U.S. ARMY WAR COLLEGE
STRATEGIC STUDIES INSTITUTE

INTEGRATED RESEARCH PROJECT

UNITED STATES ARMY WAR COLLEGE PRESS
Carlisle Barracks, PA

TRUSTING AI:

Integrating Artificial Intelligence into the Army's Professional Expert Knowledge

C. Anthony Pfaff
Project Director

Christopher J. Lowrance
Bre M. Washburn
Brett A. Carey
Researchers

DECISIVE POINT
The USAWC Press Podcast Companion Series
https://ssi.armywarcollege.edu/decisive

STRATEGIC STUDIES INSTITUTE
"The Army's Think Tank"

The Strategic Studies Institute (SSI) is the US Army's institute for geostrategic and national security research and analysis. SSI research and analysis creates and advances knowledge to influence solutions for national security problems facing the Army and the nation.

SSI serves as a valuable source of ideas, criticism, innovative approaches, and independent analyses as well as a venue to expose external audiences to the US Army's contributions to the nation. It acts as a bridge to the broader international community of security scholars and practitioners.

SSI is composed of civilian research professors, uniformed military officers, and a professional support staff, all with extensive credentials and experience. SSI's Strategic Research and Analysis Department focuses on global, transregional, and functional security issues. SSI's Strategic Engagement Program creates and sustains partnerships with strategic analysts around the world, including the foremost thinkers in the field of security and military strategy. In most years, about half of SSI's publications are written by these external partners.

Research Focus Arenas

Geostrategic net assessment – regional and transregional threat analysis, drivers of adversary conduct, interoperability between partner, allied, interagency, commercial, and Joint organizations

Geostrategic forecasting – geopolitics, geoeconomics, technological development, and disruption and innovation

Applied strategic art – warfare and warfighting functions, Joint and multinational campaigning, and spectrum of conflict

Industrial/enterprise management, leadership, and innovation – ethics and the profession, organizational culture and effectiveness, transformational change, talent development and management, and force mobilization and modernization

Trusting AI: Integrating Artificial Intelligence into the Army's Professional Expert Knowledge

C. Anthony Pfaff
Project Director

Christopher J. Lowrance, Bre M. Washburn, Brett A. Carey
Researchers

February 2023

UNITED STATES ARMY WAR COLLEGE PRESS
Carlisle Barracks, PA
STRENGTH—WISDOM

Strategic Studies Institute

Comments pertaining to this publication are invited and should be forwarded to: Director, Strategic Studies Institute and US Army War College Press, US Army War College, 47 Ashburn Drive, Carlisle, PA 17013-5244.

******

ISBN 1-58487-846-0

Cover Photo Credits

Front and Back Covers

Photo: Retinal biometrics technology with man's eye digital remix
Source: www.freepik.com
Author: rawpixel
Website: https://www.freepik.com/free-photo/retinal-biometrics-technology-with-man-s-eye-digital-remix_16016568.htm

Photo: Cruel war scenes, digital painting
Source: wwww.freepik.com
Author: liuzishan
Website: https://www.freepik.com/free-photo/cruel-war-scenes-digital-painting_15174538.htm

Table of Contents

Foreword

This study is the result of two years of observing the XVIII Airborne Corps Scarlet Dragon exercises, which include Project Ridgway, a bottom-up effort to test artificial intelligence (AI) and data technologies and integrate them with legacy targeting processes and systems. During an early iteration of Scarlet Dragon in 2020, then commander of the corps, Lieutenant General Michael "Erik" Kurilla, asked rhetorically, "How do I trust this system?" His point was important. Although soldiers need to trust the systems they employ will function reliably, AI and data technologies introduce complexities simply understanding the technologies will not overcome.

These complexities take multiple forms. First is the nature of AI and data technology itself, which can be a "black box," even to those who have relevant education and training. As a result, humans involved in the process may struggle to account for systems' output. Second, as this study relates, much of the expertise in developing and employing these technologies rests with industry, not the Department of Defense. Thus, vendors are more involved in operating and maintaining systems than has previously been the case. Third, to take full advantage of what AI and data technologies can do, everyone—from commanders and staffs to operators—will require a level of AI and data literacy.

Together, the complexities of AI and data technologies create a tension wherein knowledge critical to employing them is either inaccessible or lies outside the profession. The military's failure to establish professionals' requisite knowledge to ensure AI and data technologies are effective and accountable risks undermining the status of the military as a profession.

To address the challenge of increasing AI and data literacy in the military, this study explores the problem of trust by asking what military professionals need to know to integrate AI and data technologies into the profession's body of expert knowledge. The results of this study should interest readers who want to understand the challenges and opportunities AI and data technologies afford.

Carol V. Evans
Director, Strategic Studies Institute
and US Army War College Press

Executive Summary

Introduction

Integrating artificially intelligent technologies for military purposes poses a special challenge. In previous arms races, such as the race to atomic bomb technology during World War II, expertise resided within the Department of Defense. But in the artificial intelligence (AI) arms race, expertise dwells mostly within industry and academia. Effective employment of AI technology cannot be relegated to a few specialists. Not everyone needs to know how to fly a plane to have an effective air force, but nearly all members of the military at every level will have to develop some level of AI and data literacy if the US military is to realize the full potential of AI technologies. Thus, a critical component of future readiness will be the *AI literacy* of the force.

In this context, *AI literacy* means more than simply understanding how to use, design, and engineer AI- and data-enabled systems. Rather, data, algorithms, and the systems they support interact in complex ways that change even familiar processes, such as targeting, into something much more complicated and unfamiliar. Making matters more difficult, from a professional perspective, mastering new technology requires adequately understanding how the technology works and how its application affects organizational, ethical, and political concerns for the military and the US government, its international partners, and American society.

Challenge of Integrating AI and Data Technologies

Often, the problems associated with employing AI, especially in a lethal targeting process, arise from the perceived trade-off between taking advantage of the machine's speed and maintaining meaningful human control. To the extent humans give up control, they give up responsibility. To the extent they give up responsibility, they undermine accountability, and undermining accountability creates reasons to distrust the machine and the humans who employ it.

Thus, the central question is: On what basis commanders, staffs, and operators can trust AI technologies and the systems they enable? *Trust*, as used here, entails multiple conditions. First, one expects the system to be effective—that is, able to produce the intended effect at least as well

as, if not better, than human-only systems. Moreover, as a report by the UN Institute for Disarmament Research pointed out, AI-enabled systems must be predictable and understandable, where *predictability* entails the system consistently fulfilling its intended purpose and *understandability* entails the machine acting for intelligible reasons. In a professional context, however, professionals trusting the technology is not enough. Clients must further trust professionals to use AI in their interests and in a way that reflects their values and other ethical commitments.

Given professionals must ensure these conditions are being met, the question can be reframed as one of professional expertise, which includes educating, training, and certifying the profession's members to use the technology and evolving the profession's institutions to ensure the technology's use is effective and ethical. Knowing how the acquisition of new technologies impacts the profession's organizational culture and other stakeholders is also critical to meeting the conditions for trust.

To understand how the military can meet these conditions, this project examined Project Ridgway, an effort by the XVIII Airborne Corps to integrate currently available AI, data, and imagery to be *AI-ready*. Project Ridgway is a bottom-up effort wherein the corps engages the private sector directly to take advantage of commercially available data and algorithms to support targeting in the deep fight. This report found trusting an AI-driven system in the professional military context requires: first, understanding the context in which AI is applied; second, understanding what one is trusting AI to do; and, finally, understanding how to interact with the AI-driven system, including how the system receives input and provides output. Meeting these conditions enables one to audit and ensure the authenticity of the data, which is critical for trust.

Targeting: Why Speed Matters

In this context, targeting is a four-phase process that comprises deciding, detecting, delivering, and assessing. As currently employed in the XVIII Airborne Corps's targeting process, AI primarily applies to the detect phase, wherein sensors provide input (generally, imagery) to an algorithm, which relies on curated data to predict whether designated objects are present and, if so, their location. In the future, AI may also impact other parts of the cycle, including asset allocation and the assessment of battle damage and effects.

Targeting is iterative and interactive. The process iterates by learning during each cycle and the cycles within the larger targeting cycle to improve the AI-driven machine's performance. Targeting involves interacting with an adversary engaged in the same cycle. If an adversary is similarly equipped, the one who gets through the cycle faster wins. Since machines are faster than humans, targeting disposes humans to rely on the machine, even if doing so means taking extra risks. Speed matters.

Developing Trustworthy AI

Given this reliance on machines, one must ask oneself what one is trusting an AI-driven system to do. From a practical and an ethical perspective, lethal targeting requires one to balance the imperatives of defeating an enemy, avoiding noncombatant casualties, and protecting the force. Balancing these imperatives involves answering questions about risk. Put simply, lethal operations expose friendly combatants and noncombatants to risk, avoiding noncombatant casualties exposes friendly combatants or the operation to risk, and protecting the force exposes the operation or noncombatants to risk. Reducing risk to any one imperative thus places risk on the other two. Employing AI can reduce risk to all three. By making fires faster and more precise, AI makes defeating the enemy more likely while reducing the chance of friendly and collateral harm.

In a human-only process, trust depends on understanding the capabilities of one's soldiers and the weapons they carry, ensuring they understand and will comply with the law of armed conflict, and being able to hold them accountable when they do not comply. In an AI-driven process, trust depends on knowing how to curate and monitor data, assess and optimize algorithm performance, and secure the system from external manipulation. Artificial intelligence is a process of algorithms operating on data in a specific context. Trusting this process depends, at least in part, on trusting the components. To ensure trust, the data must be auditable and the algorithm adequately understood in its operational context.

Barriers to Trusting AI

Barriers to trusting AI include uncertainty about how to warrant confidence one has curated data correctly, trained and retrained the data and algorithms to be accurate and precise, and protected the system against spoofing or other unwanted manipulation.

Data Challenges

In the context of AI, multiple other factors that are functions of the data, the algorithm, and external interference impact trust. Algorithms are often only as good as the data on which they are trained. Through training, the machine learns to differentiate items of interest from everything else. Collecting accurate, complete, consistent, and timely data sets for the system to train on is extremely difficult and sensitive to the environment in which the targeting will occur. Keeping data sets updated is critical work that must be ongoing. The challenge is that it is extremely difficult to know when one has collected all the necessary data to optimize the system's performance. As a result, the system will make mistakes when the inputs do not closely resemble the data on which the system was trained.

Performance Issues

Performance issues usually come in the form of misclassifications, false positives, and false negatives. For example, when the inputs to AI classifiers do not resemble the training data, prediction mistakes are more likely. Prediction mistakes can occur when a classifier is trained using only images of targets taken during the summer months and then presented images of partially concealed targets taken during the winter. If a classifier that was trained using only images of tanks operating in the desert is asked to classify an image of the tank partially covered in snow, then the classifier will likely make a mistake. To counter such mistakes, continuously searching for and collecting new, informative data examples as they become available and using them to retrain and update the classifier as needed—especially relative to the environment in which one is operating—is important. Often, retraining and updating the classifier means collecting new data while the system is operating and then identifying which samples can help to improve the AI model's performance.

In short, classifiers can make mistakes given the state of the art and the difficulty of collecting comprehensive data sets. Artificial intelligence can be a "black box" because how it arrives at an output is not always discernible to humans, either due to the complexity of the algorithm or because the AI's output depends on the strength of the connections in the network. Nevertheless, commanders and operators should understand the limitations of AI and observe AI-enabled systems' performance in similar conditions, thereby enabling the commanders and operators to decide, based on risk calculations, how much control to provide to the AI in targeting operations.

Other issues include the enemy actively attempting to thwart AI by poisoning data sets or changing the enemy's asset signatures. For instance, a poisoning attack can undermine a machine learning model during the training phase by altering the model's training data. Adversarial poisoning attacks could train a target identification model to ignore one class of object entirely, enabling a high-value target to hide in plain sight. To conduct an input attack, an adversary injects noise into a model's input to produce an incorrect output.

In one example, a small piece of tape placed on a stop sign caused self-driving cars to misidentify the sign as a 60-mile-per-hour speed marker. Similarly, an adversary could visually modify a tank so a machine-learning model assesses the tank as a truck. Moreover, doing so would not be difficult. Small pixel changes, invisible to the human eye, have caused classification algorithms to misidentify images of pandas as monkeys. Both types of attacks, input and poisoning, can undermine the perceived effectiveness of fielded models and degrade trust. More to the point, one should expect AI-driven systems to be under constant attack, requiring users to find ways to detect the effects of these types of attacks.

Taken together, the sensitivity of the data sets, the complexity of the algorithms, and the potential for undetected sabotage give rise to an accountability gap. Accountability depends on intent and action. But harm, including violations of the law of armed conflict, may occur, despite commanders, staffs, and operators involved in an AI-driven system acting with good intentions and despite the system, with the exception of spoofing, working according to specification. Commanders and staffs may understand the system well but suffer from automation bias, especially with systems that are normally reliable, thus increasing the probability of unaccountable harm.

Importantly, AI performance is not all about speed. In fact, the machine provides better output when humans interact with it, even during operations. So, the idea that developing and employing AI involves a trade-off between speed and meaningful human control is a false dilemma. The question, then, is how do humans know when and where to interact with a system and provide control while optimizing the system's performance?

Developing Reliable and Capable Systems

Trust and risk are central concerns in developing reliable, capable systems. Commanders need a reliable way to know when AI can be trusted and when to execute some stages of the targeting process with less supervision for the benefit of speed but at the cost of more risk. The systems studied here

rely on neural networks that provide a measure of probabilistic confidence in each target classification. Commanders can exploit these neural networks during targeting to make informed decisions about the level of human supervision required, especially when the probabilistic confidence is combined with other information, such as the commander's risk tolerance in the context of the mission.

The commander's risk tolerance can aid in the process of deciding how to handle targets that have been classified by AI. Determining the acceptable level of risk for the operation of the AI is the commander's decision. Therefore, the commander should be given the flexibility and option to assume more risk at times if, based on his or her best judgment, the conditions merit the risk.

For instance, a commander may be risk averse when providing fire support in a counterinsurgency mission or in a dense urban environment with many civilians nearby. But a commander may be more risk tolerant if facing a high-intensity battle in mostly open terrain or performing final protective fires when friendly forces may be overrun by the enemy. To capture risk tolerance, commanders could be given a rheostat-like device that they can tune and use to convey their risk tolerance directly to the system. One can also run more than one AI model at a time; this approach, which is commonly referred to as an *ensemble*, can be used to increase confidence that inferences drawn are true or to detect errors.

Decision-making Logic within the Control System

The rheostat would interact with the system through a fuzzy-logic controller that would account for commander risk tolerance and machine certainty to determine the optimal setting for human control. Fuzzy logic can help balance machine confidence and a commander's risk tolerance. Fuzzy logic's purpose is to avoid hard coding single-value thresholds, which specify where certain values belong to certain categories (for example, 34 is moderate, and 32 is low). Rather, the idea is to program transitions between the input classes of low, moderate, and high.

Programming transitions between input classes makes fuzzy logic more tolerant of uncertainty when measuring and quantifying the inputs into linguistic sets. The regions where the moderate set overlaps with either the low or high set are the ranges where the input would be classified as belonging to multiple sets with partial membership in each, such as 80 percent high and 20 percent moderate. For instance, one could

program a freezer thermostat's controller to alert one to intervene to lower the temperature.

Given two variables (risk and certainty) and three settings (low, medium, and high), a rule base of nine recommended settings for human oversight would logically exist. The rule base would be programmed into the controller's memory using a series of if/then statements and obey the following logic: "If AI's Classification Confidence is low and Commander's Risk Tolerance is low, then human involvement is maximum. If AI's Classification Confidence is high and Commander's Risk Tolerance is high, then human involvement is minimum." Assuming two inputs with three categories each (low, moderate, and high), the complete set of nine rules can be derived by the two-dimensional rule base, expressed by machine-generated probabilities and the commander's risk tolerance.

Risk Profiles and Adaptive Teaming Based on Fuzzy-logic Controllers

What does this rule base mean in practice? The controller's decision for maximum involvement implies a human-driven targeting process in which humans lead each step. Using a human-driven targeting process does not preclude AI from assisting in these steps. In other words, AI can augment any step, but a human must explicitly verify the output before the target proceeds. On the opposite extreme, minimum involvement translates into AI automating all steps, except for the final validation and authorization process, wherein a leader in the fires cell would review the targeting information and recommendations before giving the order to proceed with a fire mission. The moderate oversight process flow is more nuanced and similar to the minimum oversight process flow, except the classification confidence of the AI algorithm and the risk assessment from the integration stage must meet stringent thresholds. If a threshold is not met in either case, then a human must inspect the output generated by the AI algorithm.

Human Development

The technical component shows soldiers will have to develop varying degrees of AI and data literacy. For this to happen, the US Army must identify what this literacy entails and how to certify it. Although identifying the varying degrees of AI and data literacy falls under the technical component, determining how to recruit, certify, and manage knowledgeable personnel will become a critical professional task. To remedy the lack of personnel

with AI and data-science education and skills, the Army has implemented plans to educate selected personnel at the leader, analyst and engineer, and technician levels. Although necessary, these plans may not be adequate to provide the range of skilled personnel required to proliferate capabilities at the corps level Army-wide, especially in the short term.

Part of the reason the Army's existing plans may be inadequate is the Army needs soldiers with the right data and AI skills and leaders who know how to employ data and AI skills effectively. Thus, the Army should also consider integrating AI and data literacy into commissioning and other entry-level education and training.

Further complicating matters, the Army's ability to manage personnel who are skilled in science, technology, engineering, and mathematics in general, much less those with AI and data-related skills, is limited. Indeed, without a more efficient management system, optimizing the assignment of personnel trained by the Army's new educational programs may not be possible, especially at the operational level. Effectively assigning newly trained personnel is critical to taking advantage of new, often commercially available technologies so the Army remains agile relative to its adversaries. Optimizing the Army's talent management will require the service to revise how it identifies educational requirements, aligns talent with operational needs, and tracks talent so personnel are available where they are most needed.

This study recommends the Army create new skill identifiers to improve the tracking of AI- and data-related expertise, consider establishing a technology corps that would be managed much like the logistics corps to provide expert knowledge where and when it is needed most, and code certain positions for more than one skill to increase assignment flexibility.

