JUNE 2025

Artificial Intelligence and War

How the Department of Defense Can Lead Responsibly

AUTHOR
Carol Kuntz

A Report of the CSIS Strategic Technologies Program

CSIS | CENTER FOR STRATEGIC & INTERNATIONAL STUDIES

JUNE 2025

Artificial Intelligence and War

How the Department of Defense Can Lead Responsibly

AUTHOR
Carol Kuntz

A Report of the CSIS Strategic Technologies Program

CSIS | CENTER FOR STRATEGIC & INTERNATIONAL STUDIES

About CSIS

The Center for Strategic and International Studies (CSIS) is a bipartisan, nonprofit policy research organization dedicated to advancing practical ideas to address the world's greatest challenges.

Thomas J. Pritzker was named chairman of the CSIS Board of Trustees in 2015, succeeding former U.S. senator Sam Nunn (D-GA). Founded in 1962, CSIS is led by John J. Hamre, who has served as president and chief executive officer since 2000.

CSIS's purpose is to define the future of national security. We are guided by a distinct set of values—nonpartisanship, independent thought, innovative thinking, cross-disciplinary scholarship, integrity and professionalism, and talent development. CSIS's values work in concert toward the goal of making real-world impact.

CSIS scholars bring their policy expertise, judgment, and robust networks to their research, analysis, and recommendations. We organize conferences, publish, lecture, and make media appearances that aim to increase the knowledge, awareness, and salience of policy issues with relevant stakeholders and the interested public.

CSIS has impact when our research helps to inform the decisionmaking of key policymakers and the thinking of key influencers. We work toward a vision of a safer and more prosperous world.

CSIS does not take specific policy positions; accordingly, all views expressed herein should be understood to be solely those of the author(s).

Center for Strategic & International Studies
1616 Rhode Island Avenue, NW
Washington, DC 20036
202-887-0200 | www.csis.org

Acknowledgments

This report is dedicated to the members of the Ukrainian military with whom I visited during a trip to Kyiv in 1992 as part of an official delegation of the U.S. Department of Defense. The trip was in the first year of Ukraine's independence from the Soviet Union.

At a reception at the Ministry of Defense in Kyiv, a military officer held out to me a sheaf of heavily creased papers—all dirty after much handling—and asked me which one was my favorite. The whole group spoke in heavily accented English so I could only understand a portion of what they were saying. I soon figured out that the candidate favorites were no. 51 about checks and balances and no. 78 about an independent judiciary.

Embers of memory from my liberal arts education flickered and I soon realized that they were asking me which one of The Federalist Papers was my favorite. The sheaf of papers was what had been called samizdat in the former Soviet Union—a single copy of forbidden literature that was passed secretly around a group of dissidents in an authoritarian regime.

I responded that I really thought checks and balances were important; otherwise, the president would have too much power. The military officer holding the sheaf of papers consulted it and then held it, carefully, like a precious object.

As the reception ended, I admitted I had not thought about these issues in recent years. "Maybe you take it for granted," the officer observed. "Yes." I said, realizing that I had never really considered which Federalist paper was my favorite and that I had never seen such an animated, joyful discussion about democracy. "Maybe I do," I replied.

I am grateful to these Ukrainian counterparts and all my colleagues during my service in the U.S. civil service. I am grateful, too, for the legal protections of the civil service and the intellectual rigor they allowed. I was assured on several occasions over the course of my more than 30-year career in government service that I would have been fired without them.

I also am grateful for the opportunity to study and teach at great universities. I left the Bush-Cheney White House to complete my dissertation at MIT, encountering no notable problems because of my realist views or my service at the upper reaches of the national security community. I teach on the historical bias in society as revealed in artificial intelligence algorithms on mortgage rates or policing without any constraint from overseers of various sorts.

For this report on artificial intelligence (AI), I am grateful to the Smith Richardson Foundation for the support that made it possible, and particularly to Marin Strmecki and Chris Griffin.

CSIS has provided a wonderful home for this report, and I am particularly grateful to Jim Lewis, Matt Pearl, and Julia Brock. It was a pleasure to work with the Publications team to polish the draft.

Colleagues at Lawrence Livermore National Laboratory shared many insights and much technical wisdom with me. I particularly appreciate their unusual ability to pair deep technical knowledge of AI with an analytical understanding of war and its vagaries.

Anne Witkowsky provided characteristically insightful comments on the draft. I am grateful to her and to Joan Rohlfing for their help and friendship over many years.

I appreciate the help of other friends and colleagues but acknowledge that any remaining errors are my responsibility.

Contents

Executive Summary

The technology of artificial intelligence (AI) is remarkable, while sometimes being remarkably fragile. AI principally has been applied in a small number of substantive domains, and its implications for other domains–particularly war–remain poorly understood despite being incredibly consequential. The Department of Defense (DOD) does not appear to be developing the analytical tools needed to assess systematically, much less predict, the accuracy of AI-enabled military capabilities; neither is it methodically developing the operational concepts needed to effectively incorporate AI into war. This need, however, is urgent: The United States is confronting the rise of a peer competitor, as well as a host of other military dangers and problems. China fields precision guided munitions, hypersonic missiles, and fighter aircraft increasingly able to pierce U.S. air superiority, capabilities that pose risks to classic U.S. power projection forces. There is a reasonable basis to believe that AI-enabled military capabilities could help rectify many deficiencies in U.S. combat power. This outcome assumes, however, that the effects of the AI algorithms in sensitive uses could be confidently anticipated, an assumption that the DOD cannot currently make. This paper demonstrates the importance of the DOD developing analytical tools to inform decisions about the use of algorithms, particularly in the context of sensitive applications like sensor and firing networks. It then provides recommendations on how the DOD should proceed to develop such analytical tools and the operational concepts to effectively employ and defend against AI-enabled military capabilities.

Context

"With the correctness of their science, history had no right to meddle, since their science now lay in a plane where scarcely one or two hundred minds in the world could follow its mathematical processes; but bombs educate vigorously . . ."[1]

—Henry Adams, 1918

Artificial intelligence is a remarkable technology, but it is vulnerable to errors, even in domains where it has long been used. The prospect of the use of AI in war raises acute concerns about the occurrence of errors and their consequences. Absent the ability to predict the effects of AI-enabled military capabilities, the United States could neither confidently nor responsibly rely on the deterrence or warfighting benefits of these capabilities.

Yet these benefits are urgently needed. The U.S. ability to deter war in regions of critical interest is fading.[2] Adversaries—particularly China—have improved their own capabilities and could now threaten classic U.S. power projection.

The DOD should develop analytical tools to measure, manage, and predict the accuracy of algorithms. Many individuals—from both inside and outside of the DOD—have called in general terms for such techniques to be developed.[3]

AI-enabled military capabilities, particularly sensor and firing networks, combined with other force and program improvements, would be among the most promising strategies for shoring up deterrence and defense capabilities in the near- and mid-term.[4]

Presently, the DOD does not appear to be developing a systematic set of analytical tools for this purpose. There is no indication that there are department-wide metrics that could inform important decisions at any of the necessary analytical levels—in the field, in the command center and the Situation Room, or in wargames. Individual programs sometimes have excellent program-specific metrics, but this isolated data cannot solve broader problems.[5]

The department has thus far framed its guidance for the development and use of AI-enabled military capabilities in a general fashion, calling for capabilities that are "responsible, equitable, traceable, reliable, and governable."[6] These general qualities, while worthy, are far from enough. Predictions about the effects of AI-enabled capabilities, particularly for sensitive uses, would need a high degree of confidence to inform decisions in the command center or the Situation Room regarding military effectiveness and law of war compliance.

Artificial intelligence could power meaningful improvements in warfighting capabilities, assuming its effects could be accurately predicted. The DOD, though, cannot make this assumption today.

Catalyzing military innovation in peacetime is very difficult. The vigorous education provided by war swiftly provides lessons about effective ways to incorporate emerging technologies. But these lessons are too often provided by the adversary, and war is invariably a harsh teacher—meting out insights with a clarity that yields an education, but one tinged with regret and loss.

Summary

Successful peacetime innovation would need to resolve interlocking puzzles confronting the United States at this strategic and technological moment:

1. AI has thus far been vulnerable to errors in the domains where it is used. These errors are likely to appear—perhaps more often and with more consequence—when it is transplanted into a new substantive domain such as war. DOD needs analytical tools to assure accuracy of uses in the field. There has been limited work, both generally and within the DOD specifically, to systematically identify, understand, and mitigate these errors.
2. If AI is to be used in war, decisions in the command center or the Situation Room would need to rely on predictions about the effects of the use of algorithms, particularly in sensitive applications. Absent such predictive tools, AI-enabled military capabilities could not be responsibly authorized in uses such as a sensor and firing network (discussed further below). DOD efforts are not on a trajectory to develop the analytical tools required to predict the effects of sensitive combat decisions made by AI, either in terms of military effectiveness or compliance with the law of war.
3. Changes in the threats confronting the United States—including the emergence of China as a peer competitor—mean that classic U.S. defense capabilities including power projection are at risk. The fielding of AI-enabled capabilities, including in sensor and firing networks, could be a significant contribution to efforts to manage and reduce emerging threats. Adversaries have incorporated new technologies and could complicate, if not defeat, U.S. efforts in some key theaters.

4. Historically, new technologies only win wars if they are incorporated into new operational concepts. The DOD does not have a methodical effort underway to develop, vet, and implement new operational concepts using emerging technologies. Identifying and developing effective operational concepts is conceptually difficult, relying as it does on an accurate idea about the nature of future war. Implementing promising new operational concepts is bureaucratically difficult, given their inevitable effect of reducing the importance of some existing weapons systems and their associated professional communities.
5. The DOD needs to make significant reforms to strengthen its ability to effectively incorporate emerging technologies. Making such reforms would be difficult in the best of times, and these are not the best of times. It is, however, urgent to strengthen deterrence and defense capabilities, including through responsibly incorporating AI-enabled technologies and building defenses against an adversary's use of them.[7]
6. Realists, liberals, and humanitarians are the traditional parties in the argument about the cause and effect of war. Realists want the warfighting advantages of precision and speed putatively offered by AI. Liberals and humanitarians are skeptical about the claims of precision and troubled by the prospect of speed. Liberals fear that speed increases the risk of unwanted escalation; humanitarians fear that autonomous firing decisions will have speed but lack precision—much less common decency—and hence oppose such firing decisions. The speed celebrated by realists, though, is largely acquired through the autonomous targeting and firing decisions. Analytical tools for accuracy and robustness are necessary to responsibly reconcile the key insights of these views, even if they cannot fully resolve their differences.

This report focuses on AI incorporated into conventional warfighting capabilities. It does not engage directly with the implications of artificial intelligence incorporated into any element of the use of nuclear weapons. It is a reasonable assumption, though, that the report's many concerns about the risk of mistakes for AI-enabled conventional weapons would apply with even greater consequence in AI-enabled decision processes about nuclear weapons.

Growing Vulnerability

U.S. conventional forces in the post-Cold War era were configured on the idea that the United States could surge power projection forces swiftly enough to supplement forward presence forces. Working with allies and friends, these combined forces would be able to deter and, if necessary, defeat conventional military threats in regions of critical interest abroad.[8]

As the Cold War receded further into history, this basic idea of using forward presence and power projection to deter or defeat regional adversaries persisted in U.S. defense planning. The changes in the ensuing years constituted adjustments to the same basic strategy, through different combinations of adversary major and minor regional contingencies. But over time, this idea has reflected less and less the strategic realities the United States actually confronts.

The United States' rivals today are vastly better than they were even a decade ago at attacking the type of forward presence and power projection forces that have been the hallmark of U.S.

conventional capabilities.[9] This is largely because adversaries—particularly China—have grown the size of their militaries, created a more favorable deployment pattern for these forces, and effectively incorporated advanced technologies. Adversaries have developed and fielded long-range precision guided missiles (PGMs) coupled with sensors that could direct these missiles effectively, as well as capable fleets of military aircraft, both fighters and bombers. China's military in particular is much larger and more capable than it was 13 years ago.[10]

Any large aggregation of U.S. military assets would be at risk from these capabilities. Large force aggregations include aircraft carriers, ports of debarkation and embarkation, military camps and airfields, and large ground force units, all key components of forward presence and power projection.

The U.S. military is developing and deploying countermeasures to these precision attacks, and doubtless some of these countermeasures will be effective. While the details matter in each dyad between a weapon and its target, there is no compelling solution to the observation that these U.S. forces would be vastly more vulnerable in the future than they have been in the past. The months-long deployment to the Persian Gulf that the United States pursued in 1991, for example, would no longer be possible today if competitors wanted to threaten it.

Adversaries also tend to be located closer than the United States to the contested theaters. This difference complicates for the United States the challenge presented by geography to power projection forces and their resupply.

In such situations, U.S. power projection forces would be vulnerable to attacks of various sorts en route to and upon arrival in theater; it is even increasingly likely that U.S. forces might arrive only after the conclusion of decisive battles, largely because of the harsh realities of distance and the improvements in adversary capabilities.[11] U.S. forces could then only lament in person the changed strategic reality on the ground and the unhappy choices now confronting the United States and its friends and allies.[12]

There are multiple changes that need to be made to make the power projection element of this strategy plausible again. Those changes should incorporate a large sensor and firing network that could detect an adversary ("red") attack on critical interests abroad and then allocate and fire U.S. or allied ("blue") weapons against those attacking forces. Much larger stockpiles of precision guided munitions would be needed, along with other reforms to U.S. forces, both forward deployed and reinforcing.

This type of broad strategy—destroying a red attack against blue allies through the rapid and integrated firing of precision-guided munitions to halt the initial invasion—is an important part of the most promising available strategy in the near- and mid-term. It would require the creation of an effective and very large—likely theater-wide—sensor and firing network. In at least some types of crises, this network would need to have a certain degree of autonomy.

These changes could help manage the greater pressure on U.S power projection forces. A sensing and firing network seems necessary, along with other force adjustments and the efforts of friends and allies, to deter—or, if necessary, defend—Taiwan from an attack from China, the eastern border

of NATO from an attack by Russia, or other missions in distant theaters. Absent these capabilities and other improvements, respected defense analysts warn that Chinese or Russian attacks could be successful.[13] It is this growing vulnerability that provides the urgency to swiftly and effectively incorporate AI.

This report acknowledges that there is a great debate within the United States today over whether there should be a fundamental redefinition of its strategic interests. The substantive resolution of this debate is difficult to predict. Whatever the outcome, the complicated issues of AI-enabled military capabilities, particularly sensitive uses like a sensor and firing network, are not going anywhere.

For example, if the United States decided that only the security of its territorial borders was critical, the risk of adversary precision guided munitions of various sorts attacking the homeland would remain. A defensive effort—whatever its precise characteristics or configuration—should involve AI-enabled military capabilities, particularly an AI-enabled sensor and firing network constructed to move swiftly to identify and destroy the incoming missiles and other weapons.[14]

Crucible For Reform

This report agrees with the assessment that broader reforms are needed in the Pentagon to strengthen its ability to incorporate emerging technology.[15] There is not, however, time to wait until the DOD's internal structure and processes are fixed. Comprehensive efforts to incorporate AI responsibly need to move forward more quickly.

To enable that more rapid progress, this report recommends that the secretary of defense empower select senior warfighters to develop operational concepts for the use of AI-enabled military capabilities, largely freed from many of the bureaucratic constraints hobbling current progress.

These senior warfighters would be a select group of the combatant commanders including at least those from Special Operations Command, Indo-Pacific Command, and European Command. They would get new authorities for funding, staffing, and reinterpreting regulations for these capabilities. They would have the funding and acquisition authority for AI, related communications, and small platforms, including drones. This funding authority would create a much more dynamic ability to select solutions from different suppliers, including the Defense Innovation Unit, tech startups, the military services, and traditional defense contractors. The new operational concepts and capabilities would be thoroughly tested in well-instrumented wargames. At the end of a two- to four-year period of experimentation, the secretary could decide whether to let the select combatant commanders retain their new funding and acquisition authority or to relocate that authority.

This report also recommends that the secretary create a competitive dynamic for the first few years on the development of analytical tools, directing the select combatant commanders and a reorganized Chief Digital and Artificial Intelligence Office (CDAO) to each develop tools. After a few years of competitive effort, the secretary should select the suite of tools most useful for the purpose and consolidate the approaches used for analysis within the DOD.

The CDAO should be supplemented with expertise in warfighting and analysis or a new organization with analytical expertise in both warfighting and AI should be created.

Finally, the secretary should ensure that JASON, the advisory group consisting of top national scientific experts who advise the national security community on sensitive technical questions, has elite academic experts on artificial intelligence.[16] This perspective is vitally important given the remarkable and unexpected advances that AI technology continues to make. The advice of these outside experts is often revelatory and influential, although it is rarely celebrated in its initial presentation.

This proposal is developed further in Appendix II.

Sensing and Firing Networks

War has a small number of recurring objectives. One of these is locating a red target and destroying it with a blue weapon. A military archer sees an attacking enemy soldier and aims an arrow. A single drone uses an image recognition algorithm, identifies a red target with required accuracy, and then detonates the blue payload of explosives an optimal distance from the target. These are very different missions, but the fundamental tasks of locating and destroying remain the same.

A "sensor and firing network" as discussed here would be an AI-enabled system to identify and characterize red targets, allocate blue weapons against those targets, and then launch those weapons. The scale envisioned is a battlefield or even a theater-wide engagement. There is no indication in the public literature that there is an existing network at anything approaching this scale and complexity.

Such a network would integrate the findings of blue sensors and information streams evaluating and characterizing hundreds or perhaps thousands of red targets, allocating blue weapons from ground, sea, and air-based assets in theater as well as long-range assets outside of the theater. It would then execute an integrated blue launch decision, either made by a human operator or pre-authorized to occur autonomously under certain conditions.

