a 30 percent error without the system). For example, AI for business process automation would likely score low on all of these risk criteria, while AI for critical network cybersecurity would likely score high on all and, therefore, necessitate stricter and more expansive TEVV requirements. For higher-risk applications, DoD may need to require systems to be designed with fail-safe systems or operated only as part of a human-machine team to help mitigate risks and govern system performance. Researchers are currently developing methods for monitoring system performance and constraining a system to a set of allowable, predictable behaviors and mitigate the risk of failure and unintended escalation. A DoD-wide AI TEVV framework will help decision-makers manage the tradeoff between the risks of failure and the value of deployment, while advancing the development of clear and consistent requirements for system design and metrics for performance evaluation.

#### DOD standards for TEVV will leverage its influence with private companies and set a global norm for safety.

**Flournoy, Haines et al 2020 - former Under Secretary of Defense for Policy and Director of National Intelligence** [Michèle and Avril, Gabrielle Cheftiz Special Assistant to the Under Secretary of Defense for Policy October, “Building Trust through Testing Adapting DOD’s Test & Evaluation, Validation & Verification (TEVV) Enterprise for Machine Learning Systems, including Deep Learning Systems” https://cset. georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf Acc 6/23/22 JZ]

4. Translate the testing framework into testable, verifiable requirements to be used by the private sector and build an integrated team to leverage this approach. DoD should establish a process for translating the testing framework into testable and verifiable requirements. DoD requirements would help standardize processes for industry contractors who develop AI for DoD and promote a faster and cheaper TEVV process by enabling the private sector to do some testing throughout development. The development of such requirements and standard processes would allow DoD to leverage the talent and expertise in the private sector, while maintaining DoD’s safety and risk standards and employing DoD’s operational knowledge and adversarial testing capabilities. Such requirements could additionally serve as the basis for standards that are promoted by the U.S. Government across the private sector and internationally, as discussed in a later recommendation. To realize this approach, DoD should build integrated, multi-disciplinary teams that reflect the entire development, testing, and sustainment life cycle. One model is the JAIC’s project manager model, which incorporates experts on product, policy, legal, test and evaluation, and requirements into one team. A similar approach is the Navy’s AI "DevRon” concept, a single entity accountable start to finish for the life cycle of capability development.29 Scaling this approach would help ensure requirements take into account the unique challenges of ML/DL TEVV. Requirements should also be written to advance the integration of ethics into the design and testing process. One case study for how this is already being done is DARPA’s Urban Reconnaissance through Supervised Autonomy (URSA) program, which is developing AI-infused drones that can help prevent friendly fire and civilian casualties in urban battles.30 This program brought in ethicists to anticipate challenges before development even began, with the aim of integrating ethics into the development loop. This URSA program offers useful lessons on how to conduct ethically aligned design of military systems that enables testing for robustness, safety, and security and ensures these systems are reliable and governable. Developing requirements and standards for ML/DL will help strengthen the link between DoD and industry, ensure a hybrid approach that leverages both private and public sector resources, and inject DoD’s safety and ethical requirements into industry development practices.

### AT Militaries are Conservative

#### The DOD has a history of pushing out technology prematurely

Scharre 2021 – Director of the Technology and National Security Program at the Center for a New American Security [Paul, 6/28/21,https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/, “Debunking the AI Arms Race Theory”, 6/18/22, LND]