Ethical

From an ethical perspective, targeting requires preventing, or at least mitigating, potential harm to noncombatants as well as friendly forces. Given the potential for friendly and noncombatant casualties, especially in large-scale combat operations, professionals will have to ensure application of the technology represents acceptable risk to protected persons, infrastructure, and other material assets. Artificial intelligence also raises questions of accountability. Having machines play a larger role in decision making may result in bad outcomes, even if both the humans and the machines perform their duties correctly. Understanding how to deal with such outcomes will be critical to applying AI. Here, we might measure success in terms of whether AI-enabled outcomes represent less harm than human-only processes.

To meet the requirements for ethical targeting, commanders must ensure staffs and operators are capable of curating and training data and that they do so at appropriate intervals to ensure the system performs as well as a human-only system. Staffs and operators must also develop familiarity with systems to the point that the staffs and operators can explain outcomes intelligibly. Introducing an interface, like the fuzzy-logic controller discussed above, would facilitate meeting the requirements for ethical targeting and allow commanders to take greater advantage of machine speeds without losing the kind of control that might give rise to ethical failure. The interface addresses accountability by making commanders accountable for the accuracy of their risk assessments and ensuring data is properly curated for the context in which commanders employ the algorithm. The interface also addresses automation bias because it provides humans a way to know when the machine itself is, in a sense, uncertain about its output. Whether these measures are good enough depends on how well the system balances the ethical imperatives discussed earlier in comparison to a human-only process. Balancing imperatives is ultimately the responsibility of the humans involved in the targeting process.

One can further improve the system's ability to avoid collateral harms—and thus perform ethically—by training data to identify legitimate targets and illegitimate targets (such as hospitals and schools). For example, if the machine could produce a result such as "80 percent tank; 10 percent school bus," the machine could alert commanders and staff that even though the target probability was within the commanders' risk tolerance, they may have additional reasons for scrutiny. Building data sets that can account for legitimate and illegitimate targets may be beyond the resources available in any given system. In these cases, commanders should account for the likelihood of illegitimate targets in their risk assessments.

Political

Political-cultural knowledge requires knowing how the use of an emerging technology will affect public expectations about the use of force, how these expectations affect society's perception of military service, and how other Department of Defense efforts to employ the emerging technology affect one's own efforts. Moreover, political-cultural knowledge requires senior military leaders to understand how shifts in public expectations will affect civil-military relations and military culture because public expectations will affect who joins the military and how they serve.

To the extent that using technology reduces risks to soldiers and noncombatants, doing so reduces the political risks associated with using

force. Thus, senior military leaders will need to manage senior civilian leaders' expectations to ensure using technology does not risk escalation into a wider conflict. In addition, senior military leaders will need to manage public expectations about collateral harms to ensure the public's support. Perhaps most importantly, senior military leaders will need to manage expectations about the effectiveness of the technology so civilian leaders do not rely too much on technology and the public does not become frustrated by a lack of results. The public is not likely to trust a military that cannot deliver results and that imposes risks on soldiers and noncombatants alike.

Conclusion

Developing and employing new military technologies is a part of being a military professional. Indeed, military history is a story of technological innovation and soldiers learning how to operate new systems. Many aspects of integrating AI are not new. Artificially intelligent technologies' capability to improve a wide range of military weapons, systems, and applications differentiates this type of technology from others. As this technology expands in application, war will be as much about managing data as it is about managing violence. Thus, commanders of the near future will need to understand how AI-enabled systems will interact with the commanders' judgments about risk to friendly forces and noncombatants. Commanders will also need to know how to ensure staffs and operators can curate and train data effectively. Finally, commanders and staffs will gain experience interacting with the private sector, which will increasingly be relied upon for AI and data technology and aspects of its operation.

— 1 —

Introduction: Professional Expert Knowledge

Integrating AI into the Army's Professional Knowledge

Integrating artificially intelligent technologies for military purposes poses a special challenge. In previous arms races, such as the race to atomic bomb technology during World War II, expertise resided within the Department of Defense (DoD). But in the artificial intelligence (AI) arms race, expertise dwells mostly within industry and academia.[1] Moreover, effective employment of artificially intelligent technology cannot be relegated to a few specialists. Not everyone needs to know how to fly a plane to have an effective air force, but almost everyone will have to develop some level of AI and data literacy if the US military is to realize the full potential of AI technologies. Thus, a critical component of future readiness will be the "AI literacy" of the force.

In this context, "AI literacy" means more than understanding how to use, design, and engineer AI- and data-enabled systems. Rather, algorithms, data, and the systems they support interact in complex ways that change even familiar processes, such as targeting, into something much more complicated and unfamiliar. Making matters more difficult, from a professional perspective, mastering new technology requires adequately understanding

Acknowledgments: The authors of this study would like to acknowledge the following subject-matter experts for their assistance. We would first like to thank Colonel Joseph M. O'Callaghan, director of the fire support coordination cell for XVIII Airborne Corps, and his team for their willingness to lend his considerable expertise and time to our efforts. In addition, we would like to thank Dr. Kathleen Moore of the US Army War College, who provided critical assistance in the data science part of the study, and Colonel Elliot Harris, who contributed to the sections on targeting. Finally, we would like to thank the members of Project Maven, who also shared their expertise.

1. Darrell M. West and John R. Allen, *Turning Point: Policymaking in the Era of Artificial Intelligence* (Washington, DC: Brookings Institution Press, 2020), 139.

how the technology works and how its application affects organizational, ethical, and political concerns for the military as well as the US government, its international partners, and American society.

Making matters more urgent, the National Security Commission on Artificial Intelligence's *Final Report* states despite the "world altering" impact of AI, the US government "is not organizing or investing to win the technology competition against a committed competitor, nor is it prepared to defend against AI-enabled threats and rapidly adopt AI applications for national security purposes."[2] Perhaps more to the point, the report points out that without AI technologies, defending against AI-enabled adversaries who can operate at "machine speeds" is "an invitation to disaster."[3] To avoid this result, the US Army has set 2035 as the deadline for successfully integrating AI and other technologies so the service can prevail throughout the spectrum of competition.[4]

The importance of AI and data technologies is not lost on US adversaries, such as China. In 2017, the Chinese government released its *New Generation Artificial Intelligence Development Plan*, which declared China's aim to become the world center for AI innovation in a broad range of sectors, including defense. Moreover, the plan seeks to harness China's government and private-sector research to serve the government's strategic ends. Companies that sign on can get preferential bidding, access to financing, and, possibly, market-share protection. On defense, the plan specifically states China seeks to use AI to make "radical breakthroughs" in military technology, which would compensate for China's lack of spending relative to adversaries like the United States.[5]

Despite China's emphasis on and advantages in AI, including its access to abundant data, the People's Liberation Army faces its own challenges taking advantage of these resources. These challenges include a fragmented bureaucracy and the requirement to clean and label data from disparate sources for use, potentially limiting its utility. Still, China's access to inexpensive data services and centralized control suggests, in times of crisis, Beijing could

2. Eric Schmidt and Bob Work, *Final Report* (Washington, DC: National Security Commission on Artificial Intelligence, 2021), 8.

3. Schmidt and Work, *Final Report*, 9.

4. Headquarters, Department of the Army (HQDA), *Army Multi-domain Transformation: Ready to Win in Competition and Conflict*, Chief of Staff Paper no. 1 (Washington, DC: HQDA, March 16, 2021).

5. Huw Roberts et al., "The Chinese Approach to Artificial Intelligence: An Analysis of Policy, Ethics, and Regulation," *AI and Society* 36 (2020): 60–62.

achieve an advantage over the United States.[6] Thus, to ensure (if not reclaim) its advantage, the United States must integrate its data and AI efforts more effectively than adversaries like China and Russia can.

Although top-down acquisition systems driven by service acquisition offices will undoubtedly provide more advanced technology, technology is available now that can give US forces a more immediate advantage and set conditions for AI literacy in the future. Beginning in the summer of 2020, XVIII Airborne Corps initiated Project Ridgway, which is intended to field AI and data technologies and develop an organizational culture to optimize these technologies' performance. The project is organized around four lines of effort: organizational culture, a data-literate workforce, data management and governance, and enabled infrastructure.[7] Along these lines of effort, the authors observed XVIII Airborne Corps's efforts to address the normalization, structuring, labeling, and classification of data as well as challenges associated with collection, targeting, and communication. The corps's efforts raised other challenges associated with exercising command responsibility, managing talent, and engaging vendors who play an important role in developing and applying AI and data technologies.

To develop and apply AI at scale, the challenge for the corps and, more broadly, the Army as a profession is to integrate AI into the full scope of their combat, combat support, and combat service support operations. While integrating AI, the Army must also maintain the trust of its client—in this case, the American people and the government that represents them. To rise to this challenge, Army leaders must first trust the technology themselves. To gain this trust, the leaders must first understand how the use of AI impacts the technical, human developmental, ethical, and political components of expert knowledge and the subsequent barriers to trust that the impacts may generate. Overcoming barriers to trust will require understanding human-machine teaming, the curation of data, talent development, and the governing of the technology's application and evolution. Finally, to overcome barriers to trust, Army leaders must address multiple stakeholders' concerns that will affect the Department of Defense internally as well as US civil-military relations externally.

6. Elsa B. Kania, "Artificial Intelligence in China's Revolution in Military Affairs," *Journal of Strategic Studies* 44, no. 4 (2021): 24.

7. Jackson Barnett, "How One Corps Is Trying to Modernize the Army," FedScoop (website), June 21, 2021, https://www.fedscoop.com/how-one-corps-is-trying-to-modernize-the-army/; and "XVIII Airborne Corps: Project RIDGWAY," All Partners Access Network (website), February 17, 2021, https://wss.apan.org/army/PROJECTRIDGWAY/Public/default.aspx.

Professions and Expert Knowledge

Optimizing AI technologies means greater dependence on the technologies, which can operate much faster than humans can effectively monitor and intervene. Greater dependence entails ceding at least some human control, which comes with certain risks. To decide how much control one wants to cede, one also must decide how much risk to tolerate. To make this decision, soldiers at all levels must adequately understand the technology necessary for the Army profession to fill its role.

A profession entails specialized knowledge in service to society that allows professionals to exercise autonomy over a specific jurisdiction.[8] The medical profession, for instance, involves specialized knowledge about human health that medical professionals apply to sustain or improve their clients' health. Because they have autonomy over a specific jurisdiction, doctors are allowed to prescribe drugs, conduct surgery, and act in ways nonprofessionals cannot.[9]

Moreover, one can only become a professional by being certified by other professionals. Doctors, nurses, and other medical professionals are certified by attending medical school and advancing through additional education and training as they progress within their chosen specialties. Finally, professions have codes of ethics to ensure professionals' practices serve a greater good.[10] Again, the medical profession is instructive because its codes obligate competency, compassion, and provision of care, among other things necessary for the medical profession to fulfill its role.[11]

Samuel Huntington famously characterized military expertise as the "management of violence."[12] In addition to tactical skill, the management of violence requires organizing, training, and equipping the force and planning and directing its operations and activities both in and outside combat.[13] Critical to a profession's health is the client's trust, which requires professionals to put their clients' needs over their own, at least when providing the

8. Samuel P. Huntington, *The Soldier and the State: The Theory and Politics of Civil-Military Relations* (Cambridge, MA: Belknap Press, 1957), 8–10.

9. Andrew Abbott, *The System of Professions: An Essay on the Division of Expert Labor* (Chicago: University of Chicago Press, 1988), 60.

10. Huntington, *Soldier*, 9–10; and Allan R. Millett, *Military Professionalism and Officership in America* (Columbus: Mershon Center of the Ohio State University, 1979), 3.

11. "Code of Medical Ethics Overview," American Medical Association (website), n.d., accessed on July 8, 2021, https://www.ama-assn.org/delivering-care/ethics/code-medical-ethics-overview.

12. Huntington, *Soldier*, 11.

13. Huntington, *Soldier*, 11.

professional service.[14] Without this trust, clients will typically look elsewhere for service, thus undermining the profession's jurisdiction. Clients may also seek to impose external regulation and oversight, undermining the profession's autonomy. The military departs from professions like law and medicine, the clients of which are typically individual members of society. Because its client is the state, the military must provide expert advice on the application of military force in defense of the society the state represents.[15]

According to Don Snider, Gayle Watkins, and Richard Lacquement's work, which was integrated into Army Doctrinal Reference Publication 1, *The Army Profession*, Army expert knowledge consists of four components: military-technical, human development, ethical, and political-cultural. Technical expertise denotes ensuring the profession is effective. Human development involves, at least in the military's case, recruiting the right people to serve and providing them the professional development necessary to become effective, certified leaders. Ethical expertise determines the norms that govern the service the profession provides and ensures the norms align, at a minimum, with clients' values, international law, and other relevant norms. Finally, political expertise, which includes cultural knowledge, covers how professions interact with external actors—which, in the case of the military, includes the US government, the American people, and partners and civilian populations where the military operates.[16]

Clients rely on professionals because clients do not have the expertise to provide a service themselves or assess whether professionals have provided the best service they could have. For clients' trust to be plausible, they must rely on other factors in the relationship, such as reliability, to assess trust. But reliability requires time, which is often absent in client-professional relationships. Thus, for trust to be plausible, clients must have a normative view of the profession, which is a belief that professions obligate professionals, among other things, to put clients' needs first.[17] For the military, trust shapes civil-military relations: The trust of the American people and the government that represents them impacts the roles the military plays, the resources it is given, and, perhaps most importantly, who joins it. Artificial intelligence

14. Millett, *Military Professionalism*, 3.

15. Huntington, *Soldier*, 11–18.

16. Richard Lacquement, "Mapping Army Professional Expertise and Clarifying Jurisdictions of Practice," in *The Future of the Army Profession*, ed. Don Snider and Lloyd Matthews, 2nd ed. (New York: McGraw Hill, 2005), 214–17; and HQDA, *The Army Profession*, Army Doctrinal Reference Publication 1 (Washington, DC: HQDA, 2015), 5-1.

17. Anne C. Ozar, "The Plausibility of Client Trust of Professionals," *Business and Professional Ethics Journal* 33, no. 1 (2014): 94–95.

can complicate matters by removing factors that are critical to professional judgment. For example, for patients to trust a doctor who uses AI to provide diagnoses, the patients must trust the doctor, trust the doctor can assess the machine's performance, and trust the machine is operating effectively. If the doctor does not fully understand the machine or critical aspects of the diagnosis fall outside the medical profession, then trust may not be warranted.

But the need for trust extends beyond the client. Professionals also must trust each other. In the case of the Army, the reason is clear. At the operational and tactical levels, the Army stands as an essential pillar of command-and-control doctrine at both the Joint and service levels.[18] Military doctrine also underscores the bidirectional nature of trust, defining *mutual trust* as "shared confidence between commanders, subordinates, and partners that they can be relied on and are competent in performing their assigned tasks."[19] Trust enables mission command, allowing leaders to delegate appropriate levels of decision making and execution to subordinates and freeing commanders to focus on decisions only they can make.[20] To optimize trust at any level, one must optimize trust at every level.

If the expertise to develop algorithms and curate data lies largely in the private sector, critical expertise may lie outside the profession. If outcomes depend on machine thinking that is inaccessible to humans, some expertise may lie outside the human. If professionals cannot access critical expertise, they risk ceding some autonomy and, consequently, jurisdiction, and such a cession impacts clients' trust in professions.

Professional Expert Knowledge, AI, and Trust

Maintaining trust in AI and data technologies first requires an understanding of that which one is trusting the technology to do—in this case, ethical targeting in large-scale combat operations (LSCO). Targeting synchronizes assets to generate effects while relying on a "targeting methodology [that] is a rational and iterative process that methodically analyzes, prioritizes, and assigns assets against targets systematically to create those effects that will contribute

18. Joint Chiefs of Staff (JCS), *Doctrine for the Armed Forces of the United States*, Joint Publication 1 (Washington, DC: JCS, updated July 12, 2017), xxiii; and HQDA, *Mission Command: Command and Control of Army Forces*, Army Doctrinal Publication 6-0 (Washington, DC: HQDA, 2019), 1-6–1-8.

19. HQDA, *Mission Command*, 1-8.

20. HQDA, *Mission Command*, 1-14.

to achieving the commander's objectives."[21] For the purposes of this discussion, targeting is a lethal process that employs Joint assets to destroy enemy weapons and equipment on corps' high-value target (HVT) lists.

The targeting process comprises four steps: decide, detect, deliver, and assess. During the decide phase, commanders provide guidance, and staffs provide recommendations and proposals to determine target priorities and match assets to targets identified in the attack guidance matrix. The detect phase synchronizes sensors to identify and track approved targets throughout the battlefield. The detect phase focuses on managing manned and unmanned platforms to find and fix enemy targets so they can be finished in the deliver phase. The deliver phase requires matching an approved target to an asset that has sufficient, but not excess, capability to destroy the target.

One should avoid using one's most capable systems so they are available should more hardened targets appear. During the deliver phase, observers identify the target and initiate a fire mission to be transmitted through the Advanced Field Artillery Tactical Data System to an available shooter. This process is time consuming compared to AI-enabled processes due to the number of potential interventions from sensor to shooter that may generate recalculations or other delays in the process. Finally, during the assessment phase, observers determine whether the targeting objectives have been achieved. Simultaneously, the staff must also assess any unintended or unanticipated outcomes and adjust plans and operations accordingly.[22]

Because the targeting process is interactive with an adversary, speed matters. To the extent the adversary is similarly equipped, the one who gets through the targeting cycle faster has an advantage. Moreover, as the enemy engages in its targeting cycle, the target that can be selected and the way in which it can be selected can change. Targets are grouped into two categories: deliberate and dynamic. Deliberate targets are developed over time, with both reconnaissance and strike assets apportioned to them. Dynamic targets are unanticipated "targets of opportunity," for which no clear priority or dedicated asset exists. Sorting dynamic targets can take time, which can give the enemy time to get out of range, find cover, or otherwise reduce its vulnerability.

Artificial intelligence and data technologies can play an important role in expediting the targeting process. They rapidly sort through large

21. HQDA, *Targeting*, Army Techniques Publication 3-60 (Washington, DC: HQDA, 2015), 1-2.

22. HQDA, *Targeting*, 1-2.

amounts of sensor data to discover, verify, or refine potential targets and can sort through the complex factors associated with matching assets to targets. But as of this writing, AI and data technologies primarily impact the detect phase, in which information from various sources, including imagery intelligence, signals intelligence, electronic intelligence, and measurements intelligence, are fused together to identify targets. Without augmentation from AI, this fusion requires multiple intelligence professionals to collaborate and make recommendations based on their respective skills and training. But the difficulty is knowing how to assess the machine output without replicating the time-consuming human involvement the machine is supposed to displace.