There are examples of such networks on a smaller, more limited scale. The Israeli Iron Dome system has had remarkable success against limited attacks by shorter-range ballistic missiles through tracking incoming missiles and autonomously launching interceptor missiles to destroy those assessed to be heading toward populated areas. The Iron Dome has been reported to have a success rate of stopping 90 percent of missiles launched against its covered areas. It reportedly has a launch-to-impact speed of 15 seconds.[17]

As the current war in the Middle East grows in duration and combatants, some of the vulnerabilities of the Iron Dome and its related systems may grow in significance.[18] The related systems seek to stop longer-range missiles, ideally further and further away from the protected areas. The per unit cost of the interceptor missiles launched by the Iron Dome is high—with estimates ranging from $60,000 to $40,000 per missile.[19] The cost of adversary missiles or drones being fired at Israel vary widely but often is significantly less. For example, in an exchange, Hezbollah missiles were largely stopped

by Israel's Iron Dome system—a seeming validation of sophisticated high-end defensive systems. But the cost of the Iron Dome defense was 10 times that of the Hezbollah attack.[20] The struggle may become one of capacity: Israel's access to more interceptor missiles against adversaries' capacity to launch yet more missiles at Israel.

The territory covered by the Iron Dome is limited in size and so the adversary does have some incentive to find valuable targets beyond its effective range. Most importantly, at some number and rate of incoming missiles, any defensive system—including the Iron Dome—would become saturated and ineffective.[21] Even with autonomous targeting and firing decisions, the total number of items to manage could overwhelm the capacity of a particular system.

These issues do not invalidate the remarkable contribution that the Iron Dome has made to the security of Israel, nor the theoretical usefulness of sensor and firing networks for defensive and even warfighting purposes. They do, however, underscore again that while specification of the mission is often clear-cut—find the red target and destroy it with a blue weapon—the details matter when planning for sustained, durable success.

Such details would be much more demanding should the attempt be made to build a theater-wide sensing and firing network. Yet such a network could contribute in essential ways to the challenge of strengthening U.S. power projection capabilities. There is both the strategic need for such a network, and plausibly, although not yet demonstrably, the technical ability to perform these functions through the use of AI.

To illustrate the strategic need, recall the Fulda Gap, that piece of territory in Germany that divided NATO from the Warsaw Pact during the coldest days of the Cold War. On each side of the gap, large numbers of tanks, troops, air power, and assorted support capabilities looked at each other across the border. If the Warsaw Pact had decided to launch a conventional attack, any advance would have been slowed, if not stopped, by the substantial NATO forces in being.

But consider a situation in which the Warsaw Pact had forces on the Fulda Gap, or perhaps a discrete distance back, and the West no longer had forces on the other side of the gap. West Germany had some military forces, but these were vastly outnumbered by the Warsaw Pact forces in being. NATO remained a stalwart ally, but each countries' forces were largely stationed on their own territory.

How, then, to stop the Warsaw Pact forces if they decided to drive toward the West through the Fulda Gap? It would be vastly preferable to stop the Warsaw Pact forces before they cross into West Germany, to prevent the civilian deaths and the destruction of property that would result from an invasion and occupation. In addition, while pushing forces out of occupied territory is certainly possible, the prospect of doing so against a nuclear-armed adversary should certainly give pause.

Once the Warsaw Pact forces started moving toward the border there would not be much time to react. Say there were about 100 miles from the Warsaw Pact barracks to the border of West Germany—roughly the distance across the strait from mainland China to Taiwan. It is not plausible

to move NATO heavy forces to the Fulda Gap in the amount of time it would take the Warsaw Pact forces to move that 100 miles into West Germany.

If a sensor and firing network such as described above exists in this thought experiment, the way to stop the Warsaw Pact invasion would be to quickly identify all the Warsaw Pact tanks moving toward the border, allocate blue weapons against each tank, and launch the blue precision guided munitions. Time would be short, though, and if the invasion force is large, an AI network, possibly set on an autonomous mode, might well be the only way to perform the process quickly enough to make a difference.

Such a network as envisioned here would have three stages. In the first phase, the network would identify, vet, and characterize red military targets. In the second phase, blue weapons would be allocated against each red target so as to maximize military effectiveness and ensure compliance with the law of war. The third phase would be the firing of the blue weapons, either by a human operator or by an algorithm consistent with rules previously programmed into it by a human.[22]

Analysts posit that an AI-driven network would have precision and speed. Integrated networks at the size of an Israeli Iron Dome have been effective. Larger networks seem plausible, but the details matter when identifying and destroying a target and ultimately could prove intractable at scale. Whether a responsible battle-wide or theater-wide network could be constructed remains undemonstrated.

This report is focused on the AI elements of such a network. It is worth noting briefly that many other gaps would remain. For example, the United States would need to vastly increase its stockpile of precision guided munitions to successfully execute such a strategy.

The call to develop analytical tools does not assume that the targeting and firing decisions in these networks would be autonomous. Careful analysis must be done in advance, to either support the human decisionmaker or to enable responsible prior authorization of autonomous firing in certain conditions.

Over the course of the wargames and other activities called for here, the DOD should develop new operational concepts that identify additional ideas about how to use AI to the United States' advantage in war. Powerful use cases are emerging that involve tactical deployment of small platforms—particularly drones—in concert with other coordinated platforms operating with at least some autonomy. These uses seem very promising. The DOD should explore these uses, either to use them against an adversary or to defend against their use by an adversary.

The Need for Analytical Tools

The DOD makes decisions in the field, in the command center and the Situation Room, and in wargames. Each of these contexts requires different analytical pathways and would need to have different tools created to enable the responsible use of AI.

Decisions in the field turn on straightforward questions. Soldiers need to be able to confirm that an AI algorithm would be accurate when performing a specific function in a particular context—for instance, that an image recognition algorithm would correctly identify an adversary tank and only an adversary tank in a particular battle on a specific day. A shot needs to hit its intended target: Even the most brave and aggressive warfighters have no desire to die from fratricide if a weapon is launched at them by their own military because of fatal errors in target identification.

Decisions in the command center or the Situation Room require a more robust understanding of the expected effects of a military measure. This requirement is necessary to assess both military effectiveness and compliance with the law of war. For AI, these questions would need to focus on whether an algorithm that has been effective in the past would be accurate in a prospective, not yet fully specified, use on a future battlefield. The current lack of understanding of the robustness of algorithms, particularly algorithms for sensitive uses, means that neither calculation can currently be made responsibly.[23]

Accuracy and robustness in decisionmaking are needed regardless of whether these sensor and firing networks are autonomous or not. Human-authorized autonomous firing would require vastly better data than is currently available. Similarly, a human with active oversight of a sensor and firing network, sometimes called being "in the loop" or "on the loop," would confront great complexity and extraordinary time pressure. Such oversight might not seem particularly meaningful without significantly better analytical tools.[24]

This report urges that any deployment of algorithms for sensitive uses, including in sensing and firing networks, should be authorized in the rules of engagement approved by the secretary of defense before a particular operation to ensure that these issues receive an appropriate degree of oversight. This would require a definition of "sensitive" use cases for algorithms, and recognition of the different requirements for accuracy and robustness.

Not all algorithms used in military capabilities would need these high levels of accuracy and robustness. Algorithms that identify candidate items for further review could have less stringent requirements for accuracy because of the prospect of human oversight without severe time pressure. For example, there are algorithms that scan millions of hours of video imagery for items that look like improvised explosive devices (IEDs). The candidate items would then be further reviewed, including by intelligence analysts, before a definitive decision that it was an IED and thus an appropriate candidate target.

This report argues that generative AI should not be used for sensitive uses, including a sensor and firing network, at this stage of its development. This is because of generative AI's vulnerability to random errors. Generative AI could be used as an aid to intelligence analysts, as long as the analysts were aware of the risks of error currently inherent in a generative AI algorithm (see the section on generative AI below).

A third set of decisions arises in the context of wargames. Wargames are one of the most promising techniques to develop and vet new operational concepts. Analytical tools for wargames

require properly anticipating and modeling the effects of red equipment, tactics, and strategies. These interactions would be difficult to model particularly for a technology whose effects are poorly understood.

AI raises strategic questions that fall outside of these three contexts as well. For example, how to manage unwanted escalation—a legitimate concern, given that much of AI's military advantage is in the posited speed of its integrated networks. There also are ethical questions surrounding any decision to permit AI algorithms to make targeting and launch decisions, particularly those that could injure an active enemy combatant or, more perilously, a noncombatant.[25] These questions have emerged in the ongoing debate among the different perspectives on war.

Realists, Liberals, and Humanitarians

Realists, liberals, and humanitarians are the traditional protagonists in the effort to explain cause and effect in international relations. These three groups rarely agree, but artificial intelligence—at its broadest level—has achieved a rare consensus: that incorporating AI into war will be transformative. The protagonists disagree, though, on the nature of that transformation.

Realists emphasize the promise of AI in effective deterrence, decisive warfighting capabilities, and even defense dominance, which would vastly reduce the risk of war against status-quo powers.[26] There is tantalizing evidence from Ukraine and Russia that these predictions are likely to be accurate.[27] Realists concede that AI algorithms could make mistakes, but they judge that, on balance, these problems could be managed and reduced to an acceptable level.[28]

Liberals warn of algorithmic mistakes and unintended escalation leading to unwanted wars.[29] Unwanted escalation occurs when neither nation-state really wants war, but both fear losing the war if the other side attacks first. Not wanting war, a nation-state may nevertheless choose to start one because it wants defeat even less. It fears that the other nation-state is likely to attack and judges that with extant military technologies at play, an attack is likely to be successful.

Humanitarians underscore the centrality of human life above all. Humanitarians fear that whatever the sophistication of the algorithms, unacceptable mistakes are inevitable because of the importance of understanding context in use of force decisions. Algorithms may learn to accurately judge what criteria make an actor a legitimate military target, but might be unable to grapple with complexity, for example, if the actor is a child.[30] This group remains troubled by the prospect of any autonomous firing decisions but especially when the target is a human being, and urge constraints on the use of autonomous weapons systems (AWS) for any purpose.[31]

The law of war, part of a broader body of international humanitarian law, seeks to assure that war is conducted with distinction, proportionality, and humanity.[32] The United States military is required to maintain consistency with the law of war.[33]

Humanitarians have been seeking the adoption of additional components of the law of war to prohibit the use of AWS to target humans.[34] These components would require "meaningful" or "effective" human control, generally defined as human involvement in the selection of the specific

target (as opposed to a target being selected by an algorithm) and in the making of a firing decision (as opposed to preapproval by a commander based on some prespecified criteria).

There is an alternative perspective that the classic requirements for distinction, proportionality, and humanity under the law of war could lead to responsible use of AI-enabled military capabilities. Liberals and humanitarians doubt, not unreasonably, that such responsible application would in fact occur, and there is too much recent evidence in favor of their position to reject these doubts out of hand.

This report concedes the vulnerability to mistakes and thus urges the development of analytical tools to minimize and manage these risks. AI-enabled military capabilities, to include algorithms to identify targets in a firing network, should not be used until such tools are developed and implemented.

This report joins many voices in the debate, including that of the U.S government, to argue that the additional prohibitions on lethal autonomous weapons (LAWs) proposed by humanitarians should not be adopted.[35] The "meaningful" or "effective" human control requirement would eliminate much of the warfighting benefit from a posited responsible autonomous sensor and firing network, since it would slow the process considerably and perhaps enable adversary efforts to move key assets out of the effective reach of U.S. capabilities or even use them first against U.S. forces. The problematic aspects of AI-enabled weapons should be governed by the classic law of war—distinction, proportionality, and humanity.[36]

But as with any discussion between the realists and humanitarians, there is the great difficulty of reconciling the perspective on human life at home with the perspective on the battlefield. Wars occur, and realists remind the nation-state that wars need to be deterred and, when necessary, won. Humanitarians draw attention to the enduring perplexity of conducting wars to protect a people, through the device of killing another people.

An AI-enabled sensing and firing network, be it large or small, would crystallize the differences among realists, liberals, and humanitarians. The risk of technical mistakes is real in every stage of a sensor and firing network but is particularly acute in the first phase, in which the algorithm is supposed to identify legitimate military targets and only legitimate military targets consistent with military requirements and distinction.

Stage two is when the blue weapons are allocated against each red target as determined in the first stage. That allocation should minimize collateral damage to ensure that proportionality and humanity requirements are met. Stage three is the firing decision made either by a human operator or by an algorithm. The challenge in each stage increases proportionally as the scale of the enterprise—the number of red targets and blue weapons—increases.

Stages two and three will be vulnerable to escalation risk because there will be military advantage to acting quickly. If war is inevitable, acting quickly to secure the resulting benefits makes sense. If there is still diplomatic hope that war could be avoided, though, acting quickly because of escalation pressures could turn a diplomatic crisis into a war wanted by neither side.

Stage three is where the decisions about autonomous firing decisions would play out in one of three ways: as the result of (1) real-time human control, (2) earlier considered human control, or (3) carelessness, the absence of any type of meaningful control.

Overall, technical issues of accuracy and robustness are most likely to plague efforts in the first phase, while escalation issues are most likely to plague those in the second and third, and the decisionmaking issues around appropriate firing authority are most likely to plague the third phase. (See Appendix I for further consideration of these issues.)

Realists, liberals, and humanitarians all agree that AI in war will be transformative. Yet the policy recommendations of these groups are irreconcilable, with their calls, respectively, for more and swifter incorporation, limits on incorporation, and a prohibition on all LAWs.

Analytical tools for accuracy and robustness must be developed and incorporated into decisionmaking. These tools and their responsible use would demonstrate respect for these different perspectives, if not their categorical observance. Otherwise, the ongoing responsibilities of the United States in war seem impossible to meet.

Remarkable Yet Fragile: AI Algorithms

The rate of advance in artificial intelligence leaves only one key observation stable: The technology is advancing more rapidly than the policy frameworks to manage its implications.

Many recent technical advances and policy debates have focused on large language models (LLMs), which burst into public notice in November 2022. The current boom in AI really started a decade earlier, in the 2012 ImageNet competition when the winning algorithm had half the errors of the second-place winner, achieving the greater accuracy through the technique of using neural networks, a supplement to an algorithm that significantly increases its accuracy but is voracious in terms of data and power consumption.[37] The theories that power AI stretch back much further than 2012; the formal start of the modern era is generally dated to an AI conference held at Dartmouth in the summer of 1956. Despite the earlier articulation of many of the critical theoretical insights, the field took off in the early twenty-first century largely because two additional, essential innovations became widely available: massive datasets, thanks to the internet, and large amounts of computing power, thanks to the ongoing advances of Moore's Law.[38]

Until very recently, it was true that a larger training database yielded a more accurate algorithm as long as there was no constraint on available computing power.[39] The amount of data needed for an accurate algorithm was usually large but varied significantly depending on the type of algorithm and the complexity of the problem. AI image recognition algorithms, for instance, often used millions of images, while LLMs used billions of pieces of data acquired from millions of webpages.[40]

The first cracks in this verity showed up at the outset of 2025. Advances by the Chinese company DeepSeek seem to have demonstrated a technical strategy to program an LLM with significantly less computing power that nevertheless yielded an algorithm as good as the best U.S. LLM algorithms.[41]

Similarly, the need for ever-larger data sets was undercut by recent indications that in certain contexts, small language models could yield accurate algorithms.[42] These most recent advances demonstrate principally that technical progress in AI remains rapid and unpredictable.

AI is remarkable but can be fragile. Artificial intelligence closely studies a training database and derives an algorithm from it; the algorithm then goes out into the real world to evaluate previously unseen examples of the same phenomena as contained in the database.[43] Depending on the type of AI, the algorithm performs various tasks with remarkable accuracy, including making predictions, identifying correlations, optimizing strategies, or generating text or other materials.

AI's fragility lies in its vulnerability to mistakes, some of which lie hidden in its stores of data or in the interworking of complex algorithms. Most technical mistakes in the use of AI algorithms arise from one of three broad categories: (1) the training database (this report looks particularly at drift and poisoned data), (2) the algorithm type (particularly generative AI, reinforcement learning, or a hybrid), or (3) the interpretation of the algorithm's findings by the user (specifically automation and user bias).[44] This set of mistakes should not be expected to be static; it is likely to change over time as technical advances eliminate some sources of mistakes and create new ones.

Technical mistakes can occur regardless of the domain within which the algorithm is being used. The frequency with which the mistakes occur can vary among domains, as can their effects. Mistakes in the domain of war, of course, can have grave consequences.

Training Database Errors

Training database problems—the first category of errors common to AI—occur when there are differences between the database and the phenomena in the real world that the algorithm is evaluating. Information from the real world is called an input, and the process of evaluating that input with the algorithm is called inference.

This issue of the relationship between the entries in the training database and the inputs in the real world can be seen in an example using a powerful type of machine learning called supervised learning. Supervised learning uses a labeled database. For example, an image recognition algorithm might be trained on a large database with many images of military tanks, military trucks, and civilian school buses. Each image is labeled. The algorithm would study the images and distill for itself rules to correctly label previously unseen images of each of these types of vehicles.

The algorithm could then be used in the real world to exam inputs of images of previously unseen vehicles. The expectation is that the algorithm would correctly identify each new input as either a military tank, a military truck, or a school bus. (The algorithm also should have an "other" category since it is likely to encounter other types of vehicles.) The entries of the training database are ideally selected to be as similar as possible to the expected inputs.

For a variety of reasons, though, the input data may be different or become different from the training database over time, in ways large or small. These differences can cause errors in AI algorithms used for inference in the real world.

In this example, the accuracy of the algorithm's findings could be undercut if, for example, the adversary military started to use civilian trucks or school buses to transport military equipment or personnel.

Drift

The emerging differences between the training database and the input data is known as drift. Drift occurs when the phenomena in the real-world change over time. Eventually, the algorithm will become inaccurate because of the differences between the database and the inputs.

In a simple example, consider efforts to train an image recognition algorithm to identify where adversary troops are making camp during an invasion into a contested region. Intelligence services are confident that the adversary uses the same units and equipment in invasions that it does in exercises.