That militaries may incur risks from moving too quickly in adopting new technology is counter to the common caricature of military culture as conservative, hidebound, and resistant to innovation. While this caricature is not entirely fair — militaries do innovate even in peacetime[49](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn49) — the lack of direct, observational feedback on performance in a realistic competitive environment, akin to marketplace dynamics for commercial companies, can mean that militaries are often slow to adapt to changing circumstances. A variety of factors can affect military adoption of new technologies,[50](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn50) and adoption rates can vary considerably depending on the technology, state, and military community. Across a range of contemporary technologies militaries lag the private sector. For example, modern-day militaries are behind the private sector in adopting information technology, human performance optimizing technologies, and personnel best practices. With an increasing amount of technological innovation occurring outside of the defense sector, this lag is likely to continue.[51](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn51) Yet, the key risk factor for military AI systems is not the timing of when militaries begin the process of adoption, but the taking of shortcuts in safety to accelerate fielding new AI capabilities. Adopting technology is a multi-stage process, involving research and development, experimentation, prototyping, technology maturation, production, testing, and fielding. It is possible for militaries to move slowly in one stage and quickly (or via shortcuts) in others. While there are many areas in which the U.S. military’s adoption of AI, autonomy, robotics, and uninhabited vehicles is moving slowly due to a variety of bureaucratic obstacles, it is also possible that the United States could rush parts of the adoption process and end up with immature technology in production or even in the field. This mixed dynamic, of moving slowly in some aspects of technology development and taking shortcuts in others, has been present in other defense programs. The F-35 fighter jet went into production before the first test flight, a decision that the Defense Department’s top acquisition official, Frank Kendall, later characterized as “acquisition malpractice.”[52](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn52) Yet, the whole acquisition program took 25 years from its initial conception to its first operational deployment.[53](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn53) The F-35 is still not in full rate production, 28 years after it was initially conceived of.[54](https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/" \l "_ftn54) The F-35 program moved too quickly in some areas, introducing unnecessary risk, even while it was overall encumbered by the laborious pace typical of major defense acquisition programs. Militaries’ sluggish bureaucracy is, therefore, no defense against shoddy testing and premature fielding.

### Potential Solvency Mechanisms

### Solvency – Transparency

#### Promoting multilateral standards for transparent AI TEVV helps establish global norms and encourage modelling without revealing specific technology secrets

Horowitz and Scharre 2021 - Director of the Emerging Capabilities Policy Office in the Office of the Under Secretary of Defense for Policy and  Vice President and Director of Studies at CNAS [Michael, Paul, January 12 2021, “AI and International Stability: Risks and Confidence- Building Measures”, <https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures>, Acc 6/18/22. M.A.]

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](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn59) 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](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn60) 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.

### Solvency – Adversarial Testing

#### Adversarial testing will improve TEVV because it helps identify vulnerabilities

Haugh, 2018 - project leader for the Institute for Defense Analysis [Brian A., September “The Status of Test, Evaluation, Verification, and Validation (TEV&V) of Autonomous Systems,” <https://www.ida.org/-/media/feature/publications/t/th/the-status-of-test-evaluation-verification-and-validation-of-autonomous-systems/p-9292.ashx>, 6/23/22 MD]

C. Adversarial Testing As noted in Challenge 6, many defense applications will require very low probabilities of high-cost events occurring. For systems with extremely large and highly nonlinear state spaces, it will not be possible to provide that assurance statistically using traditional Design of Experiments techniques. Adversarial testing, especially when combined with machine learning and automated test design methods, provides a potentially more efficient means to identify and eliminate potential failure modes. In this approach, analytical techniques are used to identify test scenarios in which the system is most likely 3-5 to perform unacceptably and focuses testing in those portions of the state space to maximize diagnostic information and inform robust design. As a specific example, one could apply reinforcement learning to the selection of environmental factors as input to the test cases in a modeling and simulation driven test bed that included the AI software running the autonomous systems. The environment could be treated as a “thinking adversary” in an asymmetric game. The environment would learn optimal strategies to defeat the autonomous system, even as the autonomous system learned improved strategies to counter the environment. This would both maximize the chance of finding key vulnerabilities and weaknesses and help discover ways to mitigate or avoid those vulnerabilities. The vulnerabilities need not be isolated within the autonomous system itself—they might also arise from the proposed CONOPS, especially in the case of significant HMT. They could also be associated with bias or incompleteness in the training data used in supervised learning to develop the autonomous capabilities. Cognitive instrumentation would be needed to distinguish inadequate algorithms or models from inadequate CONOPS or flawed training. Finally, we note that this approach requires domain expertise as well. The operating environment and mission must be well enough understood so that the apparent “weaknesses” are not associated with impossible circumstances. If the weakness occurs only during heavy rainfall at temperatures below –40° C, it will be a weakness we can live with. More subtly, the domain expertise will need to be sufficient to estimate probability of occurrence of the conditions that exposed the weakness. As mentioned before, utilization in the defense sector will require attention to the likelihood of seriously adverse outcomes.