From a practical and an ethical perspective, lethal targeting requires one to balance the demands of defeating an enemy, avoiding noncombatant casualties, and protecting the force. Balancing these imperatives involves answering questions about risk. Put simply, lethal operations expose both friendly combatants and noncombatants to risk, avoiding noncombatant casualties exposes friendly combatants or the operation to risk, and protecting the force exposes the operation or noncombatants to risk. When using current systems, reducing risk to any one imperative often places risk on the other two. Employing AI can reduce risk to all three. By making fires faster and more precise, AI makes defeating the enemy more likely and reduces the chance of friendly and collateral harm. Artificial intelligence can also assist in prioritizing targets and resources, further lowering risk to operations. The question, then, is: Under what conditions can one trust an AI-enabled system?

In a human-only process, trust depends on understanding the capabilities of one's soldiers and the weapons they carry, ensuring they understand and will comply with the law of armed conflict, and being able to hold them accountable when they do not. In an AI-driven process, trust depends on knowing how to curate and train data, assess and optimize performance, and secure the system from external manipulation.

The technology must also be able to adapt to the demands of future conflict. For instance, unlike counterterror or counterinsurgency operations in the past, which involved targeting individuals or small groups, LSCO will require engaging thousands of targets a day at rates that must be faster than the enemy's. Attaining such speeds will require the introduction of AI technologies as well as networking the technologies with space-based sensors and networked command-and-control systems. If introduced and networked successfully,

AI technologies will enable militaries to achieve decision dominance faster than ever before.[23]

The demands of future conflict will affect the four components of military expert knowledge in the following ways.

Military-Technical

In the targeting process, AI facilitates sorting through data from sensors to identify HVTs and to assign the best assets to engage. As understood here, targeting is an integrated process that requires the prioritization of targets, resources, and effects.[24] One can measure the performance of the targeting process in terms of instances it identified an HVT correctly, failed to identify an HVT correctly, and identified something as an HVT incorrectly.

Human Development

The technical component shows soldiers will have to develop varying degrees of AI and data literacy. Although identifying these varying degrees falls under the technical component, determining how to recruit, educate and train, and manage knowledgeable personnel will become a critical professional task. One can measure success in terms of how well the system proliferates and sustains a capability once developed as well as how the system expands on expert knowledge to ensure trust is maintained.

Ethical

Trust also requires preventing—or at least mitigating—potential harm to noncombatants and friendly forces.[25] Given the potential for friendly and noncombatant casualties, especially in LSCO, professionals will have to ensure that application of the technology represents acceptable risk to protected persons, infrastructure, and other material assets. Artificial intelligence also raises questions of accountability. Relinquishing more of the decision-making process to machines may result in bad outcomes, even if both humans and machines perform their duties correctly. Understanding how to respond to such outcomes will be critical to applying AI. Here, we might measure success

23. Sydney J. Freedberg Jr., "Army's New Aim Is 'Decision Dominance,'" Breaking Defense (website), March 17, 2021, https://breakingdefense.com/2021/03/armys-new-aim-is-decision-dominance/.

24. HQDA, *Fires*, Army Doctrinal Publication 3-19 (Washington, DC: HQDA, July 2019).

25. C. Anthony Pfaff, "The Ethics of Acquiring Disruptive Technologies," *PRISM* 8, no. 3 (2019): 128–45; C. Anthony Pfaff, "Five Myths about Military Ethics," *Parameters* 46, no. 3 (Autumn 2016): 8; and Christian Brose, *The Kill Chain: Defending America in the Future of High-tech Warfare* (New York: Hachette Books, 2020).

in terms of whether AI-enabled outcomes result in less harm than human-only processes do.

Political

Using AI has implications for civil-military relations and various government stakeholders that can affect AI's development and application. Understanding these impacts and mitigating negative effects are critical tasks for professionals.

To understand how to address AI's impacts on military expert knowledge, one must better understand how the functioning of AI-enabled systems can impede the process of building trust. One must also understand how to work through each barrier to gain confidence in the system and those who operate it. Artificial intelligence's barriers to trust stem from data curation, algorithm optimization, and outside manipulation.

— 2 —

Technical

Introduction

As Darrell West and John Allen point out, intentionality, intelligence, and adaptability differentiate systems that rely on AI from other kinds of systems. Systems enabled by AI are intentional because they can analyze input quickly using algorithms and large amounts of data and then act on the analysis. In addition, AI-enabled systems are intelligent because they can use machine learning (ML) to find statistical associations and patterns to support particular decisions. Furthermore, AI-enabled systems are adaptable because they can adjust to changing circumstances.[1]

Machine learning (ML) is a subfield of AI that gives computers the ability to learn without being explicitly programmed. Having evolved from the study of pattern recognition, ML explores the notion that algorithms can learn from and make predictions about data. Predictive analytics and ML go hand in hand, as predictive models typically leverage ML algorithms. Predictive modeling largely overlaps with the field of ML.[2]

The algorithms used in ML produce two types of predictive models: classification models that predict class membership and regression models that predict a number. The system designer decides which type of ML algorithm, classification or regression, suits the given application. For example, a regression model would be appropriate in forecasting a given tactical operation's chance of success, whereas a classification model may be used

1. West and Allen, *Turning Point*, 3–5.

2. Andrew W. Trask, *Grokking Deep Learning* (Shelter Island, NY: Manning Publications, 2019), 11.

to categorize targets within an image into particular classes.[3] In this paper, we will discuss the former model type.

For predictive models to be formed by ML algorithms, the models must have data from which to learn. Commonly referred to as *training data*, these data consist of historical examples collected from an application over time and include predictor variables along with the target variable one wishes to predict. Given the correlation between the predictor and target variables, an ML algorithm can be used to search for and find an optimized prediction model. But for future predictions to be accurate, the new data presented to the model must have similar characteristics to that of the training data.

The outcome of this learning process, which is referred to as *training*, is the predictive model or *trained model*. Collecting training data for ML is often a tedious curation process in which the predictor variables are conditioned and coupled with the correct output variable one wishes to predict. Predictive models can be trained over time or updated to respond to new data or values, which keeps the models accurate as conditions potentially change over time.[4]

To clarify the training process, consider the military application of using ML to classify whether objects in an image are targets. The first step is to collect enough images of the targets one wishes AI to detect automatically. The training data must sufficiently represent the targets operating under anticipated conditions—such as weather, lighting, and atmospheric effects—using the sensor of choice. The regions of pixels containing the target must be annotated with the name of the target so the ML algorithm can learn to associate the patterns of pixels with the target type. The pool of annotated or "labeled" images is then provided to the ML algorithm, which automatically generates its classification model. This model can then be used to process future images and determine whether targets exist within any region of an image. The classifier's performance depends on how closely the training data represent or are similar to the new data in which the classifier is presented.

3. "Regression vs. Classification in Machine Learning," Javatpoint (website), n.d., accessed on October 31, 2022, https://www.javatpoint.com/regression-vs-classification-in-machine-learning.

4. Michael I. Jordan and Tom M. Mitchell, "Machine Learning: Trends, Perspectives, and Prospects," *Science* 349, no. 6245 (2015): 255–60.

Barriers to Trusting AI and Data Technologies

Systems enabled by AI pose several challenges to building appropriate trust: they are literal, rigid, and function as a black box. Moreover, they think differently from humans because they reason via mathematical models that rely on and manipulate the data to which they have access. Although both humans and algorithms recognize and store information about visual patterns, models cannot move beyond their training data sets and logically abstract patterns as humans can because they are too literal.[5] Humans' and algorithms' disparate methods of reasoning mean AI-enabled machines can produce radically different decisions based on the same input humans receive.[6] The inability to conceptualize also makes models perform poorly in situations that deviate from the models' training; they are highly susceptible to mistakes when facing an unpredictable battlefield.

Data Challenge

Optimizing either a classification or a regression model requires training and retraining the model on input data sets. Algorithms are only as good as the data on which they are trained. Through training, the machine learns to differentiate items of interest from everything else. Relative to data, collecting accurate, complete, consistent, and timely data sets for the system to train on is extremely difficult and sensitive to the environment in which the targeting will take place. Keeping data sets updated is critical work that must be ongoing. Even with a focused collection effort, no data set is ever a complete representation of the world, and, as a result, algorithms will make mistakes when their operational inputs do not closely resemble the data on which the model was trained.

Performance Issues

Performance issues usually come in the form of misclassifications, false positives, and false negatives. For instance, when the inputs do not resemble the training data, prediction mistakes are more likely. Prediction mistakes happen when a classifier is trained using only images of openly exposed targets taken during the summer months and then presented images of partially concealed targets taken during the winter. If a system is trained using only images of tanks operating in the desert

5. Marcus Comiter, *Attacking Artificial Intelligence* (Cambridge, MA: Belfer Center for Science and International Affairs, 2019), 12–13.

6. Defense Science Board, *Summer Study on Autonomy* (Washington, DC: Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics, 2016), 14.

and is then asked to classify an image of a tank partially covered in snow, then the classifier will likely make a mistake. To counter such mistakes, continuously searching for and collecting new, informative data examples as they become available and then using them to retrain and update the classifier as needed—especially relative to the environment one is operating in—is important. Often, retraining and updating the classifier means collecting new data while the system is in operation and then identifying the samples that can help to improve the AI model's performance.

In short, classifiers can produce mistakes given the state of the art and the difficulty of collecting comprehensive data sets. Artificial intelligence can be described as a black box because how AI arrives at output is not always discernible to humans, either due to the complexity of the algorithm or because the output depends on the strength of the connections in the network. Commanders and operators should still understand the limitations of AI and familiarize themselves with AI-enabled systems' performance in similar conditions, thereby enabling the commanders and operators to decide how much control to provide to AI in targeting operations based on their risk calculations.

The potentially black-box nature of AI-enabled systems also impacts trust: humans have difficulty understanding the systems, and this difficulty undermines the sense of predictability humans use to build confidence in one another.[7] The prevalence of neural networks within AI-enabled object classification tasks contributes to the black-box metaphor. Neural networks are a complex form of modeling, and even the data scientists who develop the networks often cannot discern which input or series of inputs produced a certain output. Indeed, one reason adversarial attacks can be so effective against AI-enabled systems is the attacks take advantage of the black-box nature of the systems to evade detection.

The ability to audit AI inferences, from sourcing the data for acquisition to normalizing the data before applying the algorithm, is critical to overcoming some of the concerns about AI. Thus, a key aspect of trust is understanding the data pipeline from sensor to inference.[8]

Vulnerabilities

Even if one can overcome the concerns described above, the enemy will actively attempt to disrupt one's ability to employ AI by poisoning data

7. Defense Science Board, *Summer Study*, 15.

8. Colonel Joseph M. O'Callaghan, e-mail message to author, August 6, 2022.

sets or changing the signatures of the enemy's assets. A poisoning attack undermines an ML model during the training phase by altering its training data.[9] Adversarial poisoning attacks could train a target identification model to ignore one class of object entirely, enabling an HVT to hide in plain sight.[10] In a second type of adversarial manipulation called an *input attack*, an adversary injects noise into the operational data feeding into a deployed model to produce an incorrect output. In one example, a small piece of tape placed on a stop sign caused self-driving cars to misidentify the sign as a 60-mile-per-hour speed marker.[11] Similarly, an adversary can make minute, carefully crafted modifications to an object to cause neural networks to misidentify the object.[12] Input attacks can also occur in the digital space. Small pixel changes invisible to the human eye have caused classification algorithms to misidentify images of pandas as monkeys.[13] Both types of attacks, input and poisoning, can undermine the perceived effectiveness of fielded models and degrade trust.

Taken together, the sensitivity of the data sets, the pipeline the data takes, any transformations the data may undergo to meet input requirements, the complexity of the algorithms, and the potential for undetected sabotage give rise to an accountability gap. Accountability depends on intent and action, but harm, including violations of the law of armed conflict, may occur despite the commanders, staffs, and operators who are interacting with an AI-driven system acting with good intentions and the system, with the exception of spoofing, working according to specification. Commanders and staffs may understand the system well but still suffer from automation bias, especially with systems that are normally reliable, thus increasing the probability of unaccountable harm.

Importantly, AI performance is not all about speed. Indeed, the machine provides better output when humans interact with it, even during operations. For instance, much like an arms race, fighting financial fraud requires constant development of better detection methods because criminals learn to defeat the ones currently in use. The result is a cycle of better detection continuously creating better criminals. To stay ahead of this cycle, algorithms and scoring models must be updated frequently, which requires many human

9. Jared Dunnmon et al., *Responsible AI Guidelines in Practice* (Washington, DC: Defense Innovation Unit, 2021), 26.

10. Comiter, *Attacking*, 28.

11. Dunnmon et al., *Responsible AI Guidelines*, 26.

12. Anish Athalye et al., "Synthesizing Robust Adversarial Examples," *Proceedings of Machine Learning Research* 80 (2018).

13. Comiter, *Attacking*, 22.

data analysts, information technology professionals, and financial fraud experts interacting with algorithms and data to keep them ahead of criminals.[14] Therefore, the idea that developing and employing AI involves a trade-off between speed and meaningful human control is a false dilemma. The question, then, is how do humans know when and where to interact with a system to provide control while optimizing the system's performance?

Taken together, AI's vulnerabilities challenge a commander's ability to assume accountability for the system's performance, which includes both the machine components and the human components. The first challenge is that responsibility is diffused because AI's performance depends on decisions about curating, classifying, and labeling data made by developers, acquisition officials, commanders, staffs, and operators in the field. Second, the system's complexity increases the chance of automation bias, wherein operators are disposed to accept a machine's output because of their limited ability to validate it. Together, these challenges set conditions for violations or other harms, even though the system is otherwise functioning properly, thus introducing an accountability gap that must be overcome for commanders to be held accountable.

To overcome the accountability gap, one must first overcome the barriers to trust identified above. Overcoming these barriers requires a twofold process. First, commanders, staffs, and operators must understand how to ensure the system is operating properly, and, second, they must understand when the system may provide erroneous output.

Overcoming Barriers to Trusting AI

Although trusting AI-enabled systems will be imperative on the future battlefield, this trust must be calibrated. If commanders and staffs show too much trust, then the military risks succumbing to automation bias, wherein operators, staffs, and commanders blindly accept machine output.[15] If commanders and staffs show too little trust, then the military could lose the ability to exploit vast amounts of data at machine speed, resulting in the erosion of decision dominance. To build appropriate trust in emerging warfighting systems, commanders must understand the components of technical trustworthiness for ML-enabled systems—the digital infrastructure

14. H. James Wilson and Paul R. Daugherty, "Collaborative Intelligence: Humans and AI Are Joining Forces," *Harvard Business Review* (July–August 2018).

15. Brian Stanton and Theodore Jensen, *Trust and Artificial Intelligence*, National Institute of Standards and Technology Interagency Internal Report 8330 (Gaithersburg, MD: National Institute of Standards and Technology, December 2020).

that provides the how and why of ML-driven decision making.[16] To trust ML-enabled warfighting systems, organizations must familiarize themselves with a model's performance and applicability, including its performance in operational environments, as well as the quality of the operational data feeding the model.

Assessing Model Competency

Evaluating a model's competency requires evaluating whether the model can complete its designated task effectively and whether it can do so at an acceptable level of risk. Despite wide study of the desirable values for ML-enabled systems, no universal list of the factors defines a trustworthy algorithm.[17] In the United States, the National Institute of Standards and Technology has established nine characteristics for AI system trustworthiness. They are accuracy, safety, objectivity, reliability, resiliency, explainability, security, privacy, and accountability.[18]

Commanders, staffs, and operators can assess the trustworthiness of a particular model by employing a model card, which describes key model attributes related to the National Institute of Standards and Technology's trustworthiness characteristics. In this case, the card does not address privacy because this characteristic is not relevant to the targeting being considered here. Model cards narrow the gap between the data scientists who develop the algorithms and the operators who use the system by communicating information in a standardized way.[19] Understanding these metrics is akin to reading a soldier or officer's record brief and evaluation. The metrics delineate the basic qualifications and goals of a model. Model cards should start with basic administration information, including the type of model, date of development and current version, and the contact information of the development team. The model cards should have three major sections—performance, safety, and security—that cover the nine trustworthiness characteristics.[20] See figure 1

16. John Basl, Ronald Sandler, and Steven Tiell, *Getting from Commitment to Content in AI and Data Ethics: Justice and Explainability* (Washington, DC: Atlantic Council, 2021), 14.

17. Andrew Ilachinski, *AI, Robots, and Swarms* (Arlington, VA: CNA, 2017), 185; Independent High-level Expert Group on Artificial Intelligence, *The Assessment List for Trustworthy Artificial Intelligence* (Brussels: European Commission, 2020); and Stanton and Jensen, *Trust*.

18. Chad Boutin, "NIST Proposes Method for Evaluating User Trust in Artificial Intelligence," National Institute of Standards and Technology (website), May 19, 2021, https://www.nist.gov/news-events/news/2021/05/nist-proposes-method-evaluating-user-trust-artificial-intelligence-systems.

19. Matthew Arnold et al., "Fact Sheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity," *IBM Journal of Research and Development* 63, no. 4/5 (2019): 6:1–6:2.

20. Margaret Mitchell et al., "Model Cards for Model Reporting," *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta: Association for Computing Machinery, 2019), 222.

for a sample model card for an object classification model that could be used to support target identification.