The training database therefore uses pictures of the region during adversary exercises, where the adversary—say, the Russian army—moves its equipment into the contested region for all to see. The pictures in the training database capture the visual appearance of the equipment and the environment where the equipment is parked. The region is forested, containing many deciduous trees.

The feared invasion occurs. Theoretically, the algorithm should help identify where the invasion forces are located to facilitate accurate blue targeting. New photographs are taken after the invasion to be used as inputs to the algorithm that was trained on earlier adversary exercises in the region.

The images in the database that trained the algorithm, however, differ significantly from the images in the inputs. First, the Russians are using high-quality camouflage to conceal the tanks during the invasion. Second, the pictures for the training database were all taken during the winter, when the trees were without leaves—and the invasion occurs in the spring. These differences between the inputs and the training data will reduce the accuracy of the image recognition algorithm, possibly enough to render it unhelpful to targeting efforts.

That said, it remains the case that some types of patterns could reasonably be expected to repeat themselves: A vehicle that looks like a Russian tank, contains the same percentage of different metals as a Russian tank, and generates the same heat signature as a Russian tank probably is a Russian tank.

The training database for an algorithm to identify Russian tanks could contain copious entries with each of these pieces of data. Algorithms to identify the same type of adversary equipment based on a training database that contained many different types of sensor data about each entry probably would be both accurate and robust.

But how useful is such an algorithm on a battlefield where both friendly and adversary forces use Russian tanks? This potential complication demonstrates the importance of the analytical tools having qualitative in addition to quantitative elements. The qualitative components should aim to make certain characteristics of the algorithm accessible to the military personnel using it, helping them use their judgment to understand whether the intangibles of war render the algorithm unhelpful, even if its math is exquisite.

While AI algorithms are remarkable, there are aspects of the use case and the training database that can be sources of errors. Identifying these aspects will require both technical and military expertise.

STRATEGIES TO MANAGE DRIFT

Over time, generalizations about robustness—the effectiveness of the algorithm for a similar but not identical use in the future—could be distilled. Training databases that contain multimodal sensor readings for the same piece of equipment—heat, visual, metal content, generated effluents—are likely, on balance, to be accurate and robust.

However, it remains to be demonstrated systematically that there will be sufficient alignment to make AI algorithms accurate for a particular use on a particular day, much less that the algorithm will be robust enough to provide confidence that it will be accurate in some future set of related but different, perhaps even unanticipated, use cases. [45]

The commercial sector is one of the domains with the longest track record of using AI. Drift has proven to be consistent problem, and the commercial sector has developed a variety of strategies to manage it. Many businesses, for example, have developed techniques to measure the substantive difference between the training data and the input data. There also are techniques to measure the impact of drift, thereby facilitating an understanding of when retraining is necessary.[46]

Some businesses have standards that specify the rate at which they need to remove the oldest data and supplement the training database with new data to maintain the desired level of accuracy in the algorithm. This gradual replacement of the training data is called adaptive learning.[47] The military should adopt some of these techniques and perhaps develop new ones to identify, measure, and responsibly manage the problems associated with drift.

PATTERNS IN TACTICS OR IN STRATEGY

A fundamental question about war is how much of it is based on patterns. This is AI's original gift: seeing patterns unrecognizable to humans because of their extraordinary complexity.

At the same time, war is filled with unknowns, what Prussian General Carl von Clausewitz described as, "a fog of greater or lesser uncertainty."[48] He goes on to say, "A sensitive and discriminating judgment is called for; a skilled intelligence to scent out the truth."[49] AI, certainly a skilled intelligence of a sort, may well be able to lift some of the fog, but whether it can entirely eliminate it remains to be seen.

Clausewitz also refers to war as "the realm of chance."[50] It is unclear whether what appears to be chance—unseasonable weather, an unexpected technical failing, a line of infantry that held when it was expected to fail—can in the future be predicted by a sufficiently robust AI algorithm.

The presence of a thinking, reacting adversary trying to exploit any overreliance on patterns seems likely to further complicate AI's ability to understand war. Clausewitz pointed particularly to the challenges of rules predicting the course of "a continuous interaction of opposites."[51]

While there are routine tasks at the tactical level such as aiming and firing an artillery piece, there also are more singular tasks such as a commander's decision on when, where, how, and indeed whether to attack.[52] Strategic decisions seem difficult to predict, particularly because the training database seems unlikely to be large enough or representative enough of the personality and experience of the relevant commander. Generals who share formal training often pursue very different strategies in war.

Many wars, or at least battles, have been won by the unexpected decision of a strategic commander: leaving a large wooden horse just outside the adversary's well-secured city, or exploiting unsecured cockpits of large commercial jet liners traversing major urban areas. Could an algorithm anticipate the first uses of these stratagems, or move beyond pattern recognition to the type of creativity that wins wars?

This report recognizes that the technology is advancing rapidly and that its abilities and their limitations in the future will be different than any interim assessment issued now. It is reasonable, though, to question what type of military decisions could be managed by artificial intelligence in the near- and mid-term. AI, with its persistent reliance on databases of previously acquired data, may be well suited to some types of military decisions (e.g., a firing decision) and poorly suited to others (e.g., high-level military tactics).

Poisoned Data

A further type of problem with training databases or inputs is the use of poisoned data by an adversary. Poisoned data is manipulated to exploit aspects of the algorithm to cause it to yield inaccurate results. This can arise either in the training data or in the input data. Militaries should expect and be prepared to defend against poisoned data attacks.

An example might be that an adversary has a missile that it programs to always follow a certain evasive maneuver: At a particular distance from the target, it always feints to the right before homing in on its mark. The data from these observed flights would be analyzed by the United States, who would almost certainly identify the consistent feinting to the right before impact. A resulting blue targeting algorithm for the red missiles would likely predict the feint to the right and incorporate that maneuver into its steps to locate the red missile in flight for targeting.

Immediately prior to the start of the invasion, however, red programmers could alter the pattern of the missiles to feint to the left instead. The blue AI algorithm would presumably aim the defensive

missiles with the expectation that the red missiles would follow a feint to the right, as had been observed over the previous years of training and operations.

This example relies on a sudden discontinuity in behavior to confuse the algorithm, but poison data attacks often directly alter the physical information encoded into the algorithm. An adversary could alter just a few pixels in an image to cause the algorithm to incorrectly identify it. One study found that a photo of an academic building could be altered so that an image recognition algorithm identifies it instead as a dinosaur. To the naked eye, there is no difference between the two images.[53]

Poisoned data attacks seem like the type of risks that would be both persistent and constantly evolving—an adversary would be launching these attacks ceaselessly, while switching up the method. The better an adversary understands how the blue algorithm is distilling its behavior or characteristics, the more effectively the adversary could trick the algorithm. This would be a particularly likely place to see the effects of Clausewitz's "continuous interaction of opposites."[54]

Algorithm Errors

The second source of errors in AI is in the mechanisms of the algorithm itself. There are two main types of algorithms discussed here: (1) generative AI and (2) reinforcement learning. Each has a different combination of strengths and weaknesses. Generative AI is known to produce random answers or "hallucinations," and reinforcement learning is vulnerable to poor or incomplete goals. This section also discusses the challenges of assessing hybrid algorithms, which can combine several different types of algorithms.

Generative AI

For all its triumphs, generative AI—the type of AI that has fueled ChatGPT and other types of interactive models—is known to be vulnerable to mistakes, generating answers that are random or stochastic, meaning that some answers are not correct and the rate at which the algorithm will produce incorrect answers cannot be predicted with precision. This problem is more commonly known as hallucinations.

Over time, there are likely to be clever technical or mathematical strategies to manage this vulnerability. Technical strategies to assess the confidence associated with generative AI answers are emerging but are still in the research stage.[55] Among the disquieting realities about generative AI is the observation that there is a rather poor understanding around how it works.[56]

This report recommends that the DOD not incorporate generative AI into algorithms that are used in sensing and firing networks or other sensitive uses until there are significant advances in this type of algorithm to manage, mitigate, or eliminate its vulnerabilities.

Generative AI may be useful for other DOD functions, however, including screening materials for more fine-grained analysis by intelligence analysts. The department, though, needs to ensure that the use of generative AI is stated explicitly in the qualitative tools provided to military personnel and that its role is carefully evaluated in each use case because of its risk of erroneous outputs.

Generative AI, like all types of AI, is also vulnerable to reproducing errors or gaps in its training database. If the database is biased or out of date, answers generated by the algorithm also will be biased or out of date.

At least some of the recent efforts to improve the performance of generative AI algorithms have had unexpected and disquieting results, increasing the number of wrong answers given by the model and sometimes behaving in a way that a human would characterize as cheating to win in a competition.

Some of these oddities arise from the way the algorithms are trained. Generative AI algorithms are trained on huge databases of information. In the case of a large language model, the database could be millions of websites totaling billions of words, images, videos, or sounds.

The model that emerges from this training is sometimes called a foundation model. Foundation models are often optimized for a particular use through the addition of a user interface. Most famous, perhaps, are the user interfaces that function as chatbots that interact with user questions.[57] The user interface often involves additional training, which can both enhance the effectiveness of the interface for the user and introduce additional problems.

For example, chatbot interfaces often receive further training through reinforcement learning human feedback (RLHF), where a human reviewer is given two candidate answers to the same question and asked to identify the answer they prefer. This additional information about human preferences is used by the chatbot to refine the subsequent answers it generates for humans.

A recent study demonstrated—perhaps counterintuitively—that RLHF can increase the number of wrong answers an LLM provides. This is because prior to human training, the chatbot provided both correct and incorrect answers, as well as saying it did not know the answer to some questions. During RLHF training, it learned that humans did not like when it said it did not know.

This led to the model answering a larger percentage of the questions overall—some right and some wrong—and saying it did not know the answer on a lower percentage of questions. By offering answers to a greater proportion of questions it was ignorant about, the chatbot had a higher rate of incorrect answers overall.[58]

Further, recent studies indicate that even more sophisticated models called frontier LLM agents "can strategically circumvent the intended rules of their environment to achieve objectives, with more capable models showing this behavior more frequently." When playing a chess game against another model, the more sophisticated models would "often observe that the chess engine is too strong for them to win against and then hack the game environment to win." [59]

There is work underway within the DOD to understand how generative AI could be safely integrated into the warfighting enterprise.[60] However, the department should wait until the commercial technology sector makes significant progress in terms of accuracy before incorporating it into military capabilities.

The DOD, despite its many strengths, is unlikely to be the place that makes the technical breakthrough to render generative AI consistently accurate enough to be appropriate to perform safety-critical functions. It is reasonable to expect that at some point, generative AI could be incorporated into sensitive military capabilities. Based on current information, though, that point has not yet arrived.

Reinforcement Learning

The DOD should also be careful with the use of reinforcement learning algorithms, which devise the optimal strategy to achieve a specific goal within the constraints of a set of rules. Superbly suited to chess or other board games where the goal and rules are well specified, these algorithms are powerful but face more of a challenge in the real world, where a thorough-going characterization of the "rules" and the "goal" is much more difficult. Already used in autonomous cars, this is the type of algorithm most likely to be incorporated into military capabilities.

Some artificial intelligence experts warn that as algorithms become more powerful, creative, and interconnected through the internet, there is a growing risk of a dangerous "alignment problem." An alignment problem would emerge if the human programmer failed to fully specify the goal, and the algorithm misinterpreted the human's intention.[61]

A famous thought experiment about this problem is that of an AI-driven robot in a paperclip factory being told to produce paperclips. When the unexpectedly efficient robot has used up all the designated raw materials for the purpose, the robot goes to the parking lot and tears bumpers off cars to make more paperclips, consistent with its understanding of the direction it received.[62]

This example highlights the risk that humans will fail to caveat the specified goal in ways that they thought were implicit—either because the goal has changed over time, or the goal needed to be refined as preferences shifted. One leading thinker on reinforcement learning has urged computer scientists to increasingly move to program computers in a way that creates uncertainty about the goal on the part of the computer, so that it seeks regular updating and clarification from the human programmer.[63]

The DOD should incorporate this insight for its reinforcement learning algorithms. In other words, it should adopt strategies to avoid alignment problems through one of two broad strategies: by (1) building deference into the goal definition to assure that the algorithm checks back in for further direction at significant and regular junctures, or (2) limiting the specific geographic area where the targeting should take place, the duration of the engagement during which the algorithm would be allowed to operate, and other characteristics of the engagement.[64] Overly general guidance like "kill members of the opposing military" should always be avoided on principle and to avoid grievous misunderstandings by the algorithm.

This modification in goal definition is important, as alignment problems emerging in targeting algorithms due to reinforcement learning could lead to the most dystopian of scenarios. The prospect of cars, houses, or—worst of all—people with certain characteristics being designated as

targets by a roving uncontrolled robot should drive DOD programmers to incorporate the strategies in goal definition urged by thought leaders in the field of AI.

Hybrid Algorithms

The integration of multiple algorithms into a single hybrid algorithm would represent an additional level of complexity. Many real-world applications already use collaborating algorithms: For instance, an autonomous car driven by a reinforcement learning algorithm receives as input the findings of a separate image recognition algorithm which identifies items along the prospective route and assesses whether they are likely to stay still (e.g., lampposts) or to move, possibly into the route of the car (e.g., a cyclist).

So-called hybrid models bind the inner workings of algorithms even more closely together than existing collaborative models. Every algorithm, like every measurement, is likely to have an error margin, even if it is an accurate algorithm. The error range for different algorithms would in general be expected to vary independently, i.e., one error is a little high, and the other error is a little low.

Overall, this blending of different errors generally will not harm the final answer much, since the two errors often cancel each other out. There could be a situation, however, where all the errors trend in the same direction, and in that case, what were acceptably small errors in the supporting algorithms create an error of unacceptable size in the final, dominant algorithm.

Take for example a reinforcement learning algorithm that identifies an optimal search path for blue forces to move through a red village. There are no strictly military items in this village, but certain dual-use items have been used by a guerilla force in other villages.

The optimal search path is determined by several subordinate algorithms. These supervised learning algorithms evaluate images of dual-use items in the village to identify whether they are suspect or not suspect. Suspect items need to be approached cautiously by blue forces and carefully searched.

For accuracy, it would be best if the underlying incidence of guerilla infiltration in the current village was the same as the underlying infiltration of the villages searched by the experienced soldiers who labeled each of the images in the supervised learning training database. Statistics and guerrilla infiltration, though, are not always in perfect alignment.

Say the experienced soldiers were sent first to the villages with the most evident problems with guerillas. Seeing many dual-use items in use by guerillas, the experienced soldiers would accurately reflect their experience by labeling many images in the training database as being suspect. The algorithm would then reflect the expectation that all villages had significant infiltration by the guerilla forces.

Now trained, the algorithm is heading out to be used in the real world. The new village where the algorithm is being used, however, is not as intertwined with the guerrilla group as the first village. Very few of the dual-use items in fact are being used for military purposes. The dynamics created during the search by the difference between algorithmic expectation and reality could be

lethal, particularly if other factors make the search tense, such as the blue force being weary or inexperienced.

This example illustrates how an error in one algorithm could result in an even larger error in the overall hybrid algorithm. It should be noted that the original source of the error is another training database problem, namely that the original training database failed to reflect the underlying population from which the data was drawn. Qualitative tools that enable the leaders of small tactical units to thoughtfully evaluate warfare algorithms to assess whether they align with the actual conditions confronting the unit are needed.

These hypothetical scenarios show why the accuracy of the algorithms—including hybrid algorithms—need to be closely analyzed in some technical detail, particularly at the outset, before responsible generalizations can be made.

User Interaction Errors

The third source of errors in AI algorithms arises from how the human being provided with the algorithm's result interprets that result. These challenges are an issue in any domain where AI algorithms are used. The challenges of human-algorithm interaction in warfighting are likely to be particularly daunting because of the high stakes and the time pressure.

It is also likely to be challenging because of the unusually broad range of types of interaction. These may include heads-up displays in aircraft and other vehicles, controls for robotic arms, or robotic exoskeletons. This extensive range evokes the complicated topic of human-machine interface. The DOD has a long and often successful background in managing these types of human-machine interface issues. This study acknowledges these complicated problems but does not seek to uniquely advance this important debate.[65] The issues identified here are not from the unique human-machine interface context but rather arise from AI use in all types of domains.

Automation Bias

Automation bias is likely to be a significant vulnerability for military uses. The source of this error is in the readiness of a human operator to passively accept rather than rigorously reevaluate the finding of a complicated, quantitative process such as an AI algorithm.

There is substantial evidence to support the presence of this risk.[66] One study, for example, details an experiment where individuals were putatively being given a tour of a new building by a robot. As part of the study, the fire alarm suddenly goes off without explanation and smoke starts emerging from the hallways in front of the tour group. To the left of the group are clearly marked, smoke-free emergency exits.

The robot urges the tour group to follow it into the smoky hallways, stating that these hallways offer the safest route out of the building. The overwhelming number of individuals who participated in this study followed the robot into the smoky rooms, as opposed to exiting through the visible and well-marked emergency exits immediately on their left.[67] Similar automation bias

has been documented in commercial flights, car navigation systems, AI for self-driving cars, and the business sector.[68]

These "overtrust" risks appear to be heightened in time-sensitive, pressure filled contexts. The terrible rhythm of war has been discussed by several authors, and close study of these works indicates that once the battle begins there are a host of very human factors–including fear and revenge–that can compromise decisionmaking.

User Bias

While automation bias leads a user to passively accept the recommendation of an algorithm, user bias leads a user to accept the recommendation of an algorithm when it aligns with their expectation of the right answer, and to reject it when it does not.