### Solvency – Attached Funding

#### Attaching TEVV budgeting to each AI effort signals the importance of safety

Tarraf et. al 2019 - Senior Information Scientist at the RAND Corporation [Danielle with William Shelton, Edward Parker, Brien Alkire, Diana Gehlhaus, Justin Grana, Alexis Levedahl, Jasmin Léveillé, Jared Mondschein, James Ryseff, “The Department of Defense Posture for Artificial Intelligence: Assessment and Recommendations”, LMSi]

Recommendation T-5: All funded AI efforts should include a budget for AI VVT&E. This recommendation, in support of Recommendation S-2, is a forcing function that is relatively simple to implement and that can help ensure that the consideration of VVT&E is baked into the R&D of AI techniques and the design of AI solutions rather than considered as an afterthought further down the line. Although VVT&E during early R&D phases should be commonplace, we make this recommendation to explicitly reinforce its critical importance, to highlight the present lack of foundations for VVT&E in AI, and the importance of developing that science Our next recommendation is a tactical one, unrelated to VVT&E, but speaks to the need to bring developers of AI technologies together with users and operators to enable success in technology development.

### Solvency – Run Time Monitoring

#### Adding Run Time Monitoring improves traditional TEVV by extending testing beyond simulations

Haugh, 2018 - project leader for the Institute for Defense Analysis [Brian A., September “The Status of Test, Evaluation, Verification, and Validation (TEV&V) of Autonomous Systems,” <https://www.ida.org/-/media/feature/publications/t/th/the-status-of-test-evaluation-verification-and-validation-of-autonomous-systems/p-9292.ashx>, 6/23/22 MD]

D. Run-Time Monitoring Given the difficulty of assuring that a system will not exhibit specific undesired behaviors, a natural thought is to instead monitor the system during operations and intervene when bad behavior is imminent. This approach is already common in engineering practice for safety-critical systems. It is mentioned explicitly in the recent US Unmanned Systems Integrated Roadmap 2017-2042: For the most demanding adaptive and non-deterministic systems, a new approach to traditional TEVV will be needed. For these types of highly complex autonomous systems, an alternate method leveraging a run-time architecture that can constrain the system to a set of allowable, predictable, and recoverable behaviors should be integrated early into the development process. Emergent behaviors from large-scale deployment of interacting autonomous systems poses a difficult challenge. The analysis and test burden would thereby, be shifted to a simpler, more deterministic run-time assurance mechanism. The effort for new approaches to TEVV endeavors to provide a structured argument, supported by evidence, justifying that a 3-6 system is acceptably safe and secure not only through offline tests, but also through reliance on real-time monitoring, prediction, and fail-safe recovery.1 Although this mechanism might indeed be simpler than avoiding unpredictable behaviors in the first place, it is not without its own challenges. In general, any behavior whose dependability cannot be adequately assured through system design and training would need to be monitored, with a robust intervention standing by. This means not only intervening when the system is about to execute some undesired physical action (e.g., one that might risk harm to the system or to humans), but also intervening in any case where the system is making a bad decision or misinterpreting its environment. Detecting such cases and handling them gracefully will not always be easy. Research is required into architectures to support this concept, instrumentation needs and control algorithms to predict and avoid specific failure modes, systematic identification of conditions to be monitored for, robustness against attacks designed to invoke fail-safe behaviors, and so forth. It goes without saying that the fail-safe systems would themselves need to be verified and validated as well.