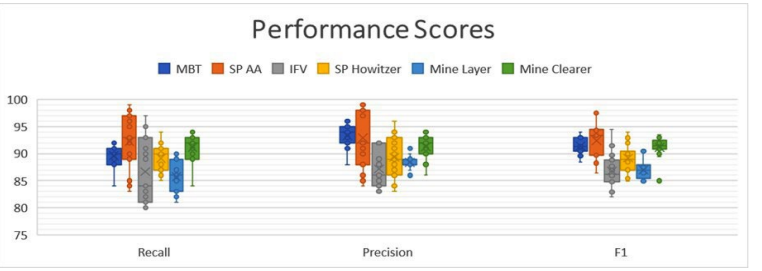
Model Card: Tracked Vehicle Identification

Administrative Information

- Developed by DoD Project X, 2020. POC susan.a.smith.mil@mail.mil. Current version 2.1 (DDMMMYY)
- Convolutional Neural Net

Performance

- Classifies tracked vehicles into 6 classes: Main battle tank (MBT), Self-propelled Anti-Air (SPAA), Infantry Fighting Vehicle (IFV), Self-propelled (SP) Howitzer, Mine Laying Vehicle, and Mine Clearing Vehicle
- Output: object bounding box, object ID, confidence score



Safety

- Intended use: ID of tracked vehicles within satellite-based electro-optical imagery across all biomes.
- Data dependencies: input chip size, preprocessing requirements, specific sensor bands
- Out of scope use: not for use with imagery below NIIRS level 6 (ground resolved distance (GRD) greater than 0.75m) and off-nadir angles above 25%; performance metrics drop-off rapidly in those conditions.
- Control features: requires in-the-loop control, with human operator confirming or denying each identification.
- Training Data:
  - AdversarialGroundVehiclesA, a set of 15,000 labeled images of commercial origin collected between 2000 and 2019 covering Adversary1, Adversary2, and Adversary3 tracked military vehicles. Dataset did not include new class X-XX tanks produced by Adversary3.
  - Test Data: test data split of AdversarialGroundVehiclesA; test data not used in training.
- **Reliability**
  - Confidence level calibrated to performance; produces confidence levels of possibly (45–54.9%) likely (55–79.9%), very likely (80–94.9%), and almost certainly (95%+).
  - Below 45% confidence, model will not classify object
  - Performance metrics evaluated by object class solely, not separated by GRD or lighting metrics within scope of use; not trained to separate friendly tracked vehicles from enemy tracked vehicles.
- Resilience: Edge operation requires local servers with XX GPU power, cloud operation requires XX bandwidth
- Explainability: Model will highlight portions of image that it considered most important for classification and will highlight portions of image that lowered confidence

Security

- Model classification and network authorities
- Training data classification
- No known capture of devices using model (CAO DDMMMYY)

Accountability

- Detected object assigned unique ID that carries locational and timestamp data, model classification and confidence, and analyst approval or disapproval logged against ID. If sent to targeting system, unique ID carried through targeting solution.

Figure 1. Model card: Sample model card for object classification model support target identification

Accuracy

Mathematical accuracy captures how often an algorithm produces the correct answer. A targeting classification algorithm that correctly identifies

9 out of 10 enemy tanks across a set of photographs is 90 percent accurate. Yet, accuracy can be a misleading performance metric, particularly when the classes of objects a model is trying to identify have unequal representation. For example, when using a data set that consists of 90 tanks and 10 self-propelled artillery pieces, a simple model that identifies every object it sees as a tank has the same overall accuracy—but different object class accuracy—as a model that correctly identifies 80 tanks and all 10 artillery pieces. Emphasizing accuracy over performance values like precision, recall, and F1 scores can result in a model that performs very poorly on underrepresented groups, even when the given object class is important—for instance, higher on a commander's HVT list.[21]

Safety

Performance metrics alone are insufficient to evaluate the ability of a model to operate competently. High-stakes applications—those impacting life or death—also require a consideration for safety, which is the next section of a model card. In this section, operators can look to understand whether the model is safe to use in a given operational context.[22] A design focus on safety inevitably leads to trade-offs between system performance in some categories to mitigate harmful outcomes in others. Indeed, improving a classification system's response to outlying data often decreases the overall accuracy levels.[23] Yet, despite a small drop in accuracy, overall performance is improved when a model can perform against unexpected data.

Risk management is the inherent driver behind AI safety, and a safe AI application operates within an acceptable level of risk when evaluated in its operational context.[24] Key safety information includes a detailed description of intended use, data dependencies, and out-of-scope usage—for example, if a model has been trained on only one biome and should not be used in others.[25] Out-of-scope usage is particularly important because it informs operators of known high-risk applications, such as if using imagery or video below a given quality level could significantly impact performance metrics. Another hallmark of safe AI system design is human oversight

21. Dunnmon et al., *Responsible AI Guidelines*, 16.

22. Arnold et al., "Fact Sheets," 6; and John D. Lee and Katrina A. See, "Trust in Automation: Designing for Appropriate Reliance," *Human Factors* 46, no. 1 (2004): 75.

23. Independent High-level Expert Group on Artificial Intelligence, *Assessment List*, 21; and Dunnmon et al., *Responsible AI Guidelines*, 16.

24. Independent High-level Expert Group on Artificial Intelligence, *Assessment List*, 10.

25. D. Sculley et al., "Hidden Technical Debt in Machine Learning Systems," in *NIPS 15: Proceedings of the 28th International Conference on Neural Information Processing Systems* (Cambridge, MA: MIT Press, 2015).

of decision making, a safeguard that is also a requirement of DoD AI-enabled weapon systems.[26]

Objectivity

Artificially intelligent systems make repeatable, systemic assessments based on patterns within their training data. Narrow training sets introduce bias into the model and result in repeatable, systemic errors. To manage the risk of such errors, model cards should describe the training data set (including size, source, variability, and collection time frame) and explicitly highlight data-set imbalances, enabling a commander to assess the likelihood of predictive bias in an operational environment that was not represented in the training.[27]

A common example of a biased classification algorithm comes from a 2018 paper by researchers from the Massachusetts Institute of Technology and Microsoft that reported the accuracy of three commercially available gender classification algorithms dropped over 34 percent for nonwhite male categories—a bias derived from training data that overrepresented white males.[28] A similarly biased facial recognition system could wreak havoc if deployed to identify military-age males in a foreign theater. Data-set size, variance, and similarity to the intended deployed environment are important considerations when evaluating bias. Labeling large data sets is a significant effort, and the requirements for security may make military databases particularly prone to size and variance limitations.[29]

Reliability

Reliability in ML systems is a function of consistent performance values and failure rates. Dependable systems have confidence measures that are calibrated to performance values; for instance, an 80 percent confidence measure should represent an 80 percent likelihood of system accuracy

26. Valerie Insinna and Aaron Mehta, "Updated Autonomous Weapons Rules Coming for the Pentagon," Breaking Defense (website), May 26, 2022, https://breakingdefense.com/2022/05/updated-autonomous-weapons-rules-coming-for-the-pentagon-exclusive-details/.

27. Timnit Gebru et al., "Datasheets for Datasets," arXiv (website), n.d., updated December 1, 2021, https://arxiv.org/abs/1803.09010.

28. Joy Buolamwini and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *Proceedings of Machine Learning Research* 81 (2018): 1–15.

29. Osonde Osoba and William Welser IV, *An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence* (Santa Monica, CA: RAND Corporation, 2017), 19; and Ilachinski, *AI, Robots, and Swarms*, 63–64.

for the detection of a given object.[30] Dependable systems also know what they do not know. For example, a dependable system would not label an object the system had a low measure of confidence in identifying. Dependable systems also know when operational data diverges from their training examples and lessen confidence levels to match their inexperience.[31] Model cards can aid a reliability assessment by clearly describing how the model derives high, moderate, and low confidence assessments.

Resiliency

The model card should also cover how it treats disruptive events. Technologically resilient systems indicate steadiness in the face of changing conditions by limiting the impact of unanticipated events.[32] A resilient system can handle variations in input data, such as stickers of pixel-adjusted pandas on stop signs. Architectural requirements are also a resiliency consideration. Some models require cloud access and have bandwidth considerations, yet others may be able to operate at the edge. Commanders must know the degree of trustworthiness of models functioning in a battlefield environment.

Explainability

The importance of the final feature of the safety section, explainability, increases with the impact level of decision making.[33] Explainable AI lessens the black-box nature of algorithms and enables humans to judge accuracy, risk, and fairness, recognizing even highly accurate models will get some decisions wrong.[34] In the life-or-death decisions that accompany AI-enabled targeting, explainability is likely a more important characteristic of trustworthiness than accuracy because explainability enables commanders to assess risk and implement control measures according to changing conditions.[35] The model card should explain how the model provides explicable results. One method of target identification algorithms is to highlight the object areas the model considered most important in assigning the classification or to highlight

30. Chuan Guo et al., "On Calibration of Modern Neural Networks," *Proceedings of Machine Learning Research* 70 (2017).

31. Massachusetts Institute of Technology, "Robust AI" (working paper, Recent Advances in Artificial Intelligence for National Security Conference, Cambridge, MA, November 15, 2021).

32. Oliver Eigner et al., "Towards Resilient Artificial Intelligence: Survey and Research Issues," in *2021 IEEE International Conference on Cyber Security and Resilience* (New York: Institute of Electrical and Electronics Engineers [IEEE], 2021), 536.

33. Independent High-level Expert Group on Artificial Intelligence, *Assessment List*, 15.

34. Osoba and Welser, *Intelligence in Our Image*, 3.

35. *Advancing the Science and Acceptance of Autonomy for Future Defense Systems, Before the House Emerging Threats and Capabilities Subcommittee*, 114th Cong. (2015).

object areas that are causing the model to lower confidence assessments and to present this information to an operator for validation.[36]

Security

The next major section of the model card is security, a vital characteristic of algorithm trustworthiness given the likelihood of adversarial attacks and the high-stakes nature of AI warfighting applications. Secure systems are resistant to adversarial tampering. Recent studies show a design for interpretability can defend against adversarial attacks that attempt to induce misclassification.[37] Similarly, ensemble models that combine several types of models into a single interface are difficult for adversaries to fool and, therefore, are more secure.[38] The classification of the developed model, its training data, and a timeline for how long the model is authorized to operate on a given network should also be included in the security section, as should an explanation of whether any devices operating the model have been captured or are known to have been hacked.[39]

Accountability

The final section of a model card should detail accountability features. Accountable systems enable auditing, which helps public and private allies and partners to gain confidence in the United States' use of AI-enabled warfighting systems. The model card should delineate who is responsible for monitoring the deployed model's performance, how identified errors are debriefed, and who will fix performance deviations.[40] Before fielding an AI-enabled warfighting system, commanders should identify rollback options—including returning to a previous version of the model or removing the AI entirely and reverting to a human-centric system—and develop decision points and procedures for implementing these options.[41]

Model cards that capture key details related to each of the National Institute of Standards and Technology's nine characteristics of trustworthiness will enable commanders and staffs to evaluate the model's competency in the context of a specific operational environment. To ensure a shared

36. Akhilan Boopathy et al., "Proper Network Interpretability Helps Adversarial Robustness in Classification," *Proceedings of Machine Learning Research* 119 (2020).

37. Boopathy et al., "Network Interpretability."

38. Boopathy et al., "Network Interpretability."

39. Comiter, *Attacking*, 37.

40. Larry Lewis, *Insights for the Third Offset: Addressing Challenges of Autonomy and Artificial Intelligence in Military Operations* (Arlington, VA: CNA, 2017), 45–46.

41. Dunnmon et al., *Responsible AI Guidelines*, 24.

trustworthiness assessment across an operational staff, the model card should be briefed to all stakeholders before model deployment. This forum will enable the commander to communicate the designated baseline sufficiency and relative importance of each characteristic, evaluate the risks in deploying the model, and determine effective control measures.[42]

Data Quality

Although understanding the competency of a model is the first step in assessing ML-enabled systems, it does not provide the full picture of trustworthiness in a given operational setting. The quality of operational data a deployed model is processing is the second evaluation metric. This metric answers the fundamental question of whether operational data inputs are fit for use in a given task. Data is the foundational ingredient through which the algorithm operates, and the quality of data is imperative to a model's deployed performance.[43] Operational commanders should monitor the inputs feeding into their deployed models with a data quality dashboard. These dashboards are similar to the dashboards of cars: the information shown on a car's dashboard is not highly detailed and does not cover all of the car's functions, but the dashboard depicts whether critical indicators are running within acceptable norms. Studies have commonly coalesced around four basic data quality considerations: accuracy, completeness, consistency, and timeliness.[44] These four markers should form the core of a data quality dashboard.

Data Accuracy

Accuracy, or ensuring data correctly represents its source and has been verified, is tightly linked to data provenance and protection.[45] The DoD Data Strategy includes the requirement for DoD data to include "protection, lineage, and pedigree metadata."[46] The security of collection, transport, and storage methods impacts the reliability of the data, and assessments of these factors depend on sourcing. Although commanders have commonly made targeting decisions based on classified information that is often presumed to be accurate

42. Dunnmon et al., *Responsible AI Guidelines*, 27.

43. Comiter, *Attacking*, 16.

44. Li Cai and Yangyong Zhu, "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era," *Data Science Journal* 14, no. 2 (2015): 2.

45. Li and Zhu, "Challenges of Data Quality," 5; and Carlo Batini et al., "Methodologies for Data Quality Assessment and Improvement," *AMC Computing Surveys* 41, no. 3 (2009): 6.

46. Department of Defense (DoD), *DoD Data Strategy* (Washington, DC: DoD, 2020), 8.

based on secure collection methodologies, cyberattacks against central data repositories could corrupt military data sets.[47]

Leveraging commercial data sets also poses accuracy risks because the military may acquire data sets but have no means to verify their accuracy. Commercial data may also be more susceptible to cyberattack or other forms of digital corruption because private-sector entities may not have the incentive or means to protect the data or rapidly share information about breaches to all of their data consumers, leaving commanders vulnerable to corrupted data. The same may also be true when leveraging information collected by allied and partner nations.[48]

Fortunately, demand is growing in the commercial industry for data owners to certify the trustworthiness of data stores while taking data quality and access into consideration.[49] Until trustworthiness certifications become the norm for military and commercial data sets, data quality dashboards should include accuracy metrics that depict data source, data security, and any known security breaches. Staffs should also ensure they have data health points of contact for each data set the staffs leverage.

Completeness

Another component of data trustworthiness is completeness, which refers to the concept of a data set having values for all required attributes and a collection of data sets includes all relevant data.[50] The completeness of a data set should be tied to the data dependencies listed on a model card: all required data fields must be present, with no blank cells. Confirming all applicable data sets are available is an additional measure. In the context of targeting identification algorithms, complete data includes all available imagery for a given area, regardless of collection source, or full-motion video from all drones with access to the target area. Although a model can only operate on one piece of imagery or video feed at a time, the same model may produce varied assessments from different inputs, leading to a more holistic assessment. Complete data sets also include all fields required for a given algorithm. If a targeting identification algorithm required a value capturing a given image's data

47. Comiter, *Attacking*, 68.

48. Erik Lin-Greenberg, "Allies and Artificial Intelligence: Obstacles to Operations and Decision-making," *Texas National Security Review* 3, no. 2 (Spring 2020): 62.

49. Arnold et al., "Fact Sheets," 4.

50. Laura Sebastian-Coleman, *Meeting the Challenges of Data Quality Management* (Oxford, UK: Elsevier Press, 2022), 232.

quality (its National Image Interpretability Rating Scale value) to assess confidence values, complete data sets would consistently include the National Image Interpretability Rating Scale value as one of their features.

Consistency

Consistency, the third component of data quality, refers to data being captured systematically, thus enabling a model to ingest and process the data.[51] For targeting identification algorithms, consistent data would present all values in a standard format that can be ingested by the model, such as locational data that is consistently presented in a latitude/longitude format. Consistent data also represent the expected range, such as National Image Interpretability Rating Scale values, which fall between zero and nine. For ML models, consistent data should also be statistically similar to training data to prevent performance degradation due to data drift.[52]

Timeliness

Data are a representation of the world—which is subject to constant change—at a specific point in time.[53] Accordingly, the timeliness of data, including the concepts of regular updates and freshness, is a critical consideration for data trustworthiness.[54] Information freshness—the idea that the time from data collection to data availability is acceptable—is also a critical component of any commander's target selection standards. The data trustworthiness dashboard should feature the most recent date of update and the expected next update for all data sources as well as a depiction of whether the data meets the freshness requirement set forth in the commander's target selection standards.

Due to the immense amount of data flowing into operational headquarters, separate data quality dashboards should be developed for critical functions, such as targeting. Monitoring the status of the dashboard provides important insight into the trustworthiness of an algorithm's decision making for a particular set of input data. If they are going to rely on algorithms to support decision making, even if it will be supervised by a human, commanders should identify the measures of accuracy, completeness, consistency, and timeliness they consider to constitute

51. Batini et al., "Methodologies," 7.

52. Andrew Burt et al., *Beyond Explainability: A Practical Guide to Managing Risk in Machine Learning Models* (Washington, DC: Future of Privacy Forum, 2018), 5.

53. Christopher Fox, Anany Levitin, and Thomas Redman, "The Notion of Data and Its Quality Dimensions," *Information Processing & Management* 30, no. 1 (1994): 15.

54. Batini et al., "Methodologies," 8.

an unacceptable risk based on their understanding of the model card. For instance, if a commander knows from the model card that a model's performance drops significantly when imagery is more than 25 percent off nadir, the commander can determine whether or not he or she is unwilling to accept trust model identifications under these conditions. If an upcoming satellite pass is more than 25 percent off nadir, the data quality standards will be unmet, and the staff will know to leverage additional collection resources to support target identification.

Assessing Trust over Time

Although model cards provide insight into the trustworthiness of a model at deployment, and data dashboards provide insight into the health of individual inputs, commanders must also understand how ML-enabled systems perform over time. Unlike weapon systems that operate the same years after having been fielded, learning systems develop similarly to humans, with operational learning and feedback loops that modify system performance over time. Any trustworthiness assessment is at best a snapshot in time that must be continually reevaluated. The military uses routine education and evaluation systems to ensure its members meet organizational requirements over time. Systems enabled by ML should be no different: continuous evaluations can serve to monitor and manage system performance while reinforcing the partnership between acquisitions and operational leaders that must be maintained across the full life cycle of a system.[55] To best enable this feedback loop, commanders should establish data red teams that capture key statistics and review them with stakeholders at regular intervals.

Technical mechanisms built into AI-enabled systems as part of accountability measures play a key role in verifying the ongoing effectiveness of deployed ML-enabled systems. A common method of evaluating these systems is to compare a metric of performance against its training baseline: An algorithm should maintain a similar performance score under deployed operations as it did during test and evaluation. But because deployed target identification models will not be used against prelabeled data sets, an approximation for accuracy must be developed. One method for doing so is to validate the machine's uncertainty characterization. For example, if analysts are not overturning roughly 20 percent of machine-generated calls that have a calibrated 80 percent confidence level, something may be amiss, and further investigation is merited.