There is a significant literature that demonstrates that quantitative findings consistent with a user's preexisting views are more likely to be accepted as opposed to a finding that goes against previously held beliefs.[69] Studies in the United States have demonstrated these effects in law enforcement, particularly in an algorithm that generated a score that purportedly reflected a defendant's risk of recidivism: "Judges were more likely to accept low scores for white defendants, while overriding similar scores for black defendants."[70]

There are indications in the literature that some techniques for presenting the findings of algorithms, particularly when they are set in a larger context of the strengths and weaknesses of algorithmic decision aids, could mitigate the ill effects of both automation bias and user bias. Efforts within the DOD should seek to investigate such strategies and implement those that show promise.[71]

Feedback Loop Bias

Feedback loop bias is a vulnerability of algorithms that use interactions with the algorithm as entries in an updated training database to further refine and update the algorithm.

Examples of such algorithms include Google search. Often, individuals who share a particular perspective on an issue are most likely to ask certain questions. When the algorithm offers nominated web sites, the bias of the individual may lead them to select a website congruent with that perspective. The next time the same question is asked, the algorithm will present the selected website as a higher nominee in the list of candidate responses to the search query. This is because the algorithm notes which nominated website was selected after a particular query was made. Selected webpages are then presented higher in the list of suggested websites in response to the same query the next time it is posed.

This creates a feedback loop bias where the biased website is more and more likely to be selected and hence to rise to higher levels in the list of websites suggested in response to the query.[72]

The Google search engine is an example of a mechanistically connected feedback loop bias. As has already been discussed, a hybrid algorithm can be vulnerable to the same problem at a slightly

greater remove. The report previously described how soldiers who had cleared villages with widespread guerilla infiltration had been asked to label the images of dual-use items that are commonly found in villages as being either suspect (and hence needing a thorough search) or not suspect (and thus not needing a search). The images and their labeling were then used as the training database for a supervised learning algorithm which contributed to a reinforcement learning algorithm that mapped the best path for a blue tactical unit clearing a red village.

As a general principal it is useful to use interactions with an algorithm to further refine and update it, but unanticipated effects in a military context should be carefully considered.

Defense experts need to move beyond wanting to add AI in general toward a more particular understanding of its strengths and weaknesses. AI algorithms have demonstrated a host of generic errors. These generally arise from errors in the training database, the mechanism of the algorithm itself, or from the user's interpretation of the algorithm result. Tragedy beckons if the DOD uses AI for warfighting but fails to recognize these many sources of potential errors.

Characteristics of Needed Analytical Tools

This report urges the creation of quantitative and qualitative tools to enable effective and responsible use of AI algorithms in war.

Quantitative tools are needed for three purposes: accuracy, robustness, and wargame instrumentation. Accuracy is assessed in the field (e.g., would a missile launched from behind this bush destroy that enemy tank and only that tank?) while robustness is more of a prediction to inform decisions in the command center and Situation Room (e.g., would prospective uses in a specific operation yield the expected military effects with acceptable collateral damage?) Wargame instrumentation is more complex because it requires modeling interaction effects– in other words, both what the effects of the blue weapons would be and the tactics and outcomes of the red forces. Well-instrumented wargames consistently have been an element of a successful peacetime military innovation.

It is critically important that these algorithms are not treated as black boxes by the military personnel using them. Military personnel do not need to be experts in AI, but they should be able to identify and evaluate whether the conditions of the real-world operation are sufficiently aligned with the design specifications of the algorithm to ensure that it can be responsibly used.

The qualitative tools should help military personnel make this assessment. Training materials and simulations should strive to highlight possible characteristics of a future operation that would render a standard algorithm too inaccurate to use in that operation. Alternative techniques to rectify problems with algorithms should be practiced in training exercises. Examples of these problems would include an operation where a friendly military is using Russian tanks, and the algorithm was trained to identify targets using the physical characteristics of Russian tanks.

The qualitative tools should also help military personnel assess whether the algorithm's confidence about distinction (i.e., accurately distinguishing noncombatants from combatants) needs to be adjusted given the characteristics of the operation.

For example, say such an algorithm is being used on an isolated military base. If new information about the context became available—for example, students from a civilian school are visiting the base—the blue military commander may need to increase the required confidence level for the algorithm to identify an input as a legitimate military target. The qualitative description of the algorithm should include the standard confidence level required for a positive identification and allow, in specified circumstances, on-scene commanders to adjust that confidence level.

This report has discussed various training database problems, including drift and poisoned data. In a particularly dynamic battlefield environment, it may be necessary to enable a tactical commander to identify when the accuracy of the pretrained algorithm becomes too low.[73] These units would ideally be equipped with the ability to build a new database in real time and train an algorithm to be much more accurate.[74] Creating such a capability at the unit level will require new analytics and equipment, but may be the only strategy to deal with fast-changing battlefield situations.

These types of qualitative information would also be valuable in more complicated uses of AI. A plausible use case would be rapidly integrating information from different streams of sensor data. The most complex type of integration might well happen for the fastest-moving adversary systems (e.g., a hypersonic missile).[75] In such a situation, there would be extreme time pressure to evaluate and ratify the assessment of the algorithm integrating each of the separate modalities of sensor data. However, the time pressure involved could severely limit the amount of real-time decisionmaking available to military commanders faced with an incoming attack. Trusting the AI might be the best—and indeed, only—course in this situation.[76] Where then, do qualitative tools come in, in circumstances like these?

Decision support systems should always contain explicit markings that a particular AI algorithm is being used, and the assumptions embedded in the algorithm should be available for study. But as in the example above, there should be no expectation that this qualitative data will always be studied during actual use. Rather, it would be available for review, and military personnel could be expected to review it during initial familiarization with sensors and before and after training sessions. Training—both classroom and exercise—should present scenarios where the assumptions in commonly used decision support algorithms are invalidated by some peculiarity of the scenario. Lessons learned sessions after the exercise should highlight the characteristics of the peculiarity that undercut the utility of the algorithm.

The objective is to build some expertise within the user base about the embedded assumptions, so that military personnel could assess whether a characteristic of a particular use invalidated some of the assumptions. This is a far from a perfect solution, but it seems like the best method to both benefit from algorithms' speed of integrating disparate sensor data and ensure that key system operators understand the assumptions embedded in these algorithms, so that they can identify those rare occasions when the assumptions might not hold and alter provisions accordingly.

Evaluating the qualitative tools for any algorithm to be used in a mission should be part of the regular operating responsibilities of the fighting unit itself, not only a review conducted outside of the unit by the DOD Test and Evaluation community as has instead been widely recommended.[77]

AI's Debut on the Battlefield

The first sustained campaigns of artificial intelligence in war have been observed in recent months on the battlefields in Ukraine and the border regions of Russia.[78] These instances provide ample evidence for the realist argument that AI-enabled military capabilities could be decisive in war.[79]

Ukraine first used small inexpensive commercial drones for surveillance and reconnaissance. A drone was likely the source of the location data that enabled a Ukrainian solider to target the shoulder-fired missile shots outside of Kyiv that disabled the first and last Russian tanks in the convoy that was headed toward Kyiv—but never arrived—in the early days of the Russian attack on Ukraine in February 2022.

Both Ukraine and Russia have since become even more skilled at using drones. The Ukrainians have largely used commercial drones developed for racing and have made parts using 3D printers to optimize for different missions, yielding a remarkable capability with a low per-unit cost.[80]

Eventually, Ukrainian soldiers began placing small explosive charges on the drones. These jerry-rigged drones were maneuvered by a Ukrainian operator over a valuable Russian target, at which time the operator would detonate the explosive. In response, the Russians jammed the communications between the operator and the drone, preventing the detonation.[81]

As with any conflict with a calculating adversary, the measure and countermeasure competition between defense and offense quickly got underway. Cost, technology, and tactics help determine who gains the upper hand in any given scenario. In this case, the Ukrainians responded to the jamming (at least in part) by having the drone use an image recognition algorithm to identify the target and detonate at a specified distance.

This may have been the first use of an AI sensor and firing network in warfare. It is a small network indeed but was often successful.

There is significant exploration underway in Ukraine and in most militaries about the next phase of operational and tactical uses for AI and small platforms, particularly drones. There has been exploration of ideas for drone "swarms" operating together to advance a common objective. These plans sometimes seek to use these swarms of drones to overwhelm defenses and pursue attacks. There is speculation that these swarms might be able to penetrate even sophisticated defenses, possibly at significantly lower cost than the cost of the defenses.[82]

While there has been meaningful progress in individual programs within the DOD, there should be a more systematic effort to understand, benefit from, and defend against these technologies. Peacetime innovation is difficult but vastly preferable to waiting for the lessons meted out in war.

War can demonstrate how new technologies can be most effectively incorporated into military capabilities. Those lessons, though, are dearly bought.

Conclusion

Henry Adams, who was quoted at the outset of this report, observed that there are limits to history's right to meddle because of the complexity of emerging technologies.[83] Meddling seems fraught indeed when one confronts the pace and complexity of advances in artificial intelligence. AI is enabling capabilities that are transforming most substantive domains more rapidly than the domain experts or the body politic at large can sort through the implications.

Bombs educate vigorously, but algorithms used for quieter purposes are more likely to creep in unnoticed, reproducing and reinforcing historical bias, or eliminating privacy, the companion to civil liberty.

The DOD could play a useful role if it catalyzes the development of tools that lower the tremendous difficulty of assessing, managing, and controlling algorithms. These tools could move into nondefense applications as well, enhancing the ability of experts in other domains to manage AI use responsibly.

Perhaps citizens are prepared to passively accept the decisions of algorithms powering their news feeds or the weapons of the military fighting in their name. Perhaps they are even prepared to passively accept the decisions of their government, forgetting the rights they secured when it was first created.

History may have no right to meddle. But citizens do and can strive to build the future they want. History will tell the story, but the people, by omission or commission, will write it.

Appendix I

A Meaningful Sensor and Firing Network

This report has discussed how a theater-wide sensor and firing network, with other reforms, could shore up the effectiveness of the U.S. power projection strategy. This appendix examines the details of a sensor and firing network and evaluates its challenges in the context of the concerns expressed by the three schools of thinkers about the cause and effect of war: the realists, the liberals, and the humanitarians.

Overall, meaningful decisions must be made about the use of force. On this point, at least, there should be broad agreement among the realists, liberals, and humanitarians. The decision should be meaningful in the sense that it aligns with well-considered assessments of military effectiveness and law of war compliance. How to secure such meaningful decisions, of course, is highly contested in AI-enabled war.

Realists tend to judge that war is generally either being deterred or fought. Given this contested environment, realists feel urgency to improve warfighting capabilities. Realists tend to be optimistic that mistakes in algorithms could be managed if not eliminated and hence are inclined to develop such algorithms and enjoy the expected warfighting benefits of speed and precision.[84]

Liberals tend to judge that it is possible to move beyond always having to deter or fight a war: Areas of cooperation between nation-states can be found, nourished, and sometimes sustained. Hence,

efforts to build responsible algorithms can proceed at a measured and deliberate pace. Liberals also are concerned that AI-enabled military capabilities are significantly vulnerable to unwanted escalation.[85] Unwanted escalation occurs when neither nation-state desired war but both fear losing if the other side attacks.

Humanitarians center their attention on human life and strive to build, nourish, and sustain norms to reduce suffering, including suffering caused by war. These norms tend to focus on the basic humanity of all people, but they acknowledge differences between combatants and non-combatants during war.

Humanitarians argue that the only way to assure meaningful decisions is to assure that a human is "in the loop" of a decision to target a particular item or to make and execute a firing decision. Under this view, relying on an algorithm, whatever its provenance, is inherently not meaningful.

The locus of concern for these different perspectives can be shown by looking closely at the different stages of a sensing and firing network. The challenge for accuracy—a blue algorithm properly identifying and characterizing a red target—principally emerges in the first stage of the firing network. The second and third stage both raise the risk of unintended escalation, and the third stage is where the conundrum about what truly constitutes "meaningful" control emerges, whether it is a human in the loop or an automated decision based on prior human approval under certain conditions.

The First Stage

The first stage of a sensor and firing network would be identification of a red target. The second stage would be the allocation of blue weapons against that red target. The third stage would be the decision to fire the blue weapons against the red targets.

As the number of targets increases, the complexity of each of the stages also increases.

For example, as the number of targets increases, the information that needs to be acquired about each red target in the first stage also increases. The additional information is needed so that the blue weapons could be allocated against the red targets optimally, maximizing military effectiveness and minimizing collateral damage.

A red target could be in an isolated area on a military base with a strictly controlled perimeter and no civilians. Or, at the other end of a theoretical continuum, it could be in densely populated civilian area. Most targets, of course, are likely to be in a blended context, with some civilian assets in the general area of a red military target.

Some red targets will have various physical characteristics that need to be considered when assessing from a military point of view how to destroy them. Some of these characteristics require special munitions, such as deeply buried or heavily protected red targets.

Red biological or chemical weapons stockpiles require special techniques to avoid lofting or otherwise dispersing the material in a manner that could contaminate surrounding populations or embedding it in the environment.

Some red targets will be considered "perishable" in the sense that their location could change quickly, and they might be very difficult to locate again. This category often includes items like mobile missile launchers and sometimes includes command posts. It also might include a military unit that is still located at a base but is expected to move into a civilian area.

Some heavily defended red military targets might pose significant risks to blue piloted aircraft and so should be preferentially attacked by drones or other unmanned techniques. These are examples of the type of information about a red target that would be needed to inform the allocation of blue weapons against those targets.

As this report has emphasized, these sensitive algorithms would require high standards for accuracy and robustness. Some algorithms—such as generative AI algorithms at their present state of technical development—are too vulnerable to errors to be used in such a network.

Required confidence for a targeting algorithm should be "dialable," with policymakers and commanders able to require a higher level of confidence for some missions in some contexts. The greater the risk that civilian and military personnel or equipment could be comingled should result in a higher level of confidence from a targeting algorithm. A lower (but still robust) level might be acceptable in a context that is confidently occupied by exclusively military equipment and personnel. This would be consistent with the requirements for distinction under the laws of armed conflict. Many argue that the precise effects of AI-enabled capabilities could, on balance, reduce collateral effects overall.[86]

Proportionality would require an assessment of the military benefit of a particular attack and its associated effects on civilian personnel or infrastructure. Certain munitions could cause more limited collateral damage than other types of munitions consistent with achieving the military objective. For example, a precision munition should be used in situations where the target is in a dense city environment and the overall calculation of military effect and collateral damage is assessed to be proportional.

This is an illustrative list of the characteristics of a red target that should be acquired in stage one to inform the allocation process in stage two. The target needs to be identified, validated, and then characterized so that this information could be used in the next stage.

These requirements for stage one reinforce the challenges for a sensing and firing network. This report argues that algorithms should be used for military purposes only after they have been thoroughly vetted for their accuracy and robustness in their intended military use case.

Recall that all three of the classic protagonists between realists, liberals, and humanitarians agree that algorithms would be challenged in war. Realists seem more prepared to judge that these challenges could be responsibly overcome. Liberals and humanitarians warn that these challenges

would be very difficult. Humanitarians judge them to be sufficiently difficult that it should not be attempted—that humans need to be in charge of authorizing individual targets and not relying only on algorithms to identify them.

In the view of this report, the realists make a compelling argument that there are many war-fighting benefits if these tasks could be performed at the speed of an algorithm. But only if the realists can demonstrate that the task is being done with proper care. Both military effectiveness and the law of war demand it.

The Second Stage

The second stage, allocating blue weapons against validated red targets, is well-suited to automation assuming that high-quality data would be available. As discussed, that is a big assumption for all but the smallest allocation of weapons against targets.

The extent to which the network would get accurate information is worth assessing, as well as its sensitivity to different levels of inaccuracy. It is currently difficult to maintain a common operating picture of blue weapons available in theater across all possible sources of those weapons. Putting together an integrated picture of red targets in theater, particularly with some of the fine-grained information this report argues for, is similarly difficult.

Escalation risks emerge because of the cost of delay. A decision to execute the firing plan would be expected to destroy all the red targets. A decision to wait—to conduct additional negotiations, for example, so as to reconcile the point of contention at the heart of the conflict—could turn the prospect of victory for blue into the reality of defeat. Red could use the time to launch its own force package, destroying many blue assets.

Even if the delay did not enable red to launch an effective attack, it would enable mobile red targets to hide, making it unlikely in many cases that their position could be found again; this might enable red to use some of their weapons against blue.

There would be the risk of new errors being introduced if the targets are not monitored after their validation in the first phase. For example, if significant time passes, a group of noncombatants traveling through the area might now surround a military target that was in an isolated area when initially identified and vetted.

It should be noted that unintended escalation is a danger only in those situations where neither state really wants a war. Uncertain about what the other side wants, and reticent to risk losing a war, they act first.[87]

This perception that military technology makes it likely that the attacker will emerge from the war victorious is called a period of offense dominance, while defense dominance means that the defender is expected, all other elements held constant, to win the battle.

Periods where the military technology is offense dominant are judged to be very vulnerable to unwanted escalation. In contrast, a period where the military technology is considered defense

dominant is generally considered to be much more stable. Most analysts predict that AI will be offense-dominant technology.

Concerns about AI's escalation risk seem warranted. If both sides had the comprehensive ability to launch a well-targeted, broad-based attack against the adversary's forces through large firing networks, victory likely would go to the attacker. There would be pressure to move swiftly to launch before the other side launched. This is the type of reasoning behind the focus on secure second-strike capabilities in the strategic nuclear domain.

Analytically, offense-defense theory is powerful. It is worth noting that what seems clear to historians and political scientists, however, appears to be significantly less clear to at least some practitioners operating with the more blurred perspective of the moment. Henry Kissinger, for one, argued that while intellectually attractive, the theory was essentially impossible to implement because of the ambiguity about whether forces were in fact offense or defense dominant.[88]

General strategies for managing escalation seem particularly worthwhile in AI because of the poor understanding of the technology in general and by military experts in particular. More fine-grained strategies seem difficult to develop now given the limited understanding of how they would be used.

The most basic tool for managing escalation risks is unclassified technical discussions between experienced defense experts, a subset of a category known as transparency- and confidence-building measures. Technical discussions could help each side understand the regular maintenance schedules of particularly sensitive weapons or create relationships that allow consultations when there are specific concerns.