### Solvency – Private Sector Coordination

**Coordinating with the private sector on TEVV will speed up adoption of AI and build DOD leadership for testing Standards**

**Flournoy, Haines et al 2020 - former Under Secretary of Defense for Policy and Director of National Intelligence** [Michèle and Avril, Gabrielle Cheftiz Special Assistant to the Under Secretary of Defense for Policy October, “Building Trust through Testing Adapting DOD’s Test & Evaluation, Validation & Verification (TEVV) Enterprise for Machine Learning Systems, including Deep Learning Systems” https://cset. georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf Acc 6/23/22 JZ]

There is insufficient coordination between DoD, the private sector, and academia. DoD needs a hybrid approach to TEVV that leverages DoD, academic, and industry research, infrastructure, and talent. The majority of ML/DL innovation will come from the private sector and academia, as will most of the insight into how to test, benchmark, and assure these systems. However, DoD has an important role to play in integrating, scaling, and deploying these solutions. Further, it can dedicate significant resources to basic and applied research and use its market power to influence the development and promotion of national standards for at least certain industries. DoD also has the unique capability to do adversarial testing, with access to threat intelligence and operational knowledge that can inform realistic modeling and simulation. DoD should, therefore, focus on unique use cases where there is no commercial relevance or where sharing the data or algorithm would reveal sensitive or classified information. DoD should, when possible, leverage commercial TEVV methods and tools, such as Microsoft Azure and Amazon Web Services secure environments and tooling.23 In many cases, however, industry methods will not be applicable, given the safety- critical application and unique classification of DoD data, requiring a hybrid model of development and testing informed by academic research. Further, DoD needs to engage the private sector to develop an intellectual property strategy both parties can live with that includes access to sufficient data for continuous testing. DoD should also engage in a sustained dialogue with commercial developers to inform how DoD defines the requirements for ML/DL testing and performance based on what is technologically feasible, now and in the future. There are some successful models for this cooperation, such as the Army’s AI Hub—a consortium of industry, government, and academic partners based at Carnegie Mellon University— which works with the JAIC and other DoD AI entities to provide independent assessments of key research questions.24 Scaling this effort will require a senior DoD champion, such as the Under Secretary for R&E, who values this work and can promote it across the Department.

### AT Military Readiness DA

#### Turn – Increasing TEVV standards helps the US military – it gives them more reliable weapons, and draws the best AI talent into our workforce.

**Horowitz and Kahn, 2021 – Senior and Research Fellows at the Council on Foreign Relations** [Michael and Lauren, The Washington Quarterly 3-19-2021 “Leading in Artificial Intelligence through Confidence Building Measures”  [https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/0163660X.2021.2018794 acc on 6-21-2022](%20https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/0163660X.2021.2018794%20acc%20on%206-21-2022%20) TM]

Leadership in standard-setting and confidence-building measures will also enhance US military capability, rather than constraining the US military, assisting the United States in strategic competition. Unsafe AI systems do not just risk unintentional conflict and inadvertent escalation—they are less likely to be effective systems. Committing to lead the world in AI safety could create a ripple effect in the US defense enterprise and the private sector about putting a premium on safe and ethical AI, in turn making it more likely that military (and civilian) uses of algorithms are reliable, improving their utility for the military. Such signaling will also likely improve the reputation of the US military with Silicon Valley and AI/ML researchers, whose concerns about military uses of AI have loomed large since a protest halted Google’s renewal of its Project Maven contract in 2018. While concerns about Silicon Valley’s opposition to the Department of Defense are overstated according to some survey research,37 other surveys of AI/ML professionals show that increasing the emphasis on safety is a high priority.38 A public commitment to safety will thus help the Department of Defense improve at attracting top STEM talent, including AI/ML talent, increasing its ability to keep the American military ahead.

#### Turn – Testing and Evaluation increases effective innovations – improvements in reliability outweigh any time delays.