55. Defense Science Board, *Summer Study*, 22.

Data drift—when operational data differs from training data—is a common reason for compromised performance. For instance, a model that comprises an algorithm that only searches for tanks and a data set that contains 90 tanks and 10 mobile artillery pieces invites data drift because the model is likely to identify artillery as tanks. In addition to monitoring uncertainty characterization, staffs should monitor the overall data set and highlight any significant differences in environment, object classification, or imagery features between the training data set and the operational data. In addition to reviewing statistics of the overall data set, operational teams should sample and trace decisions to search for potential problems in the model engineering teams need to fix.[56] Root cause analysis of known mishaps is particularly important for identifying any potential adversarial attacks.[57] The speed and scale advantages of AI risk compounding unintended outcomes programmed within the algorithm, underscoring the need to identify problems as early as possible and mitigate systemic risks.

Studies indicate many of the healthiest organizational relationships exhibit high levels of both trust and distrust—as underscored by the adage "Trust but verify."[58] When AI-enabled systems are performing reliably, regularly reviewing data of the acceptable results helps to build trust.[59] When AI-enabled systems are performing poorly, reviews enable early identification and correction. Operational reviews enable a commander to ensure AI-enabled machines continue to function according to his or her intent and form a key link from the operational force back to the development team responsible for rebaselining the systems.

The model and data characteristics described here give commanders a way of identifying the skills staff and operators must have and a framework for ensuring they have performed the necessary actions to ensure the system functions optimally. A feature of AI-driven systems is that everything can work as designed, and everyone can do what they are supposed to do, but errors may still occur. In this sense, working with AI-enabled systems can be like working with humans: no matter how well they have been trained and how good their intentions, mistakes can happen with no obvious explanation. The difference is human-only interactions are informed by a fairly

56. Dunnmon et al., *Responsible AI Guidelines*, 12.

57. Joseph Convery, *Intelligence after Next: Ensuring Decision Advantage on the Future Battlefield—Intelligence at the Speed of Hypersonic Warfare* (Bedford, MA: Mitre Corporation, 2021), 5.

58. Stanton and Jensen, *Trust*, 3.

59. Convery, *Intelligence After Next*, 5.

well-developed sense of command and individual responsibility. Systems driven by AI do not share this trait.

Teaming with AI

Of course, trust and risk are the central concerns here. Commanders need a reliable way to know when they can trust AI and when they can allow it to execute some stages of the targeting process with less supervision for the benefit of speed (but at the cost of a higher level of risk). Fortunately, precedents can help to increase the transparency of AI-enabled system performance.[60] Precedents also provide clarity in the results of human-centric measures that focus on improved interfaces and teaming aspects.[61] These precedents demonstrate that adaptive human-machine teaming that adjusts the level of human intervention as conditions change can make the targeting process more effective and safer.

To manage the appropriate level of human-machine interaction, a fuzzy-logic controller can be used to automate decisions about the number of human interventions required for a given target based on a series of conditions. Fuzzy logic provides the ability to impart human reasoning in an automated decision-making process through a series of rules that are easily understood and explained. To control the amount of interaction between humans and AI in the targeting process, the fuzzy-logic controller could consider multiple criteria before making each teaming decision, such as: (1) the commander's risk tolerance for the given mission; (2) the AI algorithm's confidence it has correctly identified a target; (3) the choice of best effect (or shooter), as selected by an AI optimization algorithm; and (4) the likelihood of collateral damage and fratricide, as assessed by an air and ground deconfliction algorithm, like the one being developed by the Defense Advanced Research Projects Agency.[62]

Such an adaptive teaming model is much like other safety-critical systems that are starting to integrate AI. For instance, the automotive industry is using AI in its advanced driver-assistance systems to automate much of the driving;

60. Laura Freeman, "Test and Evaluation for Artificial Intelligence," *INSIGHT* 23, no. 1 (2020): 27–30; and Arnold et al., "Fact Sheets," 13.

61. Patricia McDermott et al., *Human-Machine Teaming Systems Engineering Guide* (Bedford, MA: Mitre Corporation, 2018); and Maria Jesus Saenz, Elena Revilla, and Cristina Simón, "Designing AI Systems with Human-Machine Teams," *MIT Sloan Management Review* 61, no. 3 (Spring 2020): 1–5.

62. Harry Lye, "DARPA Looks to AI, Algorithms to De-conflict Airspace," Airforce Technology (website), April 9, 2020, https://www.airforce-technology.com/features/darpa-looks-to-ai-algorithms-to-de-conflict-airspace/.

the human supervises and takes control whenever the system loses confidence or when he or she observes errant behavior.[63] Similarly, military applications of AI could execute parts of the targeting process with only human supervision, but whenever the AI is not highly confident about its assessment, the process flow can be blocked until a human explicitly verifies the output of the AI.

The key is getting the human-machine pairing right because failure to do so results in slowing down the process and undermining the advantages AI is supposed to provide. Thus, integrating AI into the targeting process presents an optimization problem in which the goal is to find the right mix of human-machine teaming that maximizes the speed of AI without employing overly restrictive control measures and while ensuring the appropriate amount of human oversight for safety and risk reduction purposes. The fuzzy-logic controller proposed in this paper would help to solve this optimization problem.

Developing Reliable and Capable Systems

Artificially intelligent targeting systems rely on deep learning, which is an approach to AI that uses numerous layers of artificial neural networks to process information in a way that loosely mimics the biological brain.[64] Deep neural networks (DNNs) have proven to be effective at detecting patterns in data, including imagery.[65] In a military context, sensors that generate images, such as cameras and radars on satellites, can be processed by DNNs to detect targets. This concept is depicted in figure 2. The figure shows pixels of an image being processed through the interconnected layers of a DNN with multiple outputs. Each of the final neurons in the output layer shown in figure 2 pertain to a target class and is coupled with a confidence score that reflects the probability of the target type being present within the image.

63. Mansur Arief, Peter Glynn, and Ding Zhao, "An Accelerated Approach to Safely and Efficiently Test Pre-production Autonomous Vehicles on Public Streets," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)* (New York: IEEE, 2018), 2006–11; Xingyu Zhao et al., "Assessing Safety-critical Systems from Operational Testing: A Study on Autonomous Vehicles," *Information and Software Technology* 128 (December 2020); and Benjamin Bauchwitz and M. L. Cummings, *Evaluating the Reliability of Tesla Model 3 Driver Assist Functions* (Chapel Hill, NC: Collaborative Sciences Center for Road Safety, October 2020).

64. Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning* (Cambridge, MA: MIT Press, 2016), 78.

65. Iqbal H. Sarker, "Machine Learning: Algorithms, Real-world Applications and Research Directions," *SN Computer Science* 2, no. 3 (March 2021): 160.

These confidence scores are usually assigned by a softmax function added at the end of the neural network.[66] A softmax function normalizes the set of real numbers from the last layer of neurons in the DNN into a meaningful probability distribution that ranges from zero to one. Each output from the softmax can be interpreted as the probability its corresponding target class exists in the image. But before a DNN can make such predictions with confidence scores, it must be trained to recognize specific targets. This process involves feeding annotated examples (that is, a training set) through the DNN iteratively many times. During this training cycle, the ML algorithm uses the training data to search for the optimal set of tunable parameters to turn the DNN into a generalized prediction model that can be used against new inputs. The predictive performance of the model depends on the similarity of the new inputs to the data used during the training.

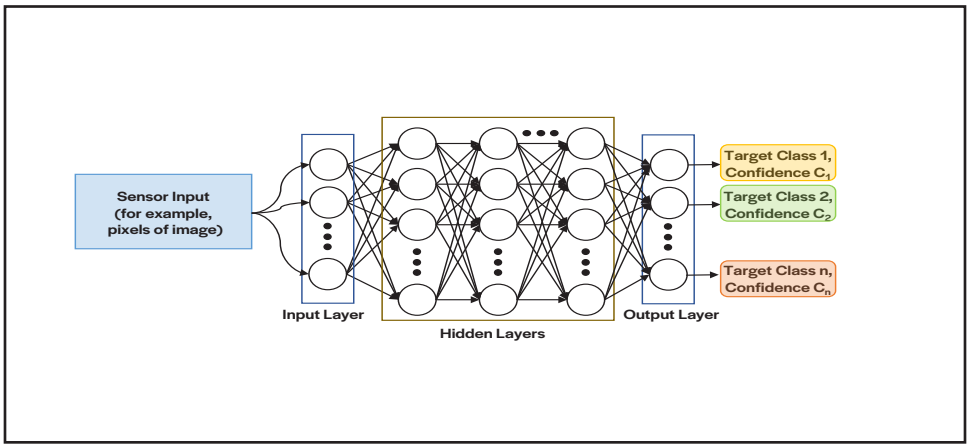


Figure 2. Deep neural networks (DNNs) and targeting: Concept diagram of DNN used in target detection. In this case, as new images are captured from a sensor, they are processed through the DNN, which predicts, with a degree of confidence, whether a particular target exists in the image.

When the inputs to AI classifiers do not resemble the training data, prediction mistakes are more likely. For instance, prediction mistakes occur when a classifier has been trained using only images of openly exposed targets taken during the summer and is then presented images of partially concealed targets taken during the winter. Figure 3 illustrates the challenge of collecting sufficient training data for AI classifiers to operate under different conditions. If a classifier has been trained using only images of tanks operating in the

66. Thomas Wood, "Softmax Function," DeepAI (website), n.d., accessed on June 16, 2022, https://deepai.org/machine-learning-glossary-and-terms/softmax-layer.

desert and is then asked to classify the image of the tank partially covered in snow, then the classifier will likely make a mistake. To try to counter such mistakes, continuously searching for and collecting new, informative data examples as they become available and using them to retrain and update the classifier as needed is important.

Often, retraining and updating the classifier means collecting new data while the system is in operation and identifying the samples that can help to improve the AI model's performance. Regardless, the main point is classifiers can make mistakes given the state of the art and the difficulty of collecting comprehensive data sets. Ultimately, commanders must understand AI's limitations and familiarize themselves with a system's performance in similar conditions, thereby enabling the commanders to make wiser decisions about how much control they are willing to provide to AI in targeting operations based on risk assessments.

Figure 3. Environmental effects: Images of tanks shown in different environments to illustrate the challenge of collecting sufficient training data for AI classifiers to operate under different conditions

Artificial intelligence filters through data rapidly and recognizes relevant patterns quickly.[67] Hence, in the case of targeting, AI can be trained to detect certain targets and then used to process the data streams of sensors in near real time. Figure 4 depicts one of the ways AI can be used in the dynamic targeting cycle of "find, fix, track, target, engage, assess" (F2T2EA), wherein AI can have a significant impact given the inherent time constraints associated with current operations and dynamic targets. As the figure conveys, AI can be used to scan sensor feeds continuously and to aid targeting analysts in the steps of "find, fix, and track" by alerting them whenever a new target is detected. Artificial intelligence can accomplish these steps more quickly than humanly possible and do so continuously, without getting fatigued or skewing its judgment under the stresses of combat.

67. Sarker, "Machine Learning."

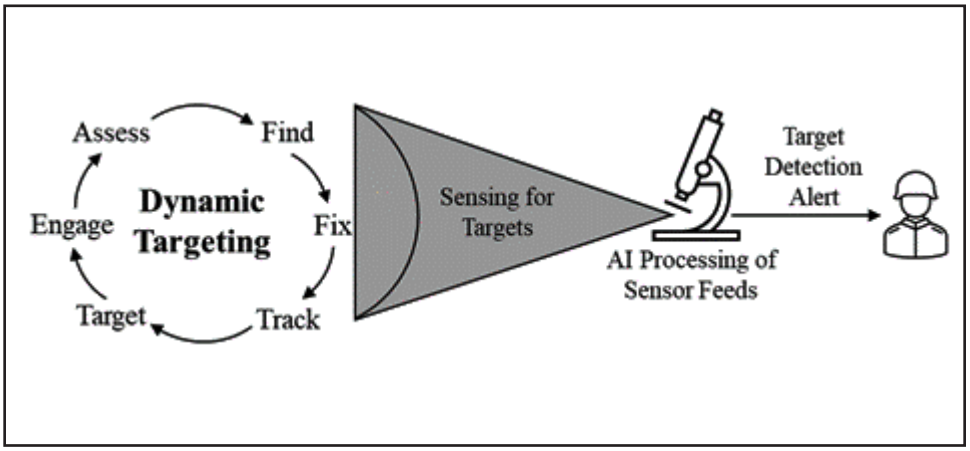


Figure 4. AI augmentation of targeting: AI can augment the dynamic targeting cycle by searching for targets and then sending alerts upon detection

One possible drawback to dynamic targeting, in which every target detection must be reinspected by an analyst, is the human bottleneck. The targets detected by AI in a short period of time could greatly outnumber the analysts available to confirm them, especially in large-scale ground combat operations. For instance, space-based sensors could quickly scan a wide area near an adversary's base of operations, and AI could process the data in seconds and produce thousands of potential targets for image analysts to verify. In this case, a target queue must be formed, and analysts must review them one by one. The longer a potential target spends in the queue, the longer it will take before effects on the target are realized.

To mitigate such bottlenecks, AI can be paired with humans in a more adaptable way that does not always require the same level of inspection in every scenario. Implementing a simple and static teaming scheme whereby all targets detected by AI must be individually reinspected by a human may be overly cautious and slow. But mediating the need for human involvement can be achieved by creating a decision matrix that accounts for machine certainty and risk tolerance, which will be discussed later.

Sometimes, a situation may demand that a commander accept greater risk for the sake of acting faster to protect the force. In such a scenario, a more agile and adaptive teaming scheme may be necessary wherein the commander's risk acceptance and the AI's confidence in its detections can be used to determine which targets can more safely pass the reinspection queue and move to the next targeting stage—especially when a final human verification stage is in place.

In addition to target acquisition, AI can also play an important role in more quickly and effectively designating assets to engage the target as well as assessing whether the selected effects may cause unintended consequences. In figure 5, the dynamic targeting steps of F2T2EA have expanded to include three more subprocesses within the target step. In addition, the steps have been color coded to indicate whether AI can augment the particular subprocess.

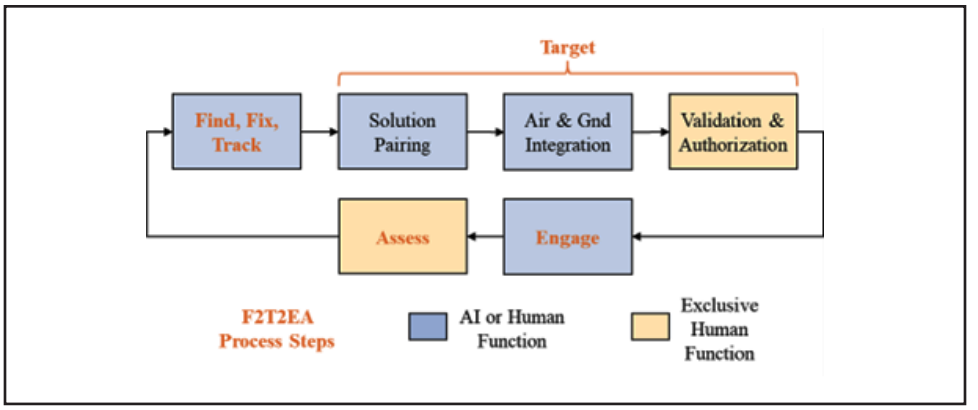


Figure 5. Find, fix, track, target, engage, assess (F2T2EA): The dynamic targeting steps of F2T2EA have expanded to include three more subprocesses within the target step. In addition, the steps have been color coded to indicate whether AI can augment the particular subprocess.

One factor of asset selection is determining the effect one wants to achieve, such as destroy, disrupt, or neutralize.[68] *Solution pairing* refers to matching a target with the appropriate effect. The process is straightforward in the case of a preplanned, deliberate target already listed in the attack guidance matrix. But for some dynamic targets, such plans may not exist. Therefore, when considering the numerous targets that will be encountered during LSCO and the limited resources available for effects, the selection of the best shooter is effectively an optimization problem for which AI is well suited.

A second factor is the risk any asset represents for collateral damage or fratricide. Today, this process is aided by prior planning, the building of fire support coordination measures and airspace coordinating measures, and the use of digital systems, including the Joint Automated Deep Operations Coordination System and the Advanced Field Artillery Tactical Data System. Fire support coordination measures and airspace coordinating measures are

68. HQDA, *Targeting*, D-5.

designed to expedite attacks on targets while protecting forces, populations, critical infrastructure, and religious sites from errant fire missions.[69]

The Advanced Field Artillery Tactical Data System checks whether a fire mission violates the fire support coordination measures or airspace coordinating measures. Nevertheless, leaders typically verify the deconfliction of fires with airspace managers and other team members within the fires cell before authorizing a mission, likely due to a lack of confidence these systems have been updated with the latest operating picture. Under a data-centric approach, systems could be designed to share data easily, allowing the various systems that manage the integration of air and ground fires to use the data.

Existing algorithms, including AI, can potentially accelerate the clearance of Joint fires and make it more reliable, without the heavy reliance on preplanned measures that account for the maneuver of friendly forces, assuming these algorithms are given real-time air and ground common operating pictures and access to other useful data, such as that contained in the Army's One World Terrain database.[70] With accurate blue-force tracking, three-dimensional digital map data from One World Terrain, and ballistics data, an intelligent algorithm can quickly estimate whether an effect is likely to have unintended consequences.

The third subprocess in the target step of figure 5 is labeled, "Validation & Authorization." During this step, leaders within the fires cell at any echelon (such as deputy fire support coordinators or targeting officers) use their best judgment to make a final decision whether to engage based on the commander's guidance, the law of war, rules of engagement, and other considerations outlined in doctrine. This step guarantees every targeting mission has at least one human who is in the loop and responsible for assessing and deciding whether to proceed to engagement using all of the available information. To avoid this stage being a potential bottleneck, the Army should continue to exercise mission command, whereby multiple targeting officers have the authority to make such engagement decisions based on staying aligned with the commander's intent. Additionally, engagement decisions should be permitted at any echelon that shares an area of responsibility, allowing busier fires cells to divert some targets for faster validation and authorization.