As such discussions are relatively low-cost and low-risk if the participants are experienced professionals, they are worth pursuing for their benefits should an escalation risk emerge.

The usefulness of such discussions seems particularly powerful at this stage in the development of AI-enabled military capabilities. There is such poor understanding of how the algorithms would work that building a better picture of how allies, then friends, and then adversaries are thinking about this issue would be valuable. Ideally, this would help manage crises between two states if neither of them wanted war.

It is hard to imagine technical characteristics of AI-enabled military capabilities that would facilitate it being perceived as defense dominant to a potential adversary. The policy literature has not identified any plausible strategy for defense dominance or for transparency, both of which are helpful for managing escalation risks. Technical experts should consider whether there are any techniques to configure AI capabilities to help with these strategic problems.

Managing AI-related escalation risks should remain at the center of the policy agenda. As a better understanding of the strengths and weaknesses of AI-enabled military capabilities develops, there will be more opportunities to design better strategies to manage escalation risks.

The Third Stage

The third stage, making or executing a decision to fire, would have both escalation risks and automation concerns. Most pertinent to this stage are the compunctions of humanitarians, who,

in the context of the United Nations, have urged the adoption of additional prohibitions in international law that would apply to AWS.

The principal element of this position has been the requirement for "meaningful" human control of the weapons systems. This requirement has been defined as prohibiting an algorithm from selecting a particular target and from authorizing the firing decision. A human would need to have a meaningful opportunity to evaluate and approve or disapprove the targeting and firing decision (being in the loop).[89]

Additional regulations sought specifically by the International Committee of the Red Cross (ICRC) would have the effect of building additional space—in time and clarity—around nonmilitary targets to further reduce the risk of mistakes.[90]

The U.S. government, in the context of these negotiations, has opposed these prohibitions, calling instead for "appropriate" human control which could occur prior to an algorithmic target selection or firing decision and would be based instead upon thorough human vetting of an algorithm prior to its deployment and use.[91] Such use of AWS would need to be, like all use of military force, consistent with the existing requirements under the law of war—distinction, proportionality, and humanity. These laws would require accurate and robust algorithms to power AWS. The United States military is required to operate under the law of war.[92]

Realists have tended to argue, when they have engaged with the humanitarian perspective, that both targeting and firing algorithms could be constructed and used in a well-considered manner that reflected the highest standards of military effectiveness and observance of the law of war. Former Secretary of Defense Ash Carter described himself as "optimistic this can be solved, like every other hard problem, with diligent technology-informed effort."[93]

A realist might argue that such careful work could yield human control that was well-considered, indeed even meaningful in the common usage of the term. Much more meaningful, perhaps, than having an exclusive focus on having a human in the loop in a time-pressured, highly stressful situation, vulnerable to automation bias. Careful, meaningful human control, then, is needed, and one could argue is most likely to be better rendered in quiet analysis and considered decisionmaking well before the crisis.

The type of analytical tools called for in this report would be needed regardless of a final decision about autonomous target selection and firing. A human in the loop would have no easy time of it. While such individuals would doubtless do their best, they would confront the confusion and stress of warfare, combined with automation bias that tends to lead to the acceptance of the finding from an automated process.

There is a wide range of perspectives in the humanitarian community. This discussion will examine more closely the position of the ICRC, which calls for two types of AWS to be entirely prohibited: those that are too unpredictable, and those that target human beings.

In addition, the ICRC urges the adoption of four legal regulations on the use of any remaining AWS. The most significant of these is the requirement for effective human supervision of target selection and firing decisions. The other three regulations seek to build more space around targets to assure that there is a high threshold around them in terms of distance, decision time, and clarity, even with the requirement for effective human supervision.

Under these regulations, even autonomous weapons with effective human control could not be used in areas where there are civilians—there could only be military personnel or equipment. No blended areas would be allowed. The regulations also would require designating a specific area and delimited time period where the supervised AWS could be used. Outside of that space, even the supervised weapon system could not go.

These additional regulations would build protective space around civilian personnel and assets. From the realist point of view, this additional space would not be needed because of the accuracy of the algorithms and would cost too much in terms of military effectiveness. The assumption about military effectiveness, though, requires that the type of analytical tools called for in this report have been developed, deployed, and used to assure compliance with the existing laws of war. Absent the development and use of such tools, it is not clear how either claims of military effectiveness or compliance with the law of armed conflict could be made.

Ban "Unpredictable Automated Weapons Systems"

The ICRC recommends that there be a ban on "unpredictable automated weapons systems."[94] As with so many proposed legal prohibitions, the critical issue would be what precisely the words mean, in this case what constitutes an unpredictable system.

Unpredictable could refer to what has been the principal concern expressed in this paper: that military commanders and senior officials need to have a sufficient understanding of the accuracy and robustness of AI-enabled military capabilities before they authorize their use in war. This understanding is essential to inform a judgment about proportionality and distinction as required under the laws of war, as well as the military utility of a planned operation.

The ICRC recommendations explain that unpredictable AWS "that are designed or used in a manner such that their effects cannot be sufficiently understood, predicted or explained" should be prohibited.[95]

The requirement that the effects be predictable seems reasonable, and indeed it is necessary to meet the already existing requirements of distinction and proportionality. However, this language could be interpreted as requiring that the functioning of the algorithm should be "explainable"—a trait that neural networks is acknowledged to lack at present.

Neural Networks

Most artificial intelligence benefits from the base algorithm being supplemented by something called a neural network, which gives it additional precision as demonstrated in analyses that use a test database.[96] Neural networks generally are described as being a "black box" or having an "explainability" problem, and thus some analysts argue they meet the criteria to not be "sufficiently understood or explained." This provision then could be interpreted to proscribe the use of neural networks.

Their greater degree of accuracy is achieved by identifying the distinguishing characteristics of the inputs and assessing the relative influence of each of these characteristics on the output. In other words, the algorithm identifies the characteristics of a cat that are most useful to distinguishing it from a dog.[97]

In the more formal language of AI, the distinguishing characteristics are called "layers," and their relative influence is called "weights." The human programmer can specify the number of layers that should be used and is able to discern the weight assigned to each layer after the algorithm has been optimized by the computer. The human programmer generally cannot, though, figure out which distinguishing characteristic of the inputs each layer signifies.

The algorithm selects the layers after closely examining the inputs and identifying their important characteristics. The algorithm then tunes or adjusts the relative weighting of the layers until their relative weighting optimizes the accuracy of the algorithm. This unpacking of the characteristics of the inputs and then their relative weighting provides the additional precision available from neural networks.

Highly skilled technical experts with significant computing power can unpack the substantive characteristics of some of the layers using post-hoc analysis. In other words, after the algorithm has optimized the weighting and completed its work, these high-end technical experts can go back and through a highly structured trial and error process generally sort out the substantive characteristic represented by the particularly high-impact layers.

It is always problematic when new provisions of international law are adopted which seem to mean different things to different people. Mandating an understanding of effects is reasonable but already required; being able to explain how neural networks work is currently impossible although they can be demonstrated to add precision. Based on the information available, neural networks appear to provide many benefits in terms of accuracy and few unique risks. Prohibiting them does not seem warranted on the available information.

This additional prohibition could leave military capabilities developed in good faith to be prohibited from use because of a different or contested interpretation.

Ban Targeting of Humans

The ICRC also recommends a prohibition on algorithms that target humans. This is for two reasons: first, the ethical concerns that "center on the interrelated loss of human agency, moral responsibility and human dignity in life and-death decisions," and second, the difficulty of distinguishing between active military combatants, inactive military combatants, and noncombatants.[98] Military combatants are considered inactive if they have been wounded, surrendered, or are sick. Inactive military combatants are considered out of combat and thus not a legitimate military target.[99]

This proposed prohibition may be redundant, as it is difficult to come up with ideas for algorithms that targeted active human combatants that would be accurate and robust without the use of extensive context clues like the individual standing behind an artillery piece and pulling the trigger. Most other ways of identifying this population would be too easy to alter, such as the military combatants changing out of uniforms and into civilian clothes.

And of course, the U.S. military already has the responsibility to have significant capabilities for distinction under the current law of armed conflict. There should be a high threshold for the use of any algorithm that targets a human being and extremely thorough vetting before the use of any such algorithm was authorized.

Regulations

In addition to the prohibitions called for above, the ICRC recommends four regulations to limit the use of any remaining autonomous weapons systems.

DISTINCTION BETWEEN MILITARY AND CIVILIAN TARGETS

The first proposed regulation states that algorithms could only be used to target (with effective human supervision) "objects that are military objectives by nature," while the third regulation would impose "limits on situations of use, such as constraining them to situations where civilians or civilian objects are not present."[100]

These regulations would significantly supplement and reinforce the requirement under the law of war for distinction. They supplement the requirement for effective human control and not targeting a human. The language about constraining AWS to situations where neither civilians nor civilian objects are present would greatly expand the safety zone around items and personnel, appearing to rule out the use of AWS in areas where there are both civilian and military targets blended together.

DURATION, SCOPE AND SCALE

The second regulation would require "limits on the duration, geographical scope and scale of use, including to enable human judgement and control in relation to a specific attack."[101]

It could be a useful best practice for this to be a technique by which some AI-enabled military capabilities operate. This approach has significant promise to avoid some of the risks of reinforcement learning algorithms, for example.

Again, however, the United States should not agree with this regulation as a legal prohibition, because there exists a small number of use cases that would not be compatible with it but still legal under the law of war and broadly useful to both the military and civilian population in an area.

An example would be a drone that followed a container with biological weapons. Maintaining positive surveillance of the container could be mission critical and would be complicated by the recommended strict limits on duration and geographical scope.

The AWS could be programmed to follow the container until it could identify a suitable method and location to destroy the container without lofting. Destruction should not occur where there would be effects from the lofting of the weapons into the environment and the surrounding populations.

EFFECTIVE HUMAN SUPERVISION

The fourth regulation would require "human-machine interaction, notably to ensure effective human supervision, and timely intervention and deactivation."[102] This raises one of the most persistent issues—whether a human needs to be in the loop of algorithm-directed targeting and firing decisions. Recall that under the ICRC prohibition, this would apply only to algorithms targeting equipment or buildings, as the targeting of humans has already been prohibited under another aspect of the proposal.

Whether or not algorithms could responsibly target and make a firing decision remains to be demonstrated through the type of analysis called for in this report. Certainly, the responsibilities under the current law of war should be taken seriously. Humanitarians certainly have some evidence on their side that this has not invariably been the case.

This report argues that appropriate tools to assess and predict the effects of a military strike need to be developed. Once developed, these tools could assist human commanders to carefully assess targeting and firing decisions much earlier, when there is still time to do careful analysis of the algorithms. This assessment could enable the responsible human preauthorization of targeting and firing by algorithms, under carefully proscribed rules. The midst of warfare very rarely provides time for the careful deliberation properly desired amidst so much violence, whatever the hopes that it could be "meaningful."

Conclusion

Realists are correct to argue that AI-enabled military capabilities—assuming they are accurate and responsible—could make the difference between victory and defeat in the wars the United States most needs to deter or, if necessary, fight effectively. Battlefields including in Afghanistan and Ukraine provide tantalizing evidence that these can be transformed into war-winning capabilities.[103]

However, the cautions brought to bear by liberals and humanitarians about escalation risks and basic humanity also are powerful. There should certainly be a presumption against the targeting of humans. If such targeting algorithms are ever adopted, they should have a particularly thorough vetting process.

The core message of this report is that vastly better tools must be developed to assess the accuracy and robustness of AI-enabled military capabilities. Significant confidence in these systems would need to be built before autonomous firing of any targets, much less of human beings, could be authorized. Once they could be responsibly authorized, however, it is difficult to deny that their speed and precision could confer significant warfighting benefits.

In this technological and strategic moment, AI-enabled military capabilities could help to deter war and, if necessary, fight it effectively. On this point, the realists seem to be right. But they forget too often that they are right only if the effects of these capabilities could be accurately predicted. Without such tools, neither wartime effectiveness nor observance of the law of war could be assured.

Appendix II

A Crucible for Reform

AI-enabled military capabilities of various sorts seem a promising technique to strengthen U.S. military forces in the near- and mid-term. Such capabilities also need to be defended against: AI-enabled military forces could be used by an adversary to further erode the force balance and weaken deterrence in key regions.

The DOD is not moving quickly enough to incorporate these emerging technologies or to embed them in operational concepts. Making thoroughgoing reforms to the DOD's processes will be complicated and time consuming. This report urges the secretary to instead launch a focused effort to incorporate AI-enabled military capabilities more swiftly.

The secretary of defense should direct select combatant commanders to develop new operational concepts using AI, empowering them through the grant of special authorities for AI, related communications, and small platforms, particularly drones. The select combatant commanders should include at least the combatant commanders of the Special Operations Command, the Indo-Pacific Command, and the European Command.

These authorities would allow the commanders to navigate around some of the most serious problems within the DOD's current processes and structures. They would align responsibility for planning for wartime missions with budget authority, acquisition authority, civilian and military personnel assignment and hiring, and freedom from many regulations on encryption and classification, among other topics. This context should create a promising environment for identifying operational concepts for the use of AI-enabled military capabilities.

The report also specifies that developing analytical techniques to understand and predict the effect of using algorithms in war and to instrument wargames would be among the responsibilities of the select combatant commanders, as well as of a reorganized organization in the OSD.

This grant of authorities to the combatant commanders should be reviewed by the secretary of defense after two to four years. The secretary could decide to keep these authorities with the combatant commanders or reallocate them in some way.

At this review point, there should the selection of an appropriate set of DOD-wide analytical tools and their consistent use thereafter should be overseen by the new OSD office. Directing multiple organizations to develop the same tools during the experimentation period may yield some duplication. However, developing the tools will be difficult analytically and the alternative approaches should help assure that many different perspectives inform the final metrics.

Specifically, the authorities proposed would empower the select commanders to develop new operational concepts and AI capabilities through:

- giving them funding and decision authority for the development and acquisition of AI, related communications, and small platforms, including drones;
- giving them authority and empty billets to build military and civilian teams with needed expertise through by-name requests and special hiring authorities;
- empowering them to waive regulations (e.g., classification, encryption);
- allowing them to identify the tactical problems for analysis;
- directing them to develop defenses against attacks on algorithms; and
- directing them to develop analytical tools and wargame instrumentation.

A well-structured competitive process as recommended here could align the disputatious nature of the Pentagon with the objective of the United States for its military forces: developing the best-possible military capabilities in this era of profound strategic and technological change.

Analysis alone, of course, would not enable a good idea to triumph in the bureaucratic battles within the Pentagon. New war-winning operational concepts rarely emerge by acclimation in large bureaucratic militaries; there are too many established skills, capabilities, and weapons systems that new concepts would render unnecessary. For important findings to be persuasive, they often need to emerge from well-run, well-instrumented war games.[104]

Moreover, a properly structured competition undergirding the wargames can create incentives to unearth good ideas and give them visibility. Visibility through the wargames may buy good ideas enough time to demonstrate their promise to observers, at least some of whom may help the ideas survive to maturity.[105]

Vesting senior military officers with the principal responsibility to devise innovative operational concepts does not ignore the historical reality that senior military officers do not in fact invariably come up with war-winning innovations—sometimes they come from senior civilians.[106]

But both history and common-sense caution that whatever the source of the good idea, it will need to be embedded in military tactics and units and so will ultimately have to be accepted by at least some senior warfighters. The wargames recommended here should provide the opportunity for a good idea—whatever its parentage—to be seen and evaluated. This would provide enough of an opening for an effective civilian leadership team to build support among at least some military allies to secure the adoption of their best ideas.

Empower Senior Warfighters

Putting select combatant commanders in charge makes sense. The combatant commanders are the best proxy in the DOD for a senior official in charge of planning for and prosecuting a specific war.[107] Combatant commanders are four-star generals or admirals assigned to one of eleven broad sets of responsibilities, such as protecting and advancing U.S. defense interests in Europe or the Pacific region. They tend to focus on specific, concrete military challenges and integrate the contributions of each of the military services.

In contrast, the separate military services—the Army, Navy, Air Force, and Space Force—each bring a valuable long-term perspective that focuses on the general challenges presented by their domain (land, sea, air, or space). Each has a proud tradition and particular expertise.

The military services and their associated departments have the responsibility to train and equip the force. In practice, this means that the services hire young people and manage their careers as they grow into more senior officers and enlisted personnel. These careers generally are shaped around training and assignments in domain specific equipment, challenges, and units.

The services also research, develop, and acquire most weapons systems. The OSD or the Joint Staff have a variety of ways to influence service decisions—some direct, some indirect, and some negligible. The vast amount of people and dollars in the DOD are controlled directly by the services, whose decisions are shaped in part by the preferences expressed by the secretary, the chairman of the Joint Chiefs of Staff, and other senior leaders.

The secretary of defense has vast formal authority but confronts extraordinary constraints from the sheer complexity of the enterprise. Former Secretary of Defense Robert Gates observed, "The very size and structure of the department assured ponderousness, if not paralysis, because so many different organizations had to be involved in even the smallest decisions. The idea of speed and agility to support current combat operations was totally foreign to the building."[108]

The combatant commanders' focus on specific military problems is ideal for the task of developing algorithms because AI will do best if they focus—at least at the outset—on specific use cases in a particular geographic context. The combatant commanders might, for example, be thinking in a detailed manner about how to contribute appropriately to stopping or slowing a Chinese attempt to secure military control of Taiwan.

As specific algorithms are validated in these more limited contexts, both the algorithms and the methods to measure their accuracy could be expanded to assess more general applications. This

work could yield general insights about the sensitivity or durability of algorithms for related but different uses.