Horowitz and Scharre 2021 - Director of the Emerging Capabilities Policy Office in the Office of the Under Secretary of Defense for Policy and  Vice President and Director of Studies at CNAS [Michael, Paul, January 12 2021, “AI and International Stability: Risks and Confidence- Building Measures”, <https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures>, Acc 6/18/22. M.A.]

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

#### Committing to TEVV will improve military AI because reliability is an important part of effective technology

**Christie, 2020 - former Deputy Head of Innovation in NATO’s International Staff**  [Edward, Nov 24 “Artificial Intelligence at NATO: dynamic adoption, responsible use” nato.int/docu/review/articles/2020/11/24/artificial-intelligence-at-nato-dynamic-adoption-responsible-use/index.html, BK]

The Alliance’s success with AI will also depend on new and well-designed principles and practices relating to good governance and responsible use. Certain Allied governments have already made certain public commitments in the area of responsible use, addressing concepts such as lawfulness, responsibility, reliability, and governability, among others. In parallel, Allies have taken part in the Group of Governmental Experts on Lethal Autonomous Weapon Systems under the auspices of the United Nations, leading to the formulation of 11 guiding principles. Importantly, there is a good case for viewing work on adopting AI and work on principles of responsible use as complementary and synergistic. In effect, there are certain essential principles or goals that will underpin and facilitate both engineering good practice, as well as responsible state behaviour. Certain national principles imply a need for specific design requirements. For example, a principle of governability may be linked to technical abilities to detect and avoid unintended consequences, and to disengage or deactivate in case of unintended behaviour. The technical characteristics required to ensure that these and other objectives are met will necessarily be part of the design and testing phases of relevant systems. In turn, the relevant engineering work will be an opportunity to refine understanding, leading to more granular and more mature principles. Further work in the area of Testing, Evaluating, Verifying and Validating (TEVV) will be essential, as will support from relevant Modelling and Simulation efforts. NATO’s well-established strengths in the area of standardization will help frame these lines of effort, while also ensuring interoperability between Allied forces.

#### NATO can still maintain technology innovation superiority even if they adopt AI safety measures

Trabucco and Stanley-Lockman, 2022 – prof of Political Science, University of Copenhagen and prof of Defense and Strategic Studies, Nanyang Technological University [Lena and Zoe, The Oxford Handbook of AI Governance, March, “NATO’s Role in Responsible AI Governance in Military Affairs” https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780197579329.001.0001/oxfordhb-9780197579329-e-69 Acc 4/16/22 TA]

In any discussion of AI as an emerging military technology, it is necessary to strike a balance between acknowledging the transformative potential of AI in the security environment, while simultaneously recognizing the “hype” that may, thus far, be unfounded. But some conclusions are clear. The risks and opportunities of military AI can pose significant challenges for future military operations, and this necessarily means there are many stakeholders with a vested interest in developing, promoting, and implementing responsible military AI. As multinational coalitions and military operations are a foundational security policy for much of the world, this means NATO is also a stakeholder with a vital interest in promoting safe and secure technology among its partners, both traditional and non-traditional. As the international security environment continues to shift, there is space for NATO to pursue its agenda to maintain technological superiority not just to protect and defend its way of life, but also to build on its pillars of AI governance to steward military innovation on a responsible trajectory.

#### Case outweighs – the instability and war escalation caused by Automated systems is much more harmful than being more accurate during war.