69. JCS, *Joint Targeting*, Joint Publication 3-60 (Washington, DC: JCS, 2013), II-31–II-32, III-7.

70. "One World Terrain (OWT)," Institute for Creative Technologies (website), n.d., accessed on January 26, 2022, https://ict.usc.edu/research/projects/one-world-terrain-owt/.

Inputs to the Control System

Recall from figure 2 that DNNs indicate how confident they are with each target classification. This information can be exploited during targeting to make an informed decision about the level of human supervision required, especially when combined with other information, such as the commander's tolerance for risk based on the mission's context. A visual representation of this confidence score—which, like any probability, ranges from zero to one—is depicted in figure 6. The figure shows decision thresholds, as indicated by the horizontal lines separating low, moderate, and high, can be set to aid in decision making. For instance, whenever the classifier is not highly confident about its classification, then the data samples in question are added to the targeting analyst queue for human reinspection.

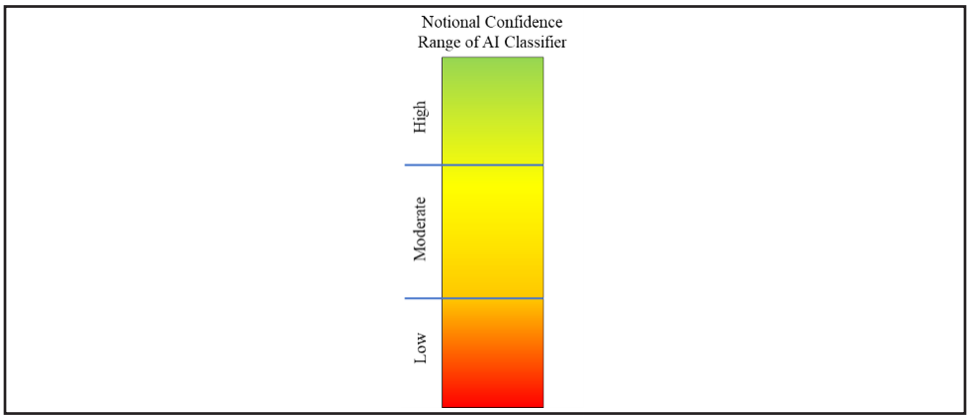


Figure 6. Classifier confidence: A visual representation of an AI classifier's confidence score, which ranges from zero to one and can be divided into linguistic sets like "low," "moderate," and "high" to aid in decision making

The concept in fuzzy logic is not to hard code single-value thresholds, as indicated by the blue horizontal lines in figure 6. Rather, the idea is to program smooth transitions between the input classes of low, moderate, and high, as indicated by the gradual change in color in the figure. This differentiation makes fuzzy logic more tolerant of uncertainty when measuring and quantifying the inputs into sets described by words or "linguistic sets." Moreover, these quantifiers allow humans to interpret the measured inputs more easily and then make subsequent decisions based on intuitive classification. Fuzzy logic facilitates this process by allowing inputs—in this case, the AI's confidence for a target type—to belong to multiple sets potentially, but with varying degrees of membership. Figure 7 provides an example of possible "fuzzy sets" constructed for the

input "AI Confidence." The shapes of these fuzzy sets are designed by a human using expert knowledge. The regions where the moderate set overlaps with either low or high are the ranges where the input would be classified as belonging to multiple sets, with partial membership in each—such as 80 percent high and 20 percent moderate.

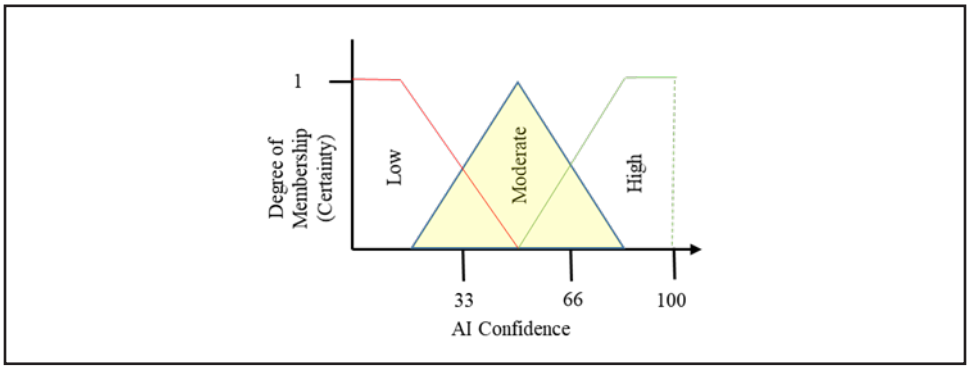


Figure 7. Fuzzy logic: An illustration of possible fuzzy sets for the input AI Confidence. The shapes of these fuzzy sets are designed by a human using expert knowledge.

Another input that can be used to aid in making decisions about how to handle targets that have been classified by AI is the commander's risk tolerance. The commander determines the level of acceptable risk at which the AI can operate. Therefore, commanders should be given the flexibility to assume more risk at times, if warranted, based on their best judgment. For example, a commander may be risk averse when providing fire support in a counterinsurgency mission or a dense urban environment with many civilians nearby, but more risk tolerant when facing a high-intensity battle in mostly open terrain or performing final protective fires when friendly forces could be overrun by the enemy. To capture risk tolerance, commanders could be given a rheostat-like interface they could tune to convey their risk tolerance directly to the system. Figure 8 illustrates the concept and shows a continuous and variable range from low to high risk tolerance.

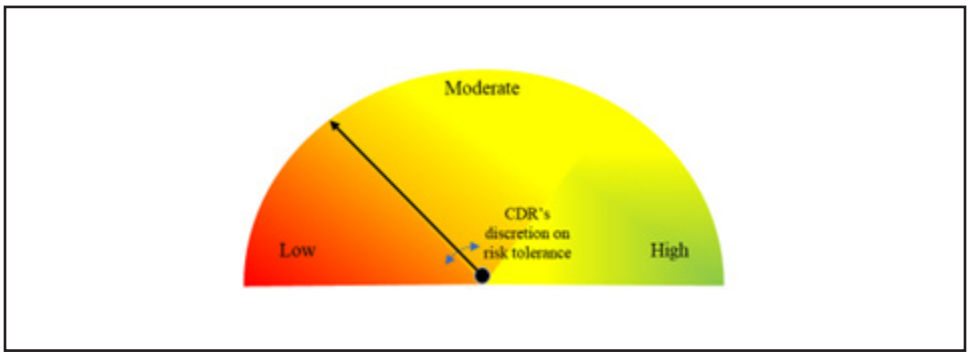


Figure 8. Commander's risk tolerance: A rheostat-like knob that commanders could use to relay their risk tolerance to the AI-enabled targeting system

Decision-making Logic within the Control System

Using the inputs above (classifier confidence and commander's risk tolerance), the goal is to produce a fuzzy-logic controller that determines, in a rapid and automated fashion, the level of oversight and direct human involvement in the targeting steps shown in figure 5. The controller's purpose is to decide the optimal mix of human-machine interaction that should take place in the F2T2EA process based on the conditions of each target. Although a commander's risk tolerance will likely remain constant for some time, the AI's confidence will probably vary for each target it detects. Hence, the controller's recommendation on the level of oversight in the F2T2EA process could change for each target.

A fuzzy-logic controller gives humans the ability to impart their decision-making logic into an automated control system via a set of preprogrammed rules.[71] Next, these human-derived rules drive the action (or output) of the system based on the conditions of the inputs and the rule base logic. Figure 9 depicts a block diagram for a programmable controller designed to gauge the confidence of AI and the risk tolerance of the commander for every newly detected target and then choose the level of oversight under which the targeting process should occur. As indicated in the figure, adding more than two criteria into the decision-making process of the controller is possible, but illustrating the concept is easier if the rule base is kept to only two dimensions. The rule base would be programmed into the controller's memory using a series of if/then statements and would

71. C. J. Lowrance, A. P. Lauf, and M. Kantardzic, "A Fuzzy-based Machine Learning Model for Robot Prediction of Link Quality," in *2016 IEEE Symposium Series on Computational Intelligence* (*SSCI*) (New York: IEEE, 2016), 1–8; and C. J. Lowrance and A. P. Lauf, "An Efficient Fuzzy-based Power Control Scheme for Ad Hoc Networks," in *2015 Wireless Telecommunications Symposium (WTS)* (New York: IEEE, 2015), 1–8.

obey the following logic: "If AI's Classification Confidence is low and CDR's Risk Tolerance is low, then human involvement is maximum. If AI's Classification Confidence is high and CDR's Risk Tolerance is high, then human involvement is minimum." Assuming two inputs with three categories each (low, moderate, and high), the complete set of nine rules can be derived by the two-dimensional rule base shown in figure 9.

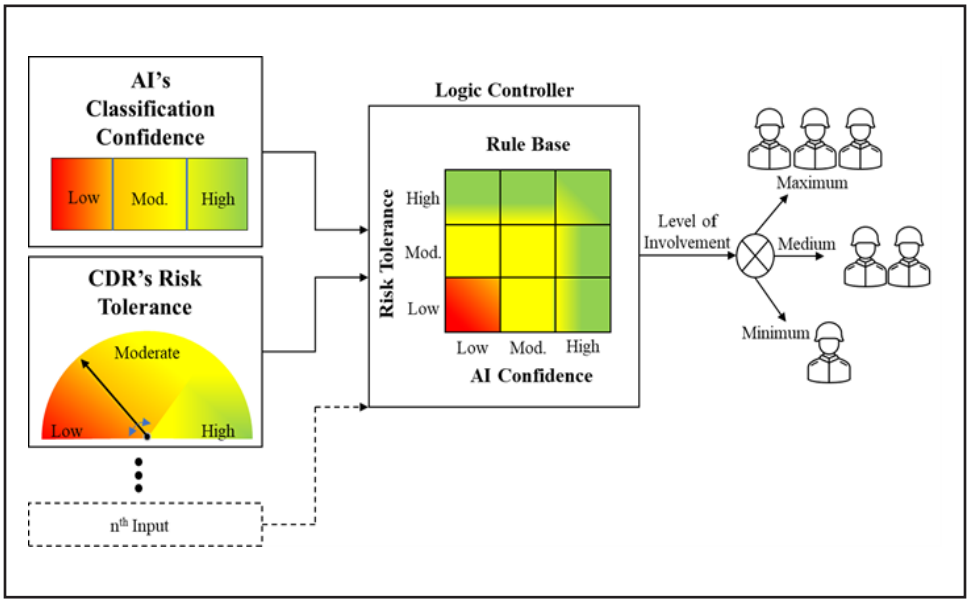


Figure 9. Combining AI's classification confidence and CDR's risk tolerance: A block diagram for a programmable controller designed to gauge the confidence of AI and the risk tolerance of the commander for every newly detected target and then choose the level of oversight under which the targeting process should occur

Risk Profiles and Adaptive Teaming Based on the Fuzzy-logic Controller's Recommendation

The following example provides a qualitative description of how fuzzy logic works, but a more in-depth and quantitative explanation of fuzzy logic and control is available from the cited references by John H. Lilly and C. J. Lowrance.[72] Assume the input "AI's Classification Confidence" was measured and found to belong partially to both high and moderate, with some arbitrary percentages that sum to 100 percent. Similarly, the CDR's Risk Tolerance was measured but found to partially belong to moderate and low. In this case, the two inputs can make a total of four combinations.

72. John H. Lilly, *Fuzzy Control and Identification* (New York: John Wiley and Sons, 2011), 1–42; and Lowrance and Lauf, "Fuzzy-based Power Control."

These combinations mean four rules would be activated, but they would only be activated to a partial degree, based on percentages the inputs were found to belong to each fuzzy set. Each rule has its own consequent output value, but if a rule is only partially activated, then the full weight of the rule or consequent is not completely applied. The final stage of fuzzy logic averages the partially activated rules together to calculate a single output value; in this case, the value would determine the level of oversight (maximum, moderate, or minimum) for a given target.

As indicated in figure 9, a target may be recommended for a maximum, moderate, or minimum level of human involvement in the targeting process. The corresponding three ways of potentially modifying the targeting process based on these output options are shown in figure 10. Each version of the targeting steps is shown to be aided by AI to some degree, but the main difference is the number of steps in which humans must verify the output from an AI-driven stage before proceeding to the next stage. Regardless, the final verification and authorization step would be reserved for humans.

The controller's decision for maximum involvement implies a human-driven targeting process, wherein humans lead each stage. But as figure 10 suggests, a decision for maximum involvement does not preclude AI from assisting in the steps for the sake of speed. In other words, AI can augment any step, but a human must explicitly verify the output before the target proceeds. On the opposite extreme, minimum involvement translates into AI automating all steps, except for the final validation and authorization process, in which a leader in the fires cell would review the targeting information and recommendations before giving the order to proceed with a fire mission. The moderate oversight process flow is more nuanced and is similar to the process flow for minimum oversight, except the classification confidence of the AI algorithm and the risk assessment from integration stage must meet stringent thresholds set by the commander. For example, the AI confidence must exceed 95 percent, and the likelihood of unintended consequences must be less than 5 percent; if a threshold is not met in either case, then a human must inspect the output generated by the AI algorithm.

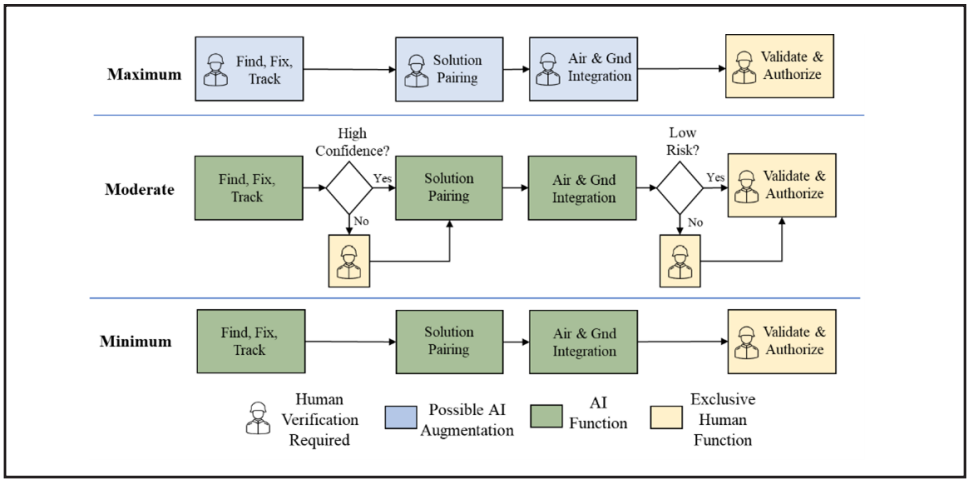


Figure 10. Risk profiles: Three risk profiles for varying the level of human oversight in an AI-enabled targeting process based on conditions assessed by a fuzzy-logic controller

Conclusion

This section has described one way to think about human-machine teaming that allows commanders to adjust the system—including both humans and machines—to optimize performance and minimize the chance of unaccountable error. The introduction of the fuzzy-logic controller addresses accountability by making commanders responsible for the accuracy of their risk assessments and ensuring data is properly curated for the context in which the commanders employ the algorithm. This way of thinking about human-machine teaming also addresses automation bias because it gives humans a way to know when the machine itself is, in a sense, uncertain about its output. Whether these measures are good enough depends on how well the system balances the imperatives discussed earlier compared to a human's balancing of the imperatives. The next challenge for the Army is developing the kinds of commanders, staffs, and operators who can ensure the system runs as designed and is tasked appropriately and who can reasonably take responsibility if the system fails to produce satisfactory results.

— 3 —

Human Development

Developing and Managing Talent

Effectively implementing data-based technologies like AI into staff processes requires skill sets ranging from basic data literacy to data science and engineering. As stated earlier, integrating AI and data technologies is difficult because they can simultaneously be highly specialized and require widespread application. But currently, few personnel at the corps level have the highly specialized skills to meet this demand. Employing these technologies requires integrating them with space-based and other sensors as well as Joint command-and-control systems, and this integration requires personnel who have knowledge of both the systems and the processes. In addition, outside vendors possess much of the AI and data expertise, increasing the difficulty of finding personnel with the proper expertise and requiring additional expertise in contracting AI capabilities.

To remedy the lack of personnel with AI and data science education and skills, the Army has implemented plans to educate selected personnel at three levels. The first level is a short course on data-driven leadership that would educate senior leaders on data management and its applications to AI. The second level consists of AI professionals who would function as data analysts, data engineers, and autonomous system engineers. Data engineers would be responsible for the more complex aspects of data management, including setting up and managing data lakes and structured databases and connecting data streams to populate these data stores. Data analysts would be military professionals who would monitor the deployed model's performance, train new AI models using new data examples, and evaluate updates before deploying them onto the AI-enabled systems. These skills would require a masters-level education. Finally, AI technicians would be trained

via certificate-level education programs for enlisted personnel, noncommissioned officers, warrant officers, and junior officers. The curriculum of these courses, the length of which would be less than a year, would include the setup and management of cloud resources, programming and scripting language basics, and AI fundamentals.[1]

Although these programs are certainly necessary, they may not be adequate to provide the range of skilled personnel to proliferate capabilities across the Army at the corps level, especially in the short term. For instance, the Army has graduated approximately 40 officers from participating universities in its first two cohorts of the AI Scholars Program, most of whom will take roles at the Army Artificial Intelligence Integration Center or other institutional positions.[2] But little talent will be left for bottom-up efforts at the corps level or lower, possibly for several more years.

Further complicating matters, the Army's ability to manage personnel skilled in science, technology, engineering, and mathematics (STEM) is limited, and even more so for personnel who have AI and data expertise. Indeed, without a more efficient management system, optimizing the assignment of personnel trained by these new Army programs, especially at the operational level, to take advantage of new, often commercially available technologies so the Army remains agile relative to its adversaries may not be possible. Optimizing Army talent management will require revising how the service identifies educational requirements and aligns talent with operational needs. Optimizing the talent management will also require tracking the talent more effectively so it is available where it is most needed. The rest of this section discusses the limits of the current Army personnel system and suggests recommendations to improve it.