The combatant commanders also make sense because they generally are experienced strategic and military leaders, comfortable specifying objectives and holding large organizations to account for meeting those objectives on a specified timeline. Most are comfortable making difficult decisions to replace ideas or subordinates who consistently fail to deliver. They want to win, particularly in high-visibility competitions.[109]

Assigning the task of developing AI-enabled operational concepts to select combatant commanders also avoids the turmoil of creating, staffing, and provisioning a new organization for this purpose.

The combatant commanders also make sense because the chain of command for military operations runs to them through the secretary of defense from the president. This would make the combatant commanders responsible for any future decisions recommending or ordering the use of sensitive AI-enabled military capabilities. Aligning responsibility for the use of AI capabilities with the development and assessment of those capabilities seems appropriate.

This report has recommended that the use of sensitive AI algorithms should require authorization in the rules of engagement for a military operation.

When considering whether to keep the funding and acquisition authority at the combatant commanders, the secretary will want to consider the longer history of Special Operations Command (SOCOM). SOCOM uniquely among the combatant commanders has its own budget and acquisition authority, in its case for "[Special Operation Forces] peculiar" equipment.

SOCOM has been pointed to as being more amenable to incorporating artificial intelligence because of their more streamlined acquisition process and innovative ethos. AI, though, should not be restricted to SOF; it would be important to include at least the commands in the Pacific and Europe in this effort. The swift acquisition track record and ethos of SOCOM is widely acknowledged but it has also sometimes raised classic concerns about the proper balance between innovation and oversight.[110]

The DOD and High-Tech Startups

A persistent problem for the DOD is that it is not an attractive customer for the tech startups that have the greatest expertise in cutting-edge fields like artificial intelligence. The DOD's processes are too bureaucratic and slow.[111]

Several new authorities would be given to combatant commanders under this proposal. These would include moving money and decision authority for the development and acquisition of AI, related communications, and small platforms to the select combatant commanders. The commanders could contract with many sources including the Defense Innovation Unit (DIU), the services, commercial startups, or more traditional DOD contractors.

The personnel authorities would enable the combatant commanders to hire respected DOD acquisition experts with a demonstrated track record of responsibly using faster acquisition tools

like Other Transaction Authorities. These tools enable the DOD to execute contracts on a time frame appropriate to the commercial technology sector.

The benefits of the hiring authorities would contribute across the board. Innovators in any field are difficult to find, and once found, are difficult to protect and allow to flourish. The authority to make by-name requests for military personnel should facilitate building the needed teams. As the term "by-name request" suggests, the combatant commander could request a specific military officer; generally, the position has to be posted in more generic terms (a 0-6 with the following military operational specialty. . .).

The combatant commanders also should be able to attract top technical and engineering experts. There should be funding and authorities to hire a variety of external civilians to enable excellent technical advisors to come on board to advise the commanders and their teams, in addition to the civilians working as part of an external team of contractors.

Focusing the development of these AI and AI-related capabilities in the combatant commands should help avoid persistent problems like the limited interoperability among the service systems. The combatant commands are joint—meaning that they integrate the contributions of the services—and thus emphasize the type of interoperability that will be needed for the integrated networks generally envisioned for AI systems such as an integrated sensor and firing network.

The DOD and Creative Destruction

There is not a single prescription for creativity in war. Military innovation, particularly in peacetime, benefits from an idea about the nature of the future of war, an understanding of specific military problems, and an iconoclastic readiness to cause a little creative destruction. It often includes a good understanding of new technologies, and is informed by rigorous analysis, including tough lessons-learned reports and well-instrumented wargames.[112]

The call for wargames and the development of metrics, both quantitative and qualitative, seeks to ensure the wide availability of quality analysis often correlated with successful peacetime innovation.

The personnel authorities should enable the combatant commanders to put together an ideal team: By-name requests for military officers and access to civilian hiring authorities would enable them to find military officers with a demonstrated interest and ability in innovation and technical experts interested in working with the military.

The competitive element between the different combatant commands and the commands and OSD seeks to overcome the inertia of any bureaucracy in favor of the existing approaches. The prospect of at least two years and not more than four years creates time pressure to demonstrate progress to make the case for keeping the authorities at the combatant command level.

The authority to waive regulations should enable investigation of interesting trade space, of at least two types. The first is how, within a fixed cost, to best trade between large numbers of platforms

with low capabilities and small numbers of platforms each with high capabilities.[113] The DOD has started some efforts to explore these issues.[114] Giving the combatant commanders the ability to waive many regulations on encryption or robustness of communication allows real experimentation with low-cost options—requirements for high-end encryption or robust communication drive up the per unit cost, while obviously providing often valuable capabilities.[115]

The second interesting trade space is how to classify algorithms trained on large databases containing only a small amount of classified data. The benefit of unclassified algorithms would be the elimination of expensive infrastructure surrounding all classified items and the limited availability of the resulting algorithm for some uses, allies, and friends. The risk is that the algorithm could be reverse engineered and so an adversary could acquire the classified information and learn about sensitive sources or methods.

Investigating this trade space is important because the default decision could be to classify any algorithm at the level of the most sensitive piece of information in its training database. The benefits and risks of different classification norms should be investigated.

The combatant commanders are directed to conduct their work to the maximum extent possible in safe "sand boxes," or controlled environments within which AI and related capabilities can be tested without having immediate real-world consequences. There is an acceptance of some risk when investigating new, promising operational concepts.

The DOD and Oversight

Spurring innovation while preserving appropriate oversight is difficult. This report has made a distinction between sensitive uses of algorithms and other uses. It proposes that sensitive uses of algorithms require explicit approval by the secretary of defense in the rules of engagement for an operation.

The rules of engagement are the rules governing the use of force by U.S. military members. The secretary of defense approves standard rules of engagement and operation-specific rules of engagement for high-profile or risky operations. Subordinate commanders also may issue supplementary rules of engagement.

Having a category of sensitive algorithms that require approval at the secretary's level would be a valuable method for assuring that these algorithms get the oversight appropriate to their use.

Sensitive uses would include time-sensitive targeting or firing algorithms, regardless of whether there was a human in the loop. Algorithms embedded in processes that are not time sensitive and include multiple levels of review, including by human experts, would not be considered sensitive. Other algorithms should be evaluated for whether they should properly be considered sensitive and thus receive this high-level of scrutiny.

Requiring approval in the rules of engagement for the use of sensitive algorithms would assure oversight. The department's top officials would review materials going to the secretary of defense

for decision from many perspectives, including professional military judgment, strengths and weaknesses of AI, and law of war.

These officials would bring their expertise and that of their staff to consider important questions, including: How well do these algorithms work in general? How well are they likely to work in this specific operation? What could go wrong with the algorithms? Could their use violate the law of war?

In time, the process and lessons learned could identify systematic problems that should be more closely examined in the context of the DOD's regular budget review or other processes. Because of the size and complexity of the DOD budget, the budget review process can identify only a small number of issues for closer review and scrutiny by the most senior officials.

DoD's review processes generally are highly structured and organize decisions for review and ultimately approval by the secretary of defense. The most prominent of these processes is called the Planning, Programming, Budgeting, and Execution process, which formally reviews selected topics from the various parts of the DoD budget as they move up the chain of command for approval before appearing in the government's budget request to the Congress.

There also are processes for major acquisition programs and for the validation of future military needs or requirements. These processes generally are characterized by meetings of increasingly senior groups of officials in the DOD, typically including representatives from the military services or departments, the combatant commanders, the Joint Staff, and the OSD.

The report has called for a reorganized OSD office to confront the unique complexities of using AI in war. It should include experts in two difficult technical areas: (1) artificial intelligence and its management, and (2) warfare and its conduct and analysis. Experts should be drawn from several existing offices and positions in the Pentagon, including the CDAO, the deputy assistant secretary of defense for plans and posture, the director of cost assessment and program evaluation; and the director of test and evaluation.

Separating the technologists from the warfighters, as current efforts largely do, would maintain distance despite the desperate need for greater communication across these substantive domains. Communication and some shared understanding between experts in these two domains will be essential to highlight where AI could be vulnerable to errors in wartime contexts.

Further, the report recommends ensuring that leading academic thinkers on artificial intelligence—ideally as a part of the JASON group of technical advisors—review how the DOD intends to use these technologies. The JASON group is a small, self-selected group of Nobel Prize winners and other leading technologists in the United States who advise the national security community on some of its most vexing strategic and technological issues.[116] This expertise is essential given the rapid pace of AI technologies and the unique risks of using AI for military purposes.

Conclusion

This proposal would create the context most likely to produce new operational concepts to effectively employ AI in war and to defend against its use by an adversary. It moves the money and authority to senior warfighters with responsibility for planning and fighting the nation's wars for at least an initial period of experimentation.

It should facilitate tapping into the companies at the forefront of global, commercial cutting-edge technologies and assembling teams of creative, iconoclastic military officers and civilian experts.

Change is hard. Creative destruction in any large bureaucracy is difficult, and it is particularly difficult for a military service that can only select its leaders from individuals who have had successful careers in the organization that needs fundamental reform.

Artificial intelligence seems a beguiling solution to complexity. Realist have expressed confidence that its risks can be responsibly managed. Former Secretary of Defense Ash Carter declared that "good engineering design can accommodate both high performance and good ethics."[117]

So let the realists get their "diligent technology-informed effort" underway.[118] Require careful assessment of their claims of accuracy and robustness and oversee closely the use of the most sensitive algorithms in war.

Artificial intelligence may prove, like war, to be less understandable and compliant than had been expected. But this too is surely better understood sooner rather than later. Education of various sorts lies ahead.

About the Author

Carol Kuntz teaches on the policy implications of artificial intelligence at Georgetown and the George Washington Universities and conducts research as an adjunct fellow (non-resident) in the Strategic Technologies Program at the Center for Strategic and International Studies (CSIS). Dr. Kuntz served in the U.S. Department of Defense (DOD) for more than 30 years. Her work particularly focused on identifying changes in the strategic and technological environment and crafting new policies and programs given those changes. In the several years before Covid-19 emerged, she helped embed cutting-edge biotechnologies into the DOD's biodefense program to strengthen its ability to protect against novel pathogens. For the five years after the 9/11 terrorist attacks on the United States, Dr. Kuntz served as the homeland security adviser to the vice president of the United States. She led two presidential initiatives to strengthen defenses against chemical, biological, radiological, and nuclear attacks through the creation of the Domestic Nuclear Detection Office in the Department of Homeland Security and of Project Bioshield in the Department of Health and Human Services. At the end of the Cold War in 1989, she worked with top DOD officials to understand and respond to the changed strategic environment, helping to craft various strategy and defense planning documents. Dr. Kuntz received her PhD in political science from the Massachusetts Institute of Technology. She received her MPA from Princeton University and her BA from Cornell University. Dr. Kuntz received numerous awards over the course of her government career, including twice receiving the Secretary of Defense Medal for Meritorious Civilian Service. Her recent publications include the CSIS report *Genomes: The Era of Purposeful Manipulation Begins*.

Endnotes

1 Henry Adams, *The Education of Henry Adams* (Boston: Houghton Mifflin Company, 1918), 496.

2 David A. Ochmanek et al., *Inflection Point: How to Reverse the Erosion of U.S. and Allied Military Power and Influence* (Santa Monica, CA: RAND Corporation, 2023), www.rand.org/t/RRA2555-1.

3 Eric Schmidt et al., *Final Report of the NSCAI* (Washington, DC: National Security Commission on Artificial Intelligence, March 2021), 137, https://reports.nscai.gov/final-report/. See also Michèle A. Flournoy, Avril Haines, and Gabrielle Chefitz, *Building Trust through Testing: Adapting DOD's Test & Evaluation, Validation & Verification (TEVV) Enterprise for Machine Learning Systems, including Deep Learning Systems* (Washington, DC: WestExec Advisors, October 2020), https://cset.georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf; and National Academies of Sciences, Engineering, and Medicine (NASEM), *Test and Evaluation Challenges in Artificial Intelligence-Enabled Systems for the Department of the Air Force* (Washington, DC: The National Academies Press, 2023), 7, https://doi.org/10.17226/27092.. The latter study "emphasizes the need to retrain AI models to meet unanticipated and changing operational conditions."

4 David A. Ochmanek, *Determining the Military Capabilities Most Needed to Counter China and Russia* (Santa Monica, CA: RAND Corporation, June 2022), https://www.rand.org/pubs/perspectives/PEA1984-1.html; and Christian Brose, *The Kill Chain: Defending America in the Future of High-Tech Warfare* (New York: Hachette Books, 2020); Jack Watling, *The Arms of the Future* (London: Bloomsbury Academic, 2024).

5 NASEM, *Test and Evaluation Challenges*, 7, 32.

6 These are the five principles specified in a 2021 memorandum from the deputy secretary of defense on "implementing responsible artificial intelligence" as essential "DoD AI Ethical Principles." Kathleen Hicks, "Implementing Responsible Artificial Intelligence in the Department of Defense," U.S. Department of Defense, May 26, 2021, https://media.defense.gov/2021/May/27/2002730593/-1/-1/0/IMPLEMENTING-RESPONSIBLE-ARTIFICIAL-INTELLIGENCE-IN-THE-DEPARTMENT-OF-DEFENSE.PDF.

7 Kennedy reflects on the sometimes unexpected differential ability of different societies to benefit from new technologies: "The relative strengths of the leading nations in world affairs never remain constant, principally because of the uneven rate of growth among different societies and of technological and organizational breakthroughs which bring greater advantage to one society than to another." Paul Kennedy, *The Rise and Fall of the Great Powers* (New York: Random House, 1987), xv.

8 Soon after the end of the Cold War, these U.S. forward presence and power projection forces were sized to be effective in fighting two major near-simultaneous regional contingencies. Then, they were assessed to be capable against one major and one minor regional contingency, and then to win one contingency and hold in a second. For two near-simultaneous major regional contingencies, see Dick Cheney, *Defense Strategy for the 1990s: The Regional Defense Strategy* (Washington, DC: U.S. Department of Defense, January 1993), https://nsarchive2.gwu.edu/nukevault/ebb245/doc15.pdf. This stance was updated by 2012: "Mr. Panetta is expected to outline plans for carefully shrinking the military . . . he will say that the military will be large enough to fight and win one major conflict, while also being able to 'spoil' a second adversary's ambitions in another part of the world while conducting a number of other smaller operations, like providing disaster relief or enforcing a no-flight zone." Elisabeth Bumiller and Thom Shanker, "Panetta to Offer Strategy for Cutting Military Budget," *New York Times*, January 2, 2012, https://www.nytimes.com/2012/01/03/us/pentagon-to-present-vision-of-reduced-military.html. By the 2020s, the commitment had further dwindled: "The Trump administration said it was one major conflict and 'deterring' a second conflict." Mark F. Cancian, "Force Structure in the National Defense Strategy: Highly Capable but Smaller and Less Global," CSIS, CSIS *Commentary*, October 31, 2022, https://www.csis.org/analysis/force-structure-national-defense-strategy-highly-capable-smaller-and-less-global.

9 Steven Kosiak, *Is the U.S. Military Getting Smaller and Older? And How Much Should We Care?* (Washington, DC: Center for New American Security, March 14, 2017), https://www.cnas.org/publications/reports/is-the-u-s-military-getting-smaller-and-older; U.S. Department of Defense, *Military and Security Developments Involving the People's Republic of China, 2024* (Washington, DC: U.S. Department of Defense, December 2024), https://media.defense.gov/2024/Dec/18/2003615520/-1/-1/0/MILITARY-AND-SECURITY-DEVELOPMENTS-INVOLVING-THE-PEOPLES-REPUBLIC-OF-CHINA-2024.PDF.

10 "China's official aim is to 'modernize' its armed forces by 2035 and make them world-class by 2049 . . . [China's navy] is not just the world's largest but also has the newest vessels . . . China's air force, called the PLAAF, is on a similar trajectory . . . The most striking progress by China has come in the area of hypersonic missiles, which fly and manoeuvre at more than five times the speed of sound . . . China has already deployed multiple hypersonic-weapon systems . . . Few peacetime precedents exist for the speed and scale of China's military modernization." "In some areas of military strength, China has surpassed America," The Economist, November 4, 2024, https://www.economist.com/china/2024/11/04/in-some-areas-of-military-strength-china-has-surpassed-america. See also Elliot Winter, "The Compatibility of Autonomous Weapons with the Principles of International Humanitarian Law," *Journal of Conflict and Security Law* 27, no. 1 (2022), 5, https://doi.org/10.1093/jcsl/krac001, which states: "China has spent an estimated $4.5 billion on drone technology, including autonomous drones, according to the Stockholm Peace Research Institute."

11 Robert M. Gates, "The Dysfunctional Superpower: Can a Divided America Deter China and Russia?" *Foreign Affairs* 102, no. 6 (November/December 2023), https://www.foreignaffairs.com/united-states/robert-gates-america-china-russia-dysfunctional-superpower; Mark Milley and Eric Schmidt, "America Isn't Ready for the Wars of the Future," *Foreign Affairs* 103, no. 5 (September/October 2024), https://www.foreignaffairs.com/united-states/ai-america-ready-wars-future-ukraine-israel-mark-milley-eric-schmidt. See "Throughout the duration of the blunt phase, establish and sustain a sensing and targeting grid over the battlespace. The grid must be able to find, identify, and track ships, aircraft, and vehicles associated with enemy invasion in the face of intensive air defenses, counter-space weapons,

cyberattacks, and sensor and communications jamming. The grid should be connected to air, land, and maritime operations centers via robust data links. The grid should also be capable of autonomously nominating and guiding weapons to targets in cases where those links are temporarily severed." Ochmanek, *Determining the Military Capabilities Most Needed to Counter China and Russia*, 7; Ochmanek et al., *Inflection Point*; Brose, *The Kill Chain*; and Watling, *The Arms of the Future*.

12 Ochmanek, *Determining the Military Capabilities Most Needed to Counter China and Russia*.

13 Gates, "The Dysfunctional Superpower"; Milley and Schmidt, "Losing the Wars of the Future"; Ochmanek, *Determining the Military Capabilities Most Needed*; Ochmanek et al., *Inflection Point*; Brose, *The Kill Chain*; and Watling, *The Arms of the Future*.