Amoroso and Tamburrini, 2021 - Prof of International Law at the University of Cagliari and Prof of Philosophy of Science and Technology at the University of Naples Federico [Daniele and Guglielmo, Feb Italian Journal of International Affairs “In Search of the ‘Human Element’: International Debates on Regulating Autonomous Weapons Systems” <https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/03932729.2020.1864995> TM]

While these arguments, involving human duties and fundamental rights, are deontological in character, others are consequentialist since they focus on expected consequences of AWS use. The main normative argument for autonomy in weapons systems is indeed consequentialist: should AWS one day ensure better ethical performances than human soldiers in terms of reduced IHL breaches in the battlefield, then permitting their deployment ought to be preferred to banning it (US 2018a; Russian Federation 2019, para. 2). Yet, this is also contested in the light of a broader consequentialist appraisal, which considers expected repercussions outside the battlefield. The spread of AWS is likely to have negative consequences for international security and peace that outweigh any military or humanitarian advantages one can envisage on the battlefield (Bode and Huelss 2018). Destabilisation risks would include a new arms race among state actors; fewer disincentives to start wars on account of reduced numbers of soldiers involved; cyber-vulnerabilities leading to unintended conflicts; and an accelerated pace of war beyond human reactive abilities (Tamburrini 2016, 127-8).

#### Plan does not hurt the US military – we are only following international standards that we have already agreed upon.

**Horowitz and Kahn, 2021 – Senior and Research Fellows at the Council on Foreign Relations** [Michael and Lauren, The Washington Quarterly 3-19-2021 “Leading in Artificial Intelligence through Confidence Building Measures”  [https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/0163660X.2021.2018794 acc on 6-21-2022](%20https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/0163660X.2021.2018794%20acc%20on%206-21-2022%20) TM]

One might argue that US leadership in multilateral AI standards could create risks for the United States if it involves commitments that prevent the US deployment and use of militarily important AI-enabled systems. These risks, however, are minimal. The standards the United States would promote involve commitments to international humanitarian law and responsible behavior that the United States already follows when it comes to the development and use of military systems. Thus, it would not require changes in US behavior that might slow down responsible military AI adoption.

### AT AIA CP

#### Autonomous Incidents Agreements fail – Incentives for breaking and untruthfulness outweigh agreement

Horowitz and Scharre 2021 - Director of the Emerging Capabilities Policy Office in the Office of the Under Secretary of Defense for Policy and  Vice President and Director of Studies at CNAS [Michael, Paul, January 12 2021, “AI and International Stability: Risks and Confidence- Building Measures”, <https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures>, Acc 6/18/22. M.A.]

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](https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures#fn68) 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.

#### Fast adoption of AI by the US leaves other allies behind, causing resentment and undermining interoperability.

Dufour 2018 - Colonel in the Canadian Army, currently working with NATO [Martin, NDC Policy Brief No. 6 December “Will artificial intelligence challenge NATO interoperability?” https://www.ndc.nato.int/news/news.php?icode=1239 Acc. 4/21/22 TA]

The 2017 report Future War NATO argues that the “technologically-driven US military strategy is advancing so fast compared to European allies that, sooner rather than later, all-important NATO military interoperability might well become a thing of the past”.11 This echoed a 2016 report from the Armament Industry European Research Group, which concluded that “a further boost in US defence technology could promote a wider US-Europe gap and the emergence of a two-tier alliance”.12 It was furthermore observed that the USD3.6 billion invested in Third Offset Strategy technologies in 2017, while representing only 5 percent of the overall US military research and development budget corresponded “to more than 40 percent of the overall EU-European R&D budgets”.13 Beyond interoperability, a profound technological mismatch between allies could also erode the political and military cohesion of the Alliance by creating resentment between countries that bear a greater share of the burden as a result, and those that do not.

### AT Ban LAWS CP

#### There are many non-lethal applications of AI for the Military.

Horowitz, 2018 – Professor of Political Science at UPen**n** [Michael, September “The Algorithms of August: The AI arms race won't be like previous competitions, and both the United States and China could be left in the dust,” https://foreignpolicy.com/2018/09/12/will-the-united-states-lose-the-artificial-intelligence-arms-race/ 6/18/22 MD]