Army Educational Requirement System

The Army identifies Army Educational Requirement System positions on either a table of organization and equipment (TOE) or a table of distribution and allowances (TDA). Force managers build these documents through the five-phase force development process that involves developing capabilities, designing organizations, developing organizational

1. Scott Maucione, "Army Futures Command Preparing an AI-ready Workforce," Federal News Network (website), October 27, 2020, https://federalnewsnetwork.com/army/2020/10/army-futures-command-preparing-an-ai-ready-workforce/.

2. Courtney Chapman, "AFC Recognizes Army Scholar and Technician Program Graduates," US Army (website), September 20, 2022, https://www.army.mil/article/260372/afc_recognizes_army_ai_scholar_and_technician_program_graduates; and Mark Phillips, e-mail message to author, February 22, 2022.

models, determining organizational authorizations, and documenting organizational authorizations.[3] Force managers generate effective dates for establishment (EDATEs) for these documents annually. As part of the review process, unit-level force managers can submit equipment and personnel changes before the TDA and TOE are published. When a unit-level force manager submits a change, documentation requirements for coding and grading positions are generated.[4]

Changing personnel force structure can be a complicated and time-consuming process because it requires coordination with the Headquarters, Department of the Army, deputy chief of staff, G-1, which conducts an annual review of the database of the Army Educational Requirement System within the Army Authorization Document System.[5] This process requires a long lead time. Once a unit decides it needs to change its personnel structure, the process can take a minimum of one to two years. If changes need to be implemented, the Headquarters, Department of the Army, deputy chief of staff, G-3/5/7 publishes modification TOEs and TDAs 12 to 18 months before the EDATE, which is when personnel managers can begin to fill these positions, to provide time to plan and synchronize resources to fill authorizations properly.[6] Thus, the time between making a skill-related change to the Army force structure and filling the new or redesignated positions can take as much as three years or more.

Army Talent Alignment Process

To improve the speed and agility of assigning officers with technical education and skills, the Army developed the Army Talent Alignment Process, which uses an online program called the Assignment Interactive Module (AIM) to match positions that require technical skills with officers who have them more effectively.[7] (The most recent version, AIM 2, is the second version of the program.)[8] Assignments under this system happen in two cycles. The

3. HQDA, *Force Development and Documentation Consolidated Policies*, Army Regulation (AR) 71-32 (Washington, DC: HQDA, March 21, 2019), 3.

4. HQDA, *Military Occupational Classification Structure Development and Implementation*, AR 611-1 (Washington, DC: HQDA, July 15, 2019), 8.

5. HQDA, *Advanced Education Programs and Requirements for Military Personnel*, AR 621-1 (Washington, DC: HQDA, December 11, 2019), 1.

6. HQDA, *Force Development*, 8, 31.

7. US Army Talent Management, *Officer's Guide to ATAP* (Washington, DC: Headquarters, Deputy Chief of Staff, Army G-1, 2019).

8. US Army Human Resources Command, "Assignment Interactive Module 2.0 (AIM 2)," US Army (website), April 4, 2017, https://www.hrc.army.mil/site/assets/directorate/OPMD/What%20is%20AIM%202.pdf.

first is a "-01" cycle that fills assignments opening up in the October through March time frame, and the second is a "-02" cycle that fills assignments opening in the April through September time frame.

Every officer, unit, and job is contained within the Army Talent Alignment Process. The officers and units engage throughout the AIM cycle, which lasts around six weeks. Officers can see what units provide under the unit information section, and the units can view each officer's resume, which is an expanded officer record brief that removes some data but adds a second page on which officers can address their knowledge, skills, behaviors, and preferences. Under this system, knowledge reflects the officer's education and training; skills are how the officer puts his or her education and training into action; behavior is the reflection of an officer's values, attitude, and temperament; and preferences indicate the path the officer wishes to pursue in the Army.[9] A unit can use AIM to conduct workplace planning and gauge the steps it must take to address future staffing needs.[10] Once these steps have been determined, units can advertise the job on AIM as well.[11]

The difficulty with this system is advertised positions are assigned against specific officer areas of concentration and functional areas (FAs). As an exception, a unit may have some officer immaterial positions, coded 01A (officer immaterial) or 02A (combat arms immaterial), that the unit could use to augment the unit with additional talent. These positions do not require an officer from a specific field, but they may be performed by any officer with certain experience, manners of performance, and demonstrated potential. But officers can only be assigned to an immaterial generalist position if the branch has filled all slots. Thus, branches such as infantry, armor, and artillery, which typically have excess officers, are traditionally able to fill these positions. Given the Army rarely has a surplus of STEM officers—especially officers with AI and data skills—these officers are generally not available for assignment this way.

9. Greg Lockhart, "The 'So What' behind KSBs," US Army (website), December 8, 2020, https://www.army.mil/article/241467/the_so_what_behind_ksbs.

10. Greg Lockhart, "The 'So What' behind KSBs – Part 2," US Army (website), March 10, 2021, https://www.army.mil/article/243797/the_so_what_behind_ksbs_part_2.

11. US Army Talent Management, *Commander's Guide to ATAP* (Washington, DC: Headquarters, Deputy Chief of Staff, Army G-1, 2020).

Recommendations

Optimizing the Army's talent management system to improve its ability to assign personnel with relevant technical expertise will likely require making changes to how personnel with technical skills are managed as well as how positions requiring technical skills are identified. Moreover, any new system will have to be more flexible and adaptive than in the past to keep pace with changes in technology.[12]

Create New Skill Identifiers

One of the most important lessons from Project Ridgway is personnel in every career field require some technical expertise, much like how field- and flag-grade officers require expertise in Joint operations. Thus, just as personnel with relevant education and experience in Joint operations receive a skill identifier (SI), assigning an SI to those who acquire technical skills may also make sense. Skill identifiers specify occupational areas that are not normally associated with a particular career field but are nonetheless required for certain positions. These skills may be related to one branch, FA, or area of concentration, but they are required to perform the duties of a special position. As with the Joint SI, personnel may require significant education, training, or experience, but possessing an SI would not require repetitive tours or provide progressive career development assignments.[13]

The military must also be prepared to create new SIs as organizations learn more about AI system operations and find new applications. For example, the Scarlet Dragon exercises and operations in Europe have identified the need for someone who understands how the process operates from the data harness to the output pipeline to keep it calibrated.[14]

Of course, simply creating an SI is insufficient for meeting Army needs. For instance, the Army recently created AI-specific SIs, such as "2U" for AI cloud technician-user, "2V" for AI cloud technician, and "2W" for autonomous systems engineer.[15] But after searching for this talent within the officer corps using the AIM Commander Dashboard, the authors

12. HQDA, *The Army Force Modernization Proponent System*, AR 5-22 (Washington, DC: HQDA, 2015), 4.

13. HQDA, *Military Occupational Classification*, 11.

14. O'Callaghan, e-mail message to author, August 6, 2022.

15. "ASI," US Army Credentialing Opportunities On-line (website), April 1, 2022, https://www.cool.osd.mil/army/index.htm; and "Chapter 4: Skill Identifiers (SI)," US Army (website), June 9, 2022, https://api.army.mil/e2/c/downloads/2022/06/09/7ff75183/chapter-4.pdf.

only identified three officers with a 2V SI and none with the other two SIs.[16] Given the number of personnel who have been sent to school for AI or related education to support educational institutions like the United States Military Academy at West Point, for example, more must be done to identify these officers for the purpose of assignment.[17]

Establish a Technology Corps

As an alternative to the relatively decentralized SI assignment process, the Army could pursue a more centralized approach by creating a technology corps, as the service did with the logistics corps, which incorporated Ordnance, Quartermaster, and Transportation Corps basic branch officers with the rank of captain or above who had graduated from the Combined Logistics Captains Career Course.[18] The logistics corps crosses multiple branches and allows a Quartermaster, Transportation, or Ordnance Corps basic branch officer to serve in the same type of position, such as the S-4 of a battalion or brigade.

The technology corps could incorporate all STEM-related FAs, including Information Network Engineering (FA 26), Electronic Warfare (FA 29), Information Operations (FA 30), Space Operations (FA 40), Operations Research/Systems Analysis (FA 49), Nuclear and Counter Weapons of Mass Destruction (FA 52), and Simulation Operations (FA 57). The purpose of the corps would be to provide Army units the ability to incorporate new technologies to meet more immediate needs that are not being met through current acquisition and educational systems. Personnel assigned to this corps would also likely need education and training in technology management and integration in addition to knowledge within their respective technical fields.

Create More Flexible TOEs and TDAs

Creating more flexible TOEs and TDAs could be accomplished by assigning multiple branches and FAs to single positions. For instance, certain technical positions could be filled by personnel from a STEM FA or a basic branch, like the Signal Corps, where technical expertise may also reside. This recommendation would allow units access to a greater variety of qualified personnel, without having to create generalist positions and hunt for qualified personnel who could only be assigned if they were excess.

16. "Assignment Interactive Module," US Army (website), n.d., https://aim.hrc.army.mil/portal/index.aspx.

17. "Academic Departments," United States Military Academy at West Point (website), n.d., accessed on November 4, 2021, https://www.westpoint.edu/academics/academic-departments.

18. US Army Combined Arms Support Command, "LG Proponent History," US Army (website), May 4, 2021, https://cascom.army.mil/S_Staff/LB/LG_History.htm.

Similarly, one could fill positions using knowledge, skills, behaviors, and preferences rather than branches or FAs. For example, if a position required AI experience, then units could use the knowledge, skills, behaviors, and preferences pool to search for qualified officers. This process would help to identify and assign personnel who have acquired the relevant technical degree to teach at the United States Military Academy at West Point but who may not receive a relevant SI that makes them easier to track and assign.[19]

Conclusion

As stated previously, the problem with developing corps-level capability with existing technology is that in the officer corps, the relevant skills are rare, and the relevant education is even rarer. For instance, XVIII Airborne Corps currently has three space officers (FA 40) and two simulation/knowledge management officers (FA 57).[20] Because of the officers' positions within the targeting process, their skills may make them good choices to act as AI integration officers, but they still would need additional on-the-job training to be effective. Of course, these officers could take on the provision of relevant AI and data science training, but doing so may make the Army too reliant on the officers' branches for these skills and dilute the skills the officers' primary education is supposed to impart.

Of course, the right people need the right rules. In addition to the technological roles discussed throughout this section, implementation of AI will also create new challenges for operating within long-standing ethical norms. The next section discusses these challenges.

19. "Academic Departments."

20. O'Callaghan, e-mail message to author, November 15, 2022.

— 4 —

Ethics

Ethics and the Professions

Professional practice cannot be divorced from the ethics that govern it. As Samuel Huntington pointed out in *The Soldier and the State*, professionals are "practicing experts" who perform an essential service for the functioning of society. Military professionals provide defense much as other professions provide health, education, or justice.[1] For society to trust the professionals who provide defense, the former must believe both that professionals are certified in the relevant expert knowledge and, when required, the professionals will set aside self-interest to provide defense. Moreover, professional practice is corporate. No individual can provide the relevant service in isolation; thus, military professionals must also trust each other.

These needs express themselves in the kind of moral-ethical expertise that specifies the moral obligations members of the military have to both fellow professionals and the society they serve. The moral-ethical expertise further specifies the norms that should govern the provision of the service. These obligations and norms apply to both the application of force and civil-military relations.[2] Moreover, as Huntington posited, these obligations may consist of "unwritten norms" or be codified into "written canons of professional ethics" that are promulgated through professional education and reinforced in practice by professional institutions tasked with governing the profession.[3]

1. Huntington, *Soldier*, 9.
2. Lacquement, "Army Professional Expertise," 217.
3. Huntington, *Soldier*, 9–10.

Martin Cook observes that professionals do not require a great deal of ethical reflection when they share common assumptions about the nature and purpose of their work and practice under conditions of relative stability. But Cook states that when the "nature, function, and security of the profession" undergo great change, the common view of professional ethical behavior may be inadequate to address the change.[4] As the previous discussion has clearly indicated, the introduction of AI and data technologies is forcing a great deal of change in how the military fights.

To understand how this change impacts the professional ethic, one must understand the ethical concerns lethal targeting raises. In combat, military ethical decision making is often about trading off between military necessity (what it takes to win), force protection, and avoiding collateral harm. Obviously, military operations place soldiers and civilians at risk. Soldiers can reduce the risk they face by limiting their exposure to harm. But if soldiers limiting their exposure makes fires less precise, the risk to noncombatants, the mission, or both may increase. One can prioritize avoiding harm to noncombatants, but doing so means soldiers must increase their exposure to engage more directly with an enemy or place the mission at risk.[5]

Artificial intelligence impacts these trade-offs in multiple ways, the first of which may make warfighting more humane. A well-trained, AI-enabled system can identify both legitimate and illegitimate targets and allow for greater situational awareness. Moreover, AI-enabled systems do not suffer from fatigue or other cognitive impairment as the systems process thousands of targets—sometimes in a single day. Additionally, motivations such as self-preservation, fear, anger, revenge, and misplaced loyalty do not impact the systems' output, thereby giving humans a more reliable and ethical starting point at the least for targeting decisions.[6]

In combination with space-based and other remote sensors, these systems can also allow soldiers to engage the enemy effectively without necessarily increasing the soldiers' exposure. Assuming command-and-control networks can be secured, most soldiers involved in the targeting process do not need

4. Martin Cook, "Army Professionalism: Service to What Ends?," in *The Future of the Army Profession*, ed. Don M. Snider and Gayle L. Watkins, 1st ed. (New York: McGraw Hill, 2002), 337–38.

5. C. Anthony Pfaff, *Resolving Ethical Challenges in an Era of Persistent Conflict* (Carlisle, PA: Strategic Studies Institute, US Army War College Press, 2011), 5–9.

6. Ronald G. Arkin, *Governing Lethal Behavior in Autonomous Robots* (Boca Raton, FL: Chapman and Hall, 2009), 30.

to be in theater to operate these systems. If done properly, the application of these technologies can lower the risk to soldiers without increasing it for noncombatants. To the extent the system provides greater speed, precision, and accuracy, the system can also make winning more likely, which further reduces the risk to soldiers. In addition, to the extent AI is more precise and accurate, it limits overall destruction, reducing the risk to noncombatants.

Nevertheless, these technologies are associated with some ethical risk. The ethical benefits described previously only happen if the data is properly curated, the sensors are adequately sensitive, and the output is reviewed to ensure the machine has not made errors. Fortunately, the US military took up such ethical concerns early in the AI development process. In the 2012 version of DoD Directive 3000.09, Autonomy in Weapon Systems, the Department of Defense states AI-enabled systems must meet standards for reliability, must be subject to human control, and must only be used in ways that conform to the laws of war.[7] Although the Department of Defense intended for this guidance to apply to autonomous and semiautonomous systems, these rules would also apply to decision support systems, like the one employed in Project Ridgway.

Although a good start, simply requiring AI-enabled systems to be reliable and the humans who use them to obey the laws of war does not fully account for the range of ethical challenges the use of these technologies represents. Perhaps the principles the Defense Innovation Board has established are more applicable to a wide range of the military's AI-enabled systems. Affirming the applicability of the laws of war, the board recommended AI systems be designed to be subject to appropriate levels of human oversight, avoid unintended bias that could cause unintentional harm, be transparent and auditable by experts, be tested and assured within a specified domain throughout their life cycle, be able to detect and avoid unintended harm, and have the option of being deactivated should systems demonstrate unintended escalatory or other undesired behavior.[8]

Systems enabled by AI may work as designed, and the humans who design, manufacture, and employ the systems can do so with the best of intentions,

7. Ashton B. Carter, *Autonomy in Weapon Systems*, DoD Directive 3000.09 (Washington, DC: Office of the Secretary of Defense, November 21, 2012).

8. Defense Innovation Board, *AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense* (Washington, DC: DoD, 2020).

but errors can still occur.[9] The possibility of such errors means some ethical concerns remain. Because of the nature and complexity of the systems, humans may not be able to evaluate machine output to avoid collateral or other harms, which gives rise to an accountability gap that undermines the law of war and incentivizes the inappropriate use of the systems.

Moreover, because of machine inscrutability, humans may be prone to automation bias, which can set conditions for unintentional and entirely avoidable harms, even though humans are involved in the decision-making process. This combination of less accountable and more biased human operators threatens to dehumanize warfare in ways that increase the likelihood of harm to both noncombatants and one's own soldiers.

Inscrutability is not the only issue. Even where the model is explainable, operators sometimes focus on AI output as opposed to its role in the process. As a result, the operators may miss the ways in which AI impacts other aspects of the system, leading to error. For example, the process employed by XVIII Airborne Corps is adept at comparing data from sensors with a foundational database to determine whether certain kinds of objects are in the sensor data. But the process is not designed to determine the significance of these objects. Although it may identify a group of tanks, the process is not yet able to gauge factors like adversary intent and type of unit that might be critical to a targeting decision.[10]

Accountability Gap

Understanding that which makes one accountable helps one to understand the importance of accountability. In the context of criminal law, whether civil or military, responsibility rests on an individual's intent to violate a law (mens rea) as well as his or her act of violating that law (actus reus).[11] This standard was employed by the post–World War II Nürnberg trials as well as the Rome Statute that governs the International Criminal Court. As the Nürnberg trial document states, for an accused person to be responsible for a war crime, "there must be a breach of some moral obligation fixed by international law, a personal act voluntarily done with the knowledge

9. Hin-Yan Liu, "Refining Responsibility: Differentiating Two Types of Responsibility Issues Raised by Autonomous Weapons Systems," in *Autonomous Weapons Systems: Law, Ethics, Policy* (Cambridge, UK: Cambridge University Press, 2016), 340.

10. O'Callaghan, e-mail message to author, November 15, 2022.

11. Yoram Dinstein, *War, Aggression, and Self-defence*, 4th ed. (Cambridge, UK: Cambridge University Press, 2005), 136.

of its inherent criminality under international law."[12] This criminal responsibility can also extend to anyone, such as a commander, who orders the commission of any crime.