14 While the *Economist* article does not explicitly refer to a plan to include AI in the plans for the golden dome system, it is very difficult to create a plausible alternative technique to move swiftly enough given the constrained decision timelines to make the system work absent at least some AI-authorized firing decisions. "The plan to protect America by shooting down missiles mid-air," *The Economist*, May 22, 2025.

15 Milley and Schmidt, "Losing the Wars of the Future"; and Schmidt et al., *Final Report*.

16 Ann Finkbeiner, "Jason- a secretive group of Cold War science advisors-is fighting to survive in the 21st century," *Science*, June 27, 2019, https://www.science.org/content/article/jason-secretive-group-cold-war-science-advisers-fighting-survive-21st-century.

17 "How Israel's 'Iron Dome' works," *The Economist*, July 15, 2014, https://www.economist.com/the-economist-explains/2014/07/15/how-israels-iron-dome-works; Lara Jakes and Aaron Boxerman, "How does Israel Defend Against Iran's Missiles?" *New York Times*, June 13, 2025, https://www.nytimes.com/2025/06/13/world/middleeast/israel-iran-missiles-defense.html; and Agnes Chang and Samuel Granados, "How Missile Defense Works (and Why It Fails)," *New York Times*, Nov 2, 2024, https://www.nytimes.com/interactive/2024/11/02/world/middleeast/missile-defense-israel-iran.html.

18 The Iron Dome "was built to stop short-range rockets, and is too slow and limited when it comes to ballistic missiles. For that, Israel relies on several more advanced layers of defenses designed to counter ballistic missiles at different stages of flight." These include the Arrow 2 and 3. Chang and Granados, "How Missile Defense Works (and Why It Fails)."

19 "Each interception costs about $60,000…." "How Israel's 'Iron Dome' works," *The Economist*; Other articles put this cost at about $40,000 per interceptor missile. For example, "…the Tamir inceptor missile costs as little as $40,000 per missile." Michael Ray, Iron Dome, Britannica, updated June 14, 2025, https://www.britannica.com/topic/Iron-Dome.

20 Mara Karlin, "The Return of Total War," *Foreign Affairs* 103, no. 6 (November/December 2024), https://www.foreignaffairs.com/ukraine/return-total-war-karlin; and Michael C. Horowitz, "Battles of Precise Mass," *Foreign Affairs* 103, no. 6 (November/December 2024), https://www.foreignaffairs.com/world/battles-precise-mass-technology-war-horowitz.

21 "Israel launches waves of devastating strikes on Iran," *The Economist*, June 13 2025, https://www.economist.com/middle-east-and-africa/2025/06/13/israel-launches-waves-of-devastating-strikes-on-iran; "The Israel-Iran war is now a brutal test of staying power," *The Economist*, June 15, 2025, https://www.economist.com/middle-east-and-africa/2025/06/15/the-israel-iran-war-is-now-a-brutal-test-of-staying-power; and Natan Odenheimer, Farnaz Fassihi, and Aaron Boxerman, "Live Updates: Israel Launches New Strikes on Tehran," *New York Times*, June 15, 2025, https://www.nytimes.com/live/2025/06/15/world/iran-israel-nuclear.

22 There are different ways of describing these very classic military functions, such as "prepare, sense and understand, decide, and execute faster." Schmidt et al., *Final Report*, 22. See also Ochmanek et

al., *Inflection Point*, ix, xi; and Brose, *The Kill Chain*, 6. The myriad authors of the works surveyed for this report agree that there should be some sort of threshold before autonomous firing would be allowed, but these thresholds vary. For example, Ochmanek states that "we envisage that under normal circumstances the sensing and targeting mesh would have communication links to operations centers so that humans would be overseeing the function of the mesh. Should those links be severed, the mesh could function on its own, identifying targets and directing weapons to them but that would not be the normal mode of operation" (email to author, September 6, 2024). Other authors seem to see robust pre-testing as essential to authorizing certain degrees of autonomy. Consider Brose, who thinks it comes down to "training, testing, and trust" (Brose, *The Kill Chain*, 122–136) or the recommendations of the NSCAI report cited above (Schmidt et al., *Final Report*).

23 A rough estimate is that DOD is spending about $2 billion per year on AI, mostly in research and development. In its 2025 budget, the DOD requests $1.8 billion in artificial intelligence and significant additional funding for AI related efforts, including $1.4 billion for the Combined Joint All Domain Command and Control and resources for the Department-Wide Replicator Initiative to accelerate the delivery of innovative capabilities to warfighters at speed and scale. U.S. Department of Defense, "Department of Defense Releases the President's Fiscal Year 2025 Defense Budget," press release, March 11, 2024, https://www.defense.gov/News/Releases/Release/Article/3703410/department-of-defense-releases-the-presidents-fiscal-year-2025-defense-budget. There is a wide range of estimates for this number, though, depending on definitions; consider this estimate: "The total that the DoD might spend on AI-related contracts if each contract were extended to its fullest terms grew . . . to $4.3 billion in the period leading up to August 2023." Will Henshall, "The U.S. Military's Investments into Artificial Intelligence are Skyrocketing," *Time*, March 29, 2024, https://time.com/6961317/ai-artificial-intelligence-us-military-spending/.

24 Warfare seems to be a complicated enterprise regardless of the technical sophistication of the weapons being used. Consider Agincourt in 1415: "Those in the next rank would have found that they could get within reach of the English only by stepping over or on the bodies of the fallen . . . they would have no choice to do so; yet in so doing, would have rendered themselves even more vulnerable to a tumble than those already felled, a human body making either an unstable platform or a very effective stumbling block for the heels of a man trying to defend himself from a savage attack to his front." John Keegan, *The Face of Battle: A Study of Agincourt, Waterloo and the Somme* (New York: Penguin House, 1976), 101.

25 Decision speed could alter the scope of an attack because of the different psychology needed for a quick decision as opposed to that needed to sustain a protracted decision: "Against defenseless people there is not much that nuclear weapons can do that cannot be done with an ice pick. . . . Nuclear weapons can do it quickly. That makes a difference. . . . It is not in the number of people they can eventually kill but in the speed with which it can be done, in the centralization of decision, in the divorce of the war from the political processes, and in the computerized programs that seem to take war out of human hands once it begins." Thomas Schelling, *Arms and Influence* (New Haven: Yale University Press, 1966), 19–20.

26 See, for example, Ash Carter, "The Moral Dimension of AI-Assisted Decision-Making: Some Practical Perspectives from the Front Lines," *Daedalus* 151, no. 2 (Spring 2022): 299–308, https://doi.org/10.1162/daed_a_01917; and Schmidt et al., Final Report.

27 "The Ukrainian drone strike on bombers far inside Russia on June 1st will be ranked among the greatest military raids in history. The operation, combining old-fashioned sabotage with the iconic weapon of the Ukraine war, illustrated two things. One is that new technology, deployed inventively, can be lethal. The other is that even major powers are vulnerable to attacks on critical infrastructure deep inside their own territory, overturning the assumptions of the 1990s and 2000s." "The West is rethinking how to fight wars," *The Economist*, Jun 3, 2025, https://www.economist.com/leaders/2025/06/03/the-west-is-rethinking-how-to-fight-wars. "The drone attacks that have filled the skies over Ukraine and Russia the past few weeks have not only cemented a new era of warfare, they have also shown Western countries

how ill-prepared they are for it." Lara Jakes, "As Drones Transform Warfare, NATO May Be Vulnerable," *New York Times*, June 4, 2025, https://www.nytimes.com/2025/06/04/world/europe/ukraine-russia-drones-nato.html.

28 "As a technologist as well as a government leader, I believe strongly in both the wonders of AI and the importance of morality in engineering. I am also optimistic this can be solved, like every other hard problem, with diligent technology-informed effort. This will be essential for national defense." Carter, "The Moral Dimension of AI-Assisted Decision-Making," 301.

29 See, for example, Paul Scharre, "Killer Apps: The Real Dangers of an AI Arms Race," *Foreign Affairs* 98, no. 3 (May/June 2019), https://www.foreignaffairs.com/articles/2019-04-16/killer-apps.

30 "A young girl of maybe five or six headed out of the village and up our way, two goats in trail. Ostensibly she was just herding goats, but she walked a long slow loop around us, frequently glancing in our direction. It wasn't a very convincing ruse. She was spotting for Taliban fighters." Paul Scharre, *Army of None* (New York: W.W. Norton, 2018), 3.

31 "ICRC Position on Autonomous Weapon Systems," International Committee of the Red Cross (ICRC), May 12, 2021, https://www.icrc.org/en/document/icrc-position-autonomous-weapon-systems; and "Human Rights Watch Submission to the United Nations Secretary-General on Autonomous Weapons Systems," Human Rights Watch, May 2024, https://www.hrw.org/sites/default/files/media_2024/05/HRW_UN_AWS_052024_3.pdf.

32 Distinction refers to the principle that only active military personnel and military equipment are legitimate targets. Noncombatants and civilian infrastructure should not be attacked. Proportionality requires that legitimate military targets not be attacked if the collateral damage on noncombatants and civilian infrastructure would be excessive when compared to the military benefit of the attack. Humanity means reducing unnecessary suffering through methods such as warning civilian populations before the initiation of hostilities.

33 Office of General Counsel, *Department of Defense Law of War Manual* (Washington, DC: Department of Defense, June 2015, updated July 2023), https://media.defense.gov/2023/Jul/31/2003271432/-1/-1/0/DOD-LAW-OF-WAR-MANUAL-JUNE-2015-UPDATED-JULY%202023.PDF. The "law of war" is often considered to be as follows: distinction (military from non-military targets), proportionality (collateral damage weighed against military benefit), and humanity or distinction (seek mitigation of means and methods and through warnings). This is a summary from Winter, "The Compatibility of Autonomous Weapons with the Principles of International Humanitarian Law," For defense experts who are not lawyers, however, it may come as a surprise to learn there is no specific delineation of "the law of war." They are a subset of customary international humanitarian law which is an evolving set of norms, treaties, rules, and practices. See "Customary International Humanitarian Law Database," International Committee of the Red Cross, https://ihl-databases.icrc.org/en/customary-ihl.

34 Kelley M. Sayler, "International Discussions Concerning Lethal Autonomous Weapon Systems," Congressional Research Service, *In Focus* 11294, February 25, 2025, https://www.congress.gov/crs-product/IF11294

35 Kelly Sayler, "Defense Primer: U.S. Policy in Lethal Autonomous Weapon Systems," Congressional Research Service, *In Focus* 11150, January 2, 2025, https://www.congress.gov/crs-product/IF11150.

36 "Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy," U.S. Department of State, November 9, 2023, https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy-2/; and "DoD Directive 3000.09 Autonomy in Weapons Systems," U.S. Department of Defense, January 25, 2023, https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf.

37 Martin Ford, ed., *Architects of Intelligence* (Birmingham, UK: Packt Publishing, 2018), 76–77, 148.

38 Nils J. Nilsson, *The Quest for Artificial Intelligence*, (New York: Cambridge University Press, 2010), Chapter 3. Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach, Fourth Edition* (London: Pearson, 2020), 14–15: "The modern digital electronic computer was invented independently and almost simultaneously by scientists in three countries embattled in World War II. . . . Since that time, each generation of computer hardware has brought an increase in speed and capacity and a decrease in price – a trend captured in Moore's Law. Performance doubled every 18 months or so until about 2005, when power dissipation problems led manufacturers to start multiplying the number of CPU cores rather than the clock speed. Current expectations are that future increases in functionality will come from massive parallelism . . . "

39 DeepSeek and small language models seem to have changed this verity, as discussed in the next several paragraphs. "LLMs will transform medicine, media and more," *The Economist*, August 13, 2024, https://www.economist.com/schools-brief/2024/08/13/llms-will-transform-medicine-media-and-more.

40 NASEM, *Test and Evaluation Challenges*, 37.

41 "DeepSeek sends a shockwave through markets," *The Economist*, January 27, 2025, https://www.economist.com/business/2025/01/27/deepseek-sends-a-shockwave-through-markets.

42 "The real meaning of the DeepSeek drama," *The Economist*, January 29, 2025, https://www.economist.com/leaders/2025/01/29/the-real-meaning-of-the-deepseek-drama; and Stephen Ornes, "Small Language Models Are the New Rage, Researchers Say: Larger models can pull off a wider variety of feats, but the reduced footprint of smaller models makes them attractive tools," *Wired*, April 13, 2025, https://www.wired.com/story/why-researchers-are-turning-to-small-language-models/.

43 Artificial intelligence, strictly speaking, is a larger category of which machine learning is a subset. AI is a discipline in the labs of academic departments and high-end commercial companies positing and evaluating a wide range of technical strategies to enable computers to perform functions using human-like intelligence in at least a narrow substantive domain. While a variety of strategies are studied in these labs, only machine learning has performed well enough to emerge from the lab into the real world. In the context of this paper, as in most discussions, artificial intelligence and machine learning are used as synonyms.

44 This structure of errors owes a significant debt to Selena Silva and Martin Kenney, "Algorithms, Platforms, and Ethnic Bias: An Integrative Essay," *Phylon* 55, no. 1 & 2, Special Volume (Summer/Winter 2018): 9–37, https://www.jstor.org/stable/26545017, which contains a thoughtful taxonomy of AI problems.

45 Carl von Clausewitz, *On War*, edited and translated by Michael Howard and Peter Paret (Princeton, NJ: Princeton University Press, 1976), 170. Clausewitz takes a dim view of the proposition that many functions above the most applied in war would follow predictable patterns: "It is, of course, obvious that an iron cannonball, impelled by powder to a speed of 1,000 feet per second, will smash any living creature in its path. One needs no experience to believe that. But there are hundreds of relevant details determining this effect, some of which can only be revealed empirically. Nor is the physical effect the only thing that matters: the psychological effect is what concerns us, and experience is the only means by which it can be established and appreciated."

46 Samuel Ackerman et al., "Automatically Detecting Data Drift in Machine Learning Classifiers," *Proceedings of Engineering Dependable and Secure Machine Learning Systems (EDSMLS) Workshop* (November 10, 2021), https://arxiv.org/abs/2111.05672. Part of the difficulty of building robust approaches to manage drift is that it is referred to by different terms in the artificial intelligence literature. Drift is a common problem for machine learning algorithms and is referred to by different names including "drifting data distributions" (Samek et al., "Explaining Deep Neural Networks and Beyond"), "concept

drift" (Gama et al., "A Survey on Concept Drift Adaption"), and "domain drift" (NASEM, *Test and Evaluation Challenges*).

47 On adaptive learning, see Wojciech Samek et al., "Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications," *Proceedings of the IEEE* 109, no. 3 (March 2021): 247, https://doi.org/10.1109/JPROC.2021.3060483; on the regular updating of data, see Joao Gama et al., "A Survey on Concept Drift Adaption," *ACM Computing Surv* 46, no. 4, Article 44 (March 2014): 37, http://dx.doi.org/10.1145/2523813. NASEM, *Test and Evaluation Challenges*.

48 Clausewitz, *On War*, 101.

49 Ibid.

50 Clausewitz, *On War*, 86, 101. "In short, absolute, so-called mathematical, factors never find a firm basis in military calculations. From the very start there is an interplay of possibilities, probabilities, good luck and bad that weaves its way throughout the length and breadth of the tapestry. In the whole range of human activities, war most closely resembles a game of cards."

51 Ibid., 134–136. "Efforts were therefore made to equip the conduct of war with principles, rules, or even systems. . . . All these attempts are objectionable. . . . They aim at fixed values; but in war everything is uncertain, and calculations have to be made with variable quantities. They direct the inquiry exclusively toward physical quantities, whereas all military action is intertwined with psychological forces and effects. They consider only unilateral action, whereas war consists of a continuous interaction of opposites."

52 Ibid., 153. "In short, routine will be more frequent and indispensable, the lower the level of action. As the level rises, its use will decrease to the point where, at the summit, it disappears completely. Consequently, it is more appropriate to tactics than to strategy."

53 Andrew J. Lohn, *Hacking AI: A Primer for Policymakers on Machine Learning Cybersecurity* (Washington, DC: Center for Security and Emerging Technology, December 2020), https://cset.georgetown.edu/publication/hacking-ai/.

54 Clausewitz, *On War*, 136.

55 This capability has been demonstrated and described in an academic paper published by Open AI but has not been incorporated into its publicly available models." Stephanie Lin, Jacob Hilton, and Owain Evans, "Teaching Models to Express Their Uncertainty in Words," *arXiv preprint arXiv:2205.14334* (June 2022), https://arxiv.org/abs/2205.14334.

56 Cade Metz and Karen Weise, "A.I. Is Getting More Powerful, but Its Hallucinations are Getting Worse," *New York Times*, May 5, 2025, https://www.nytimes.com/2025/05/05/technology/ai-hallucinations-chatgpt-google.html. "It's no secret that large language models work in mysterious ways. Few—if any – mass-market technologies have ever been so little understood." Will Douglas Heaven, "Anthropic can now track the bizarre inner workings of a large language model," *MIT Technology Review*, March 27, 2025, https://www.technologyreview.com/2025/03/27/1113916/anthropic-can-now-track-the-bizarre-inner-workings-of-a-large-language-model/.

57 Florence G'Sell, *Regulating Under Uncertainty* (Stanford, CA: Stanford Cyber Policy Center, 2024), 264–290, https://fsi9-prod.s3.us-west-1.amazonaws.com/s3fs-public/2024-12/GenAI_Report_REV_Master_%20as%20of%20Dec%2012.pdf.

58 "These findings highlight the need for a fundamental shift in the design and development of general-purpose artificial intelligence, particularly in high-stakes areas for which predictable distribution of errors is paramount." Lexin Zhou et al., "Larger and more indestructible language models become less reliable," *Nature* 634 (October 2024): 61–68, https://www.nature.com/articles/s41586-024-07930-y.