Most military applications of AI will be a far cry from the killer robots depicted in Hollywood films. For example, computer-run algorithms could aid militaries in better designing recruiting campaigns, more effectively training personnel, cutting labor costs through better logistical planning and operations, and improving surveillance. Or consider image recognition algorithms, which have a range of applications--from tailoring ads in the commercial sector to monitoring disputed territory. The race to develop such applications will be crowded for several reasons. For developed countries that have plenty of capital to invest but face challenges in recruiting and retaining talented personnel for their [armed forces](https://go-gale-com.proxy.lib.umich.edu/ps/i.do?p=AONE&u=umuser&id=GALE%7CA556838648&v=2.1&it=r), as well as autocracies that do not trust their populations, there will be an especially strong incentive to leverage AI for their militaries. Doing so will allow them to replace personnel with [automation](https://go-gale-com.proxy.lib.umich.edu/ps/i.do?p=AONE&u=umuser&id=GALE%7CA556838648&v=2.1&it=r) whenever possible.

### AT Ban AI CP

#### The spread of military AI is inevitable – setting standards for reliability is key to avoiding disaster.

Horowitz, 2018 – Professor of Political Science at UPen**n** [Michael, September “The Algorithms of August: The AI arms race won't be like previous competitions, and both the United States and China could be left in the dust,” https://foreignpolicy.com/2018/09/12/will-the-united-states-lose-the-artificial-intelligence-arms-race/ 6/18/22 MD]

GIVEN AI'S MANY POTENTIAL MILITARY USES, policymakers need to rethink the idea of an AI arms race and what it will mean for international [politics](https://go-gale-com.proxy.lib.umich.edu/ps/i.do?p=AONE&u=umuser&id=GALE%7CA556838648&v=2.1&it=r). The fundamental dilemma facing most attempts at arms control is that the more useful a technology is at providing [armies](https://go-gale-com.proxy.lib.umich.edu/ps/i.do?p=AONE&u=umuser&id=GALE%7CA556838648&v=2.1&it=r) with an edge, the harder it is to effectively regulate. There is, after all, no arms control agreement that meaningfully restricts countries from developing tanks, submarines, or fighter jets. Effective agreements tend to restrict the use of less important weapons that don't decide [wars](https://go-gale-com.proxy.lib.umich.edu/ps/i.do?p=AONE&u=umuser&id=GALE%7CA556838648&v=2.1&it=r)--such as landmines and blinding lasers--or ones that have rarely been used, such as nuclear weapons. [Military history](https://go-gale-com.proxy.lib.umich.edu/ps/i.do?p=AONE&u=umuser&id=GALE%7CA556838648&v=2.1&it=r) suggests that those applications of AI with the greatest relevance for fighting and winning wars will also be the hardest to regulate, since states will have an interest in investing in them. Countries with advanced AI companies will be able to leverage those businesses to provide them with some military capabilities, either through adapting commercial technology or by offering financial incentives for talented researchers to focus on defense applications of AI. In these areas, the competition will be fierce because many actors could develop similar algorithms. Some AI applications, such as operational planning for a complex battlefield or algorithms designed to coordinate swarms of planes or boats trying to attack an enemy target, may appeal exclusively to militaries (though even swarms have nonmilitary applications, including firefighting). But even though there are clear military applications, AI cannot be compared to nuclear or [biological weapons](https://go-gale-com.proxy.lib.umich.edu/ps/i.do?p=AONE&u=umuser&id=GALE%7CA556838648&v=2.1&it=r) or even military mainstays such as tanks. AI is not itself a weapon. Just as there was not an arms control regime for combustion engines or electricity, it's hard to imagine an effective regime for containing the coming AI arms race. Mitigating the military risks involved should therefore focus on specific potential uses of AI, rather than the broad technology category. For example, the Convention on Certain Conventional Weapons, which focuses on weapons systems with the potential to cause excessive or indiscriminate injury, is convening discussions in Geneva with countries around the world about lethal autonomous weapons systems. It is critical, however, not to let the specter of killer robots obscure the broader ways AI could reshape militaries, just as the general purpose technologies of previous centuries did. Rapid technological advances that outpace the ability of governments to administer them, fear of falling behind other countries, and uncertainty about the range of what is possible magnify the challenge of effectively regulating AI. Moreover, given the potential for economic investments in AI to spill over into potential military applications, many countries beyond the United States, China, and Russia may balk at regulatory approaches that limit their ability to develop more effective defense forces. Still, governments can create norms and practices surrounding the use of AI both inside and outside of militaries. Setting standards for AI reliability is one possibility, just like international standard setting in other arenas, such as Wi-Fi. Focusing on AI safety is a promising avenue to help ensure that, whatever forms of AI a defense team chooses to adopt, those applications work as intended. Predicting how AI will impact the future of warfare is difficult; it means assessing technologies that are still mostly immature. Much could change in just a few years. If a decade from now current predictions about the military uses of AI turn out to have been more fiction than fact, one reason might be because militaries failed to design algorithms that were robust and secure enough to withstand efforts by sophisticated adversaries to deceive and distort them.