59 Alexander Bondarenko et al., "Demonstrating specification gaming in reasoning models," *arXiv preprint arXiv: arxiv.org/pdf/2502.13295* (2025), 1, https://arxiv.org/pdf/2502.13295.

60 U.S. Department of Defense "DoD Announces Establishment of Generative AI Task Force," press release, August 10, 2023, https://www.defense.gov/News/Releases/Release/Article/3489803/dod-announces-establishment-of-generative-ai-task-force/.

61 The alignment problem is discussed in Stuart Russell, "If We Succeed," *Daedalus* 151, no. 2 (Spring 2022), https://direct.mit.edu/daed/article/151/2/43/110605/If-We-Succeed. Bengio points to general problems and that there are "machine-learning methods to mitigate such problems," albeit still imperfect: Yoshua Bengio, "AI and Catastrophic Risk," *Journal of Democracy* 34, no. 4 (October 2023), https://www.journalofdemocracy.org/ai-and-catastrophic-risk/. For a broader discussion about "control" issues, consult Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014), 129, 150–153.

62 Bostrom, *Superintelligence*.

63 Russell, "If We Succeed."

64 Such caveats are also recommended by the International Committee of the Red Cross (ICRC), as subsequent section discusses, but this paper would not be as categorical as the ICRC recommends.

65 See a useful discussion of these issues in NASEM, *Test and Evaluation Challenges*, 43–48.

66 "Automation bias" is a well-recognized decisional support problem that has emerged from studies in aviation and healthcare, areas that have traditionally heavily relied on automated tools. Automation bias refers to undue deference to automated systems by human actors that disregard contradictory information from other sources or do not (thoroughly) search for additional information." Saar Alon-Barkat and Madalina Busuioc, "Human-AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice," *Journal of Public Administration Research and Theory*, 2023, 33, 153–169, https://doi.org/10.1093/jopart/muac007.

67 Paul Robinette et al., "Overtrust in Robots in Emergency Evacuation Scenarios," HRI '16: The Eleventh ACM/IEEE International Conference on Human Robot Interaction, 101-108, March 7, 2016, https://dl.acm.org/doi/10.5555/2906831.2906851.

68 Alon-Barkat and Busuioc, "Human-AI Interactions in Public Sector Decision Making," 155.

69 "Following a motivated reasoning logic, this growing body of literature has established that decision-makers are prone to selectively seek and interpret information in light of preexisting stereotypes, beliefs, and social identities. They assign greater weight to information congruent with prior beliefs and contest inputs that contradict them." Alon-Barkat and Busuioc, "Human-AI Interactions in Public Sector Decision Making," 155. Studies show that some judges interpret the same numerical finding more harshly for people of color than for white people when receiving recommendations about bail or sentencing. The study argues that every other factor of the case was held constant so an underlying bias against the people of color could be the only explanation for the higher bail or sentence set in the case of that particular defendant. See Silva and Kenney, "Algorithms, Platforms, and Ethnic Bias."

70 Alon-Barkat and Busuioc, "Human-AI Interactions in Public Sector Decision Making," 156; and Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, "Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks." ProPublica, May 23, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

71 This study outlines the existing literature that demonstrates the presence of automation bias and user bias (called here "selective adherence") but then reports on several small "n" studies that show techniques to mitigate these ill effects. Alon-Barkat and Busuioc, "Human-AI Interactions in Public

Sector Decision Making," 166.

72 This phenomenon can be seen in the case of Dylann Roof. Roof, a white man, murdered eight black members of Mother Emanuel church during a prayer meeting. Roof explained to the FBI after being arrested that he had searched "black on white crime" and was led down an increasingly racist set of nominated pages like the Ku Klux Klan and Neo-Nazi groups. See Safiya Umoja Noble, *Algorithms of Oppression*, (New York: New York University Press, 2018) 110–118.

73 NASEM, *Test and Evaluation Challenges*.

74 This idea was first shared with me by Michael D. Schneider, PhD., associate program leader of the Decision Superiority Laboratory at Lawrence Livermore National Laboratory, in an interview in July 2024.

75 "In the simplest scenario, an AI system could potentially enhance the speed and completeness of the analysis done at the satellite and radar data sites before submission to NORAD. Additionally, AI could be used at NORAD to fuse the two incoming data streams with other Intelligence, Surveillance, and Reconnaissance (ISR) information to add to the confidence of the warning. It has been reported that NORAD already is partnering with the Defense Innovation Unit to create an AI system called Pathfinder to fuse data from military, commercial, and government sensors to help with early warning." Jill Hruby and M. Nina Miller, *Assessing and Managing the Benefits and Risks of Artificial Intelligence in Nuclear-Weapons Systems* (Washington, DC: Nuclear Threat Initiative, August 2021), 16, https://www.nti.org/analysis/articles/assessing-and-managing-the-benefits-and-risks-of-artificial-intelligence-in-nuclear-weapon-systems/.

76 It should be noted that these decision-support algorithms are assumed to not be powered by generative AI but rather to use other types of AI that are significantly less vulnerable to random errors. Decision-support tools, particularly for time-sensitive tasks leading to a firing decision, should not use generative AI at its current stage of technical development, as discussed in the generative AI section of this report.

77 This recommendation is in contrast to the suggestion that AI tools be a supplement to standard test and evaluation (T&E) measures through efforts called verification and validation (V&V). Most analysts recognize that algorithms and other types of software will need more frequent updating than equipment approvals. This additional updating is referred to as verification and validation to underscore that it should occur more frequently than traditional testing and evaluation. Schmidt et al., *NSCAI*; Flournoy et al, *Building Trust through Testing*; and NASEM, *Test and Evaluation Challenges*.

78 Reuters reported the first use autonomous drones by Turkey in Libya in 2021. Winter, "The Compatibility of Autonomous Weapons," 9.

79 Horowitz, "Battles of Precise Mass"; Stacie Pettyjohn, *Evolution Not Revolution: Drone Warfare in Russia's 2022 Invasion of Ukraine* (Washington, DC: the Center for a New American Security, February 8, 2024), https://www.cnas.org/publications/reports/evolution-not-revolution; "How racing drones are used as improvised missiles in Ukraine," *The Economist*, March 24, 2023, https://www.economist.com/the-economist-explains/2023/03/24/how-racing-drones-are-used-as-improvised-missiles-in-ukraine; "How Ukraine uses cheap AI-guided drones to deadly effect against Russia," *The Economist*, December 2, 2024, https://www.economist.com/europe/2024/12/02/how-ukraine-uses-cheap-ai-guided-drones-to-deadly-effect-against-russia; "Fighting the war in Ukraine on the electromagntic spectrum," *The Economist*, February 5, 2025, https://www.economist.com/science-and-technology/2025/02/05/fighting-the-war-in-ukraine-on-the-electromagnetic-spectrum; and "The added dangers of fighting in Ukraine when everything is visible," *The Economist*, February 6, 2025, https://www.economist.com/europe/2025/02/06/the-added-dangers-of-fighting-in-ukraine-when-everything-is-visible.

80 "How drones dogfight above Ukraine," *The Economist*, February 7, 2023, https://www.economist.com/the-economist-explains/2023/02/07/how-drones-dogfight-above-ukraine; and "Ukraine is betting

on drones to strike deep into Russia," *The Economist*, March 20, 2023, https://www.economist.com/europe/2023/03/20/ukraine-is-betting-on-drones-to-strike-deep-into-russia.

81 Paul Mozur and Adam Satariano, "A.I. Begins Ushering in an Age of Killer Robots," The *New York Times*, July 12, 2024.

82 Kateryna Bondar, "Ukraine's Future Vision and Current Capabilities for Waging AI-Enabled Autonomous Warfare," CSIS, March 6, 2025, https://www.csis.org/analysis/ukraines-future-vision-and-current-capabilities-waging-ai-enabled-autonomous-warfare; "Drone Swarm Technologies," Government Accountability Office, GAO-23-106930, September 2023, https://www.gao.gov/assets/gao-23-106930.pdf; and "How drone swarms will defend Britian," *The Economist*, October 15, 2020, https://www.economist.com/britain/2020/10/15/how-drone-swarms-will-defend-britain.

83 Adams, *The Education of Henry Adams*, 496.

84 Carter, "The Moral Dimension of AI-Assisted Decision-Making."

85 Scharre, "Killer Apps."

86 "Humanitarian benefits of emerging technologies in the area of autonomous weapon systems," Office of the Secretary of Defense, U.S. Department of Defense, April 3, 2018, https://ogc.osd.mil/Portals/99/Law%20of%20War/Practice%20Documents/US%20Working%20Paper%20-%20Humanitarian%20benefits%20of%20emerging%20technologies%20in%20the%20area%20of%20LAWS%20-%20CCW_GGE.1_2018_WP.4_E.pdf?ver=O0lg6BIxsFt57nrOuz3xHA%3D%3D.

87 Robert Jervis, "Cooperation under the Security Dilemma," World Politics 30, no. 2 (January 1978): 167–214, https://www.jstor.org/stable/2009958. This phenomenon is called the "security dilemma: many of the means by which a state tries to increase its security decrease the security of others." A classic example is the Stag Hunt, which is described by Jean-Jacques Rousseau to explain the calculations of two hunters as they try to decide whether they trust each other enough to work together and trap a deer, resulting in more meat for each of them, or lack trust and thus decide to hunt alone and each trap a much smaller rabbit.

88 Henry Kissinger wrote: "Thus, the danger of surprise attack was in fact exaggerated by two groups with conflicting objectives: those who wanted substantial defense budgets to protect against the danger of surprise attack, and those who invoked the fear of a surprise attack as a reason for shrinking the defense budgets. Since the issues were so complex, a premium was placed on skills in briefing. And emotions ran so deep that it was not easy to tell whether experts had been led to their conclusions by scientific study or whether they invoked science to support preconceived conclusions - too often the latter. Pity the policymaker who became hostage to the advice of scientists with widely divergent views, and who had devoted more years of study to nuclear issues than the stateman had hours in which to consider them. Debates about such esoteric subjects as vulnerability, accuracy, and calculability attained the complexity of medieval disputes while being, in fact, surrogates for long-standing philosophical disagreements dating back to the earliest days of containment." Henry Kissinger, *Diplomacy* (New York: Simon & Schuster, 1994), 715–716.

89 The content of the "law of war" is complicated. A shorthand description would be distinction (military and non-military targets), proportionality (collateral damage and military benefit), and precaution or humanity (mitigation of means and methods and warnings). The "law of war" is a subset of "international humanitarian law." Some of the debate is whether the responsibilities of militaries using AWS creep pass the traditional three elements and into broader aspects of IHL such as human dignity and compassion. One author observes that "Broadly, there are two positions on the subject. The first is that the deployment of autonomous weapons would be unlawful because such machines would be unable to comply with one or more aspects of IHL. The second is that we should reserve judgement on the matter because it may turn out that autonomous weapons are able to abide by humanitarian

law adequately or, indeed, better than humans in some scenarios." See Elliot, "The Compatibility of Autonomous Weapons with the Principles of International Humanitarian Law."

90 "ICRC Position on Autonomous Weapon Systems," International Committee of the Red Cross; and "Human Rights Watch Submission to the United Nations Secretary-General on Autonomous Weapons Systems," Human Rights Watch.

91 Sayler, "Defense Primer"; "Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy," U.S. Department of State, November 9, 2023, https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy-2/.

92 Kenneth Anderson and Matthew C. Waxman, "Law and Ethics for Autonomous Weapon Systems: Why a Ban Won't Work and How the Laws of War Can," Stanford University, The Hoover Institution Jean Perkins Task Force on National Security and Law Essay Series, 2013, https://scholarship.law.columbia.edu/faculty_scholarship/1803. Thomas W. Simpson and Vincent C. Müller, "Just War and Robots' Killings" *The Philosophical Quarterly* 66, no. 263 (April 2016): 302–322, https://www.jstor.org/stable/24672810; Michael N. Schmitt, "Autonomous Weapon Systems and International Humanitarian Law: A Reply to the Critics," *Harvard National Security Journal* 5 (February 2013), https://harvardnsj.org/2013/02/05/autonomous-weapon-systems-and-international-humanitarian-law-a-reply-to-the-critics/; and WH Boothby, "Highly Automated and Autonomous technologies" in WH Boothby, ed. *New Technologies and the Law in Law in War and Peace* (Cambridge: Cambridge University Press, 2019), 159.

93 See Carter's assessment: "I am also optimistic this can be solved, like every other hard problem, with diligent technology-informed effort." Carter, "The Moral Dimension of AI-Assisted Decision-Making," 301.

94 "ICRC Position on Autonomous Weapon Systems," International Committee of the Red Cross.

95 Ibid.

96 Use of a test database is a classic way to compare the accuracy of different algorithms. For example, most of the AI community viewed neural networks skeptically until an algorithm powered by such a network won a prestigious image recognition competition. The algorithm using a neural network had half the error rate of the closest competitor (which did not use neural networks). The competition used a subset of a large dataset that had been created under the leadership of now Stanford professor Fei-Fei Li. The dataset is called ImageNet and contains 15 million images organized with 22,000 labels. The different competing algorithms were trained on a subset of the database. Each of the algorithms then labeled previously unseen images from the same database. Their accuracy on labeling the inputs were then compared. Martin Ford, ed., *Architects of Intelligence* (Birmingham, UK: Packt Publishing, 2018), 76–77, 148.

97 Deep learning "is a broad family of techniques for machine learning in which hypothesis take the form of complex algebraic circuits with tunable connection strengths. The word 'deep' refers to the fact that the circuits are typically organized into layers, which means that computation paths from inputs to outputs have many steps." Russell and Norvig, *Artificial Intelligence: A Modern Approach*, 653. See also, "The term deep learning refers to machine learning using multiple layers of simple, adjustable computing elements." Ibid., 26.

98 Italics in the original. Spelling of "centre" changed to American spelling. "ICRC Position on Autonomous Weapon Systems," International Committee of the Red Cross, 32.

99 Individuals who are here called "inactive" are more commonly called "hors de combat," or combatants who are now outside of combat because they are wounded, sick, surrendering, or shipwrecked.

100 Ibid.

101 Ibid.

102 Ibid.

103 Carter, "The Moral Dimension of AI-Assisted Decision-Making."

104 Watts and Murray, "Military Innovation in Peacetime." See also the discussion of importance of competitions in AI development, including Image Net and AI City, in NASEM, *Test and Evaluation Challenges*, 42–43.

105 "A good idea rarely wins on its merits alone." Brose, *The Kill Chain*, 227.

106 Posen emphasizes the role of civilian leaders in coming up with new ideas. Barry R. Posen, *The Sources of Military Doctrine* (Ithaca, NY: Cornell University Press, 1984), https://www.jstor.org/stable/10.7591/j.ctt1287fp3.

107 "Symbolically, there was no one of high rank in Defense whose specific job it was to ensure that the commanders and troops in the field had what they needed." Robert M. Gates, *Duty: Memoirs of a Secretary at War*, (New York: Vintage Books, 2014), 116.

108 Ibid.

109 "The goal should be to take advantage of the many deep-seated rivalries in the defense establishment." Brose, *The Kill Chain*, 236. Brose also proposes a competitive process for this purpose but structures the competition differently.

110 Special Operations Command has funding and acquisition authorities under its control of Major Force Program (MFP) 11 to acquire "SOC peculiar" equipment. "Title 10 Authorities," Special Operations Command, https://www.socom.mil/about/title-10-authorities; John Hamre, "Reflections: The Relationship between Special Forces and General Forces" CSIS, Defense360, April 5, 2016, last modified April 5, 2016, https://defense360.csis.org/reflections-relationship-special-forces-general-forces/; and Sydney J. Freedberg Jr., "Artificial Intelligence: Will Special Operators Lead the Way?," *Breaking Defense*, February 13, 2019, https://breakingdefense.com/2019/02/artificial-intelligence-will-special-operators-lead-the-way/.

111 Paul Scharre, *Four Battlegrounds: Power in the Age of Artificial Intelligence* (New York: Norton, 2023), 191–226; RAND Project Air Force, *Addressing Challenges in Defense Acquisition* (Santa Monica, CA: RAND Corporation, 2023), https://www.rand.org/content/dam/rand/pubs/corporate_pubs/CPA2500/CPA2580-2/RAND_CPA2580-2.pdf; Milley and Schmidt, "America Isn't Ready for the Wars of the Future"; and Michèle A. Flournoy, "AI is Already at War," *Foreign Affairs*, November/December 2023, https://www.foreignaffairs.com/united-states/ai-already-war-flournoy.

112 Williamson Murray, "Armored Warfare: The British, French, and German experiences," in *Military Innovation in the Interwar Period*, ed. Williamson Murray and Allan R. Millett (New York: Cambridge University Press, 1996), 17, 24–27; and Barry Watts and Williamson Murray, "Military Innovation in Peacetime," in *Military Innovation in the Interwar Period*, 369–415.

113 Karlin, "The Return of Total War"; and Horowitz, "Battles of Precise Mass."

114 "Replicator," Defense Innovation Unit, https://www.diu.mil/replicator; and Jaret C. Riddick, "Delivering Battlefield Autonomy At-Scale for DoD: Big Announcements from the Pentagon and Congress," CSET blog, https://cset.georgetown.edu/article/delivering-battlefield-autonomy-for-dod/.

115 Scharre, *Four Battlegrounds*. See specifically Chapter 25, "The Wrong Kind of Lethality."

116 Ann Finkbeiner, "The world's most independent defence science advisers," *Nature* 477 (2011): 397–399, https://doi.org/10.1038/477397a.

117 Carter, "The Moral Dimension of AI-Assisted Decision-Making," 307.

118 Ibid., 301.

COVER PHOTO GENERATIVE AI VIA ADOBE STOCK

CSIS | CENTER FOR STRATEGIC & INTERNATIONAL STUDIES

1616 Rhode Island Avenue NW
Washington, DC 20036
202 887 0200 | **www.csis.org**