#### Military AI is inevitable – you cannot stop it – you can only make it safer with CBMs

Horowitz and Scharre 2021 - Director of the Emerging Capabilities Policy Office in the Office of the Under Secretary of Defense for Policy and  Vice President and Director of Studies at CNAS [Michael, Paul, January 12 2021, “AI and International Stability: Risks and Confidence- Building Measures”, <https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures>, Acc 6/18/22. M.A.]

Military use of AI poses several risks, including the 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. 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.

### AT Security K

#### The Aff is not based on Realist assumptions – our call for global norms and limits on arms reject realist assumptions. We rely on Utility Theory to explain the status quo

Sosanya, 2022 - AI researcher and a policy analyst at the Day One Project [Andrew, Jan 3, Peace Review A Journal of Social Justice “Autonomous Weapons Are Here to Stay” <https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/10402659.2021.1998856> TM]

To understand why states ban weapons, we must first examine the dynamics of conflict and cooperation. The international system composed of states is commonly viewed through two lenses: realist and constructivist. From a realist stance, states exist in an anarchic world, and their actions are solely driven by security concerns, power, and material interests. While cooperation may occur through international institutions such as the United Nations (UN), realists believe that arms control is merely a tool of statecraft that solely reflects the distribution of power among states. To a realist, states agree to ban weapons only if they have little to no military utility. On the other hand, constructivists view the world as a product of social interactions within a web of different actors that gradually builds up preferential outcomes and behaviors. A central tenet of constructivism is the idea of norms that bind standards of behaviors influencing what these actors believe they “ought” to do, rather than making the most self-serving choice. Thus, states are conditioned to take a collaborative, humanitarian approach to arms control and ban weapons that their leaders and citizens believe to be inhumane, in which there exists a stigma or moral abhorrence typically associated with the use of the weapon itself. While both constructivism and realism have their merits in explaining the dynamics of the world, they cannot individually capture why weapons get banned. Realism cannot adequately explain why states ban weapons at all, nor can constructivism explain why states still develop and deploy certain weapons that are considered to be inhumane, even amongst those who use them. To predict what path of prohibition states will take with autonomous weapons, I introduce a framework that strikes the optimal balance between the two schools of thought: utility theory. Utility theory posits that states cannot cooperate to ban a class of weapons, inhumane or not, if they are considered by states to be decisive, meaning that they are crucial to winning a war. If those weapons are seen as indecisive and inhumane, however, states can cooperate on prohibition because the weapons’ stigma outweighs their usefulness. Utility theory predicts that autonomous weapons will not be banned because key stakeholders—state decision makers—are optimistic about their tactical and strategic benefits. These stakeholders consist of military leaders, political leaders, and others in the defense industry, and they have invested heavily in the development of autonomous weapons and auxiliary technologies, showing no sign of stopping anytime soon. With autonomous weapons, states can project more power with more efficiency and speed than states without them.