In the context of military operations, commanders can even be held responsible when they did not have a particular intent or act in a way that resulted in a crime or other kind of violation. Because of their position, commanders are also responsible for crimes they should have prevented, regardless of whether they intended for them to occur. But this accountability is limited. For commanders to be held accountable for a war crime they neither intended nor committed, the commander must both have been in a position to know about the crime and have had responsibility for those who perpetrated the crime.[13]

On the first condition, whether a commander had knowledge of a particular crime does not matter; what matters is whether he or she should have known. As the documents from the Nürnberg trials state, "[A]n army commander will not ordinarily be permitted to deny knowledge of reports received at his headquarters, they being sent there for his special benefit."[14] To the extent a commander has responsibility over an organization, he or she also takes on the affirmative duty of being aware of the actions of his or her subordinates. As the trial documents also state, "If he [a commander] fails to require and obtain complete information, the dereliction of duty rests upon him and he is in no position to plead his own dereliction as defense."[15] These points suggest commanders must avoid ordering illegal or immoral acts, take steps to ensure they are knowledgeable of their subordinate activities, limit unnecessary harm, and, if those preventive measures fail, take steps to hold violators accountable.[16]

Thus, accountability turns on that which one intends and does. For commanders, accountability also depends on whether they ensure adequate awareness of the intentions and actions of one's subordinates as well as whether

12. International Military Tribunal, *Trials of War Criminals before the Nuernberg Military Tribunals under Control Council Law No. 10, Nuremberg, October 1946–April 1949* (Washington, DC: US Government Printing Office, 1950), quoted in Sanford Levinson, "Responsibility for Crimes of War," in *War and Moral Responsibility*, ed. Marshall Cohen, Thomas Nagel, and Thomas Scanlon (Princeton, NJ: Princeton University Press, 1974), 117.

13. Levinson, "Responsibility for Crimes of War," 118.

14. International Military Tribunal, *Trials of War Criminals*, quoted in Levinson, "Responsibility for Crimes of War," 119.

15. International Military Tribunal, *Trials of War Criminals*, quoted in Levinson, "Responsibility for Crimes of War," 119.

16. Michael Walzer, *Just and Unjust Wars*, 5th ed. (New York: Basic Books, 2015), 316–20.

positive measures to ensure meaningful control over the subordinates are present to limit the likelihood of crimes, other violations, or unintended harm. Systems driven by AI turn sensor input into data, which impacts a decision whether to strike a target. Furthermore, the quality of future output depends on retraining the data set by either confirming its accuracy or accounting for changes in the environment resulting from the strike. Therefore, ethical command and control also depends on poststrike feedback that further refines the data set.

Employing AI can pose a challenge for this conception of accountability. Systems enabled by AI can sometimes force a trade-off between taking advantage of the machine's speed and providing meaningful human control. But that which qualifies as meaningful human control is in dispute. Consider the stringent standard proposed by the International Committee for Robot Arms Control that requires human operators to have full contextual and situational awareness of the target area as well as sufficient time for deliberation on the nature of the target, the necessity and appropriateness of attack, and the likely collateral harms and effects. Moreover, the operators must also have the means to abort the attack if necessary to meet the other conditions.[17]

Of course, as described earlier, the difficulty is that human intervention can create time-consuming bottlenecks in an AI-enabled process. Moreover, this standard for meaningful human control holds AI systems to a higher standard than nonautonomous weapon systems already in use. Rarely in war do soldiers and their commanders have "full contextual and situational awareness of a target area," and even when they do, soldiers who fire their rifles at an enemy have no ability to prevent the bullet from striking the point at which they aimed the gun.[18] Thus, banning future weapons based on higher standards than the current systems meet makes less sense when one realizes some of the capabilities that come along with AI-enabled systems can set conditions for better moral decision making and more humane warfare.[19]

Nevertheless, some features of AI-enabled systems make dismissing this concern difficult. As noted earlier, the nature and complexity of these systems can make them a black box. Few commanders, staffs, or operators

17. Noel Sharkey, *Guidelines for the Human Control of Weapons Systems* (Sheffield, UK: International Committee for Robot Arms Control, April 2018).

18. Sharkey, *Guidelines*, 3.

19. Paul Scharre and Michael C. Horowitz, *Artificial Intelligence: What Every Policymaker Needs to Know* (Washington, DC: Center for a New American Security, June 2018), 16.

have the technical capability to understand how sensors, data, and algorithms interact to provide targeting suggestions. In addition, some elements of a system—particularly, the algorithms—often remain the property of their developers. Thus, even if commanders, staffs, and operators did collectively have the right expertise, they may not always have access.

These conditions give rise to two concerns. First, machine output that results in unintended harm can still occur, even if the system is functioning as designed and the human operators are acting with the best of intentions.[20] As Wendell Wallach and Colin Allen state, "As either the environment becomes more complex or the internal processing of the computational system requires the management of a wide array of variables, the designers and engineers who built the system may no longer be able to predict the many circumstances the system will encounter or the manner in which it will process new information."[21]

Thus, even if commanders, staffs, and operators had access to all features of a system, fully accounting for machine behavior in terms of decisions made by human beings would be difficult, if not impossible, especially given that the complexity of interactions increases as processes include increasingly more data.[22] If one cannot tie machine output to human decisions, then an ethical violation may arise for which no one is responsible. Thus, the inability to account fully for machine behavior introduces a "responsibility gap" that threatens to undermine the application of the war convention and dehumanize warfighting.[23]

Communities use norms, such as those expressed by the war convention, to communicate individual accountability to outsiders. But when these norms are not upheld, they frequently die. For instance, if norms associated with timeliness and meeting deadlines were not upheld, then people would likely ignore them. Eventually, if enough people fail to uphold such a norm enough times, the norm will effectively, if not actually, cease to exist.[24]

Similarly, the employment of AI systems risks eroding the war convention. The more AI applications absolve humans of accountability, the greater the

20. Liu, "Refining Responsibility," 340.

21. Wendell Wallach and Colin Allen, "Framing Robot Arms Control," *Ethics of Information Technology* 15, no. 2 (2013): 127.

22. O'Callaghan, e-mail message to author, November 15, 2022.

23. Heather Roff, "Killing in War: Responsibility, Liability, and Lethal Autonomous Robots," in *The Routledge Handbook of Ethics in War*, ed. Fritz Allhoff, Nicholas G. Evans, and Adam Henschke (New York: Routledge, 2013), 355.

24. Pfaff, "Disruptive Technologies."

risk their use will establish a dysfunctional incentive to employ the applications more often and to blame them when something goes wrong, even when a human is responsible. Repeated unaccountable violations could result in the rules rarely being applied, even to humans. Thus, the permissibility of using these systems cannot merely rest on the ability to employ them discriminately and proportionately. Rather, the permissibility must rest on a process of accountability that ensures human responsibility without undermining machine efficiency.[25]

Unfortunately, holding commanders, staffs, and operators responsible for every harm, even when it is not a function of their intention or action, creates dysfunctional incentives. To the extent commanders assess the possibility of such harms as likely—or at least likely enough—commanders will be disincentivized to employ AI systems. This reluctance poses at least two concerns. First, failure to employ the technology could place US forces at a disadvantage, placing soldiers' lives and vital national interests at risk. Second, employing the machine allows operators to continue to update and refine data sets and algorithms to be more precise and deliver better outcomes. Just as soldiers can learn to be more discriminate, proportional, and humane, so can the AI-enabled systems with which they work. Thus, a standard of accountability that is too high fares no better than one that is too low.

Establishing the right level of accountability depends first on defining the effective and ethical functioning of the system. As described earlier, effective and ethical targeting requires one to maximize accuracy and precision in the identification of legitimate targets and to engage the targets in a way that avoids collateral harm, ensuring unavoidable harm is proportional to the value of the military objective. In a human-only targeting system, these conditions mean combatants do not deliberately target protected persons, like noncombatants, or prohibited targets, like hospitals and cultural sites. Combatants must also take steps to minimize any collateral harm, even if doing so means accepting extra risk themselves—though they are not required to assume so much risk an otherwise legitimate operation will fail or they are unable to continue the war effort.[26]

Automation Bias

Of course, one can further close the accountability gap by placing humans at critical points in the process (the current configuration of

25. Pfaff, "Disruptive Technologies."

26. Walzer, *Just and Unjust Wars*, 156.

the systems under consideration here) to ensure they make all decisions. But this configuration potentially slows the process down and raises other concerns. To properly evaluate machine outcomes, humans must be able to trust the information they receive. Sometimes, this trust can be taken too far, and humans may inappropriately subordinate their judgment to that of a less capable machine. Moreover, systems do not have to be very advanced for this inappropriate subordination, known as *automation bias*, to occur.

For example, in the USS *Vincennes* incident, a US Navy ship shot down an Iranian airliner due to automation bias. In 1988, sailors were monitoring the Aegis air defense system—which had the ability to be fully autonomous but was configured for its lowest degree of autonomy at the time—aboard the USS *Vincennes*. The system detected an Iranian jet whose path and radio signature were consistent with civilian airliners; nonetheless, the system registered the aircraft as an Iranian F-14 and, thus, an enemy. Despite indicators the aircraft may have been civilian, the crew trusted the machine and fired.[27] Here, the complexity of machine thinking along with the pressure to act, especially in combat, disposes humans to trust the machine, especially if doing so allows them to avoid responsibility for the action in question. Moreover, at least one study has shown this trust can emerge independent of the machine's reliability. A study conducted in Korea found the effects of institutional pressure, mature information technology infrastructure, and top management support were more significant in building feelings of trust than the effects of machine performance.[28]

Thus, humans who are included in the decision-making process can, but often do not, prevent inappropriate system behavior. Rather than considering machine output to be a judgment requiring justification and explanation, humans often interpret the output as fact. This certainly seemed to be case with the USS *Vincennes*: an aircraft was approaching the ship, but the system judged the aircraft to be an enemy. Based on the context—specifically, the flight path and radio signature—the humans on board should have questioned the machine and aborted the attack.[29]

27. P. W. Singer, *Wired for War: The Robotics Revolution and Conflict in the 21st Century* (New York: Penguin Books, 2010), 125.

28. Hyun-Ku Lee and Hangjung Zo, "Assimilation of Military Group Decision Support Systems in Korea: The Mediating Role of Structural Appropriation," *Information Development* 33, no. 1 (2017), quoted in James Boggess, "More Than a Game: Third Offset and the Implications for Moral Injury," in *Closer Than You Think: The Implications of the Third Offset Strategy for the US Army* (Carlisle, PA: Strategic Studies Institute, US Army War College Press, 2017), 133.

29. David Evans, "Vincennes: A Case Study," *Proceedings* 119/8/1086 (August 1993); and Pfaff, "Disruptive Technologies."

As machine judgments become more complex, conditions for automation bias will likely only get worse. To avoid these conditions, commanders, staffs, and operators will need sufficient expertise to identify possible machine mistakes and investigate them in a timely manner. To facilitate this expertise, designers will have to do their best to ensure the output of AI systems is explainable to at least the operator, if not commander.[30]

This point means, first, those involved will need to be familiar enough with the system to know when information requires corroboration and the system can help. The application of the fuzzy-logic controller both helps close the accountability gap and, by identifying when the system is uncertain relative to the commander's risk level, helps to avoid automation bias by signaling to human operators when corroboration or another kind of verification should be used. Whether these measures are good enough depends on how well the system balances the ethical imperatives discussed earlier in comparison to a human-only process.

Assessing Ethical Performance

Taking for granted a human-only targeting process can find such a balance, we can judge AI-enabled processes by how well they perform relative to the human-only process. If these systems perform as well as or better than the human-only process, then one has good ethical reasons to use them, even if collateral harms occur. In the case collateral harms occur, humans must take steps to ensure the AI-enabled system performed as well as the human-only process.

From an ethical perspective, being clear about what it means to perform more ethically than a human-only system is important, especially when the purpose of the system is to kill and destroy. For instance, a more lethal system may enable a more rapid defeat of the enemy, but, due to performance errors, the system could pose a more significant risk to noncombatants. One could argue a more rapid defeat of the enemy over the long term will result in fewer noncombatant deaths, but such speculation is inadequate to justify the use of AI-enabled systems. For example, similar arguments were made for the strategic bombing of urban areas in World War II. These bombings often did little to undermine enemy resolve and, in some cases, may have strengthened it.[31] Thus, without a demonstrable and necessary

30. Scharre and Horowitz, *Artificial Intelligence*, 11.

31. Gian P. Gentile, *How Effective Is Strategic Bombing?* (New York: New York University Press, 2001), 67, 73, 77; and Barrie Paskins and Michael Dockrill, *The Ethics of War* (Minneapolis: University of Minnesota Press, 1979), 45.

connection between the use of an AI technology and decreased harms to protected persons, the technology's effectiveness will be insufficient to justify its use.

Simply lowering the quantity of collateral harm may not be adequate if, in doing so, one places certain members of a protected population or infrastructure at greater risk. For instance, if a classifier is more likely to mistake noncombatants of a certain ethnicity for enemy soldiers or buildings associated with a particular religion (like churches and mosques) as legitimate targets, then the application of the classifier may reasonably be considered to be unfair. Moreover, such unfairness does not need simply to apply to persons based on ethnicity or faith. For example, classifiers that mistake hospitals and schools as legitimate targets pose a disproportionate risk to medical personnel, patients, educators, and students. Such a system may also be considered unfair, even if it reduces overall collateral harm.

Thus, for AI to be said to perform more ethically than a human process, the AI's results must be fairer. For the AI to be fairer, it must result in fewer rights violations. Defeating the enemy aims to restore the rights of the victims of the enemy's aggression. Avoiding friendly and noncombatant casualties aims to preserve their individual rights to life. If one acts in such a way that some persons affected are better off and no persons affected are worse off, then one is acting fairly, even if the act violates the rights of some of those persons.[32]

The difficulty with AI-enabled systems is that they may result in different collateral harms than human-driven processes because the systems make different kinds of mistakes than humans do. Thus, biases like the kind discussed above may endure in AI- and data-enabled systems. Where these systems are biased, they cannot fulfill the fairness condition discussed earlier because some persons, by virtue of their kind, face a greater disadvantage than others. But, to the extent AI evenly lowers the risk protected persons face, the use of the AI can be said to be fair, even though protected persons still experienced some harm. This point suggests one should look for and resolve system biases and, when doing so is not possible, take the bias into account when setting risk tolerance.

Improving Ethical Performance

Machines of any kind can possess operational and functional traits that demonstrate a commitment to ethical norms. For instance, placing a safety

32. Arthur Isak Applbaum, *Ethics for Adversaries: The Morality of Roles in Public and Professional Life* (Princeton, NJ: Princeton University Press, 1999), 151.

device on a rifle is an operational measure that reflects a concern for individual well-being. Functional traits, on the other hand, can allow the machine to assess ethically significant aspects of how it operates. For example, autopilot takes passenger comfort into account when the machine makes course corrections, limiting the kinds of maneuvers it can make. Autopilot performs this function because the designer cared about passenger comfort, not because the maneuvers might prevent the plane from reaching its destination.[33]

One can improve the ethical performance of AI- and data-enabled systems by ensuring they are robust, specific, and assured. *Robustness* is an operational trait that refers to the system's ability to determine when it is not confident about a prediction and then alert operators to its lack of confidence. *Specification* is another functional trait that refers to the ability of human-machine teams to align machine behavior with human expectations. Finally, *assurance* is a functional trait that refers to the human ability to understand why the system behaves a certain way and how this behavior aligns with the system's purpose.[34]

In the context of the corps targeting process, one can make the system more robust by curating data sets at the right intervals to increase the probability the system will identify legitimate targets. One can also improve specification by building data sets that identify illegitimate targets so they may be avoided. Building such data sets may require refining how commanders establish risk. For instance, as currently designed, such systems establish a probability an image fits one of the categories of legitimate targets, such as a tank. If a system assesses an 80 percent probability a target is a tank, then the probability the target is not a tank is 20 percent. This 20 percent chance does not mean the probability that the target is illegitimate is 20 percent. Thus, in an operational environment where the risk of collateral harm is low, a commander could reasonably choose to engage the target.

Although the risk of collateral harm is low, the risk school buses are operating in the area, for example, is slightly higher. In this case, an 80 percent probability a target is a tank and a 20 percent probability the target is a school bus could reasonably lead the commander to choose not to engage the target, even though the risk of collateral harm is low overall.

33. Wallach and Allen, "Robot Arms Control," 25–26.

34. Tim G. J. Rudner and Helen Toner, *Key Concepts in AI Safety: An Overview* (Washington, DC: Center for Security and Emerging Technology, 2021).

The point here is not that commanders must ensure models can detect illegitimate targets. Given limited resources and time, training data sets to recognize objects like school buses may not be immediately possible. But commanders would be accountable for doing what they can, including avoiding illegitimate targets, to ensure data sets will perform as well as possible. Given commanders' high success rate in this regard, they would also be accountable for where they set risk, which would be a function of the value of the military objective as well as the probability of both collateral harms as well as more specific ones. If a corps were operating in an environment where school buses could be mistaken for tanks, the commander should probably set the risk tolerance lower than if the corps were operating in an environment where the risk of such a mistake was negligible.

One can improve system assurance by ensuring AI literacy among commanders, staff, and operators. In this context, AI literacy means more than understanding how to use AI- and data-enabled systems. Literacy also requires more than understanding how to design and engineer such systems. Rather, data, algorithms, and the systems they support interact in complex ways that change even familiar processes, such as targeting, into something much more complicated and unfamiliar. Making matters more difficult, from a professional perspective, mastering a new technology requires gaining an adequate understanding of how the technology works as well as how its application affects organizational, ethical, and political concerns—both for the military and for the US government, its international partners, and American society. Of course, such knowledge cannot reside in one person. Figuring out where this distribution should lie is a task that should be taken up at both the institutional and operational levels.

Conclusion

Based on observations from three Scarlet Dragon exercises, the process employed by Project Ridgway conforms to these principles. The system has an error rate comparable to, if not lower than, human targeteers. Moreover, humans oversee the operation by confirming targets are legitimate and validating that the corps has appropriate assets to engage. Data is routinely retrained to avoid error and bias, and new data sets and algorithms are rigorously tested. The introduction of the fuzzy-logic controller would not make the system more ethical as much as it would enable commanders to take greater advantage of machine speeds without the kind of loss of control

that gives rise to ethical failure. Of course, no warfighting system is free of risk. But as long as data is properly curated, algorithms regularly updated, and systems properly supervised, they should perform—from an ethical perspective, at least—as well as human-only systems.

— 5 —

Political-Cultural

Political-cultural aspects of military expert knowledge apply to managing the military's relationship with the broader defense community, which includes both public and private organizations as well as society.[1] Thus, political-cultural knowledge applies to civil-military relations as well as servicemembers and DoD civilians. As the sociologist Morris Janowitz opined, the military's future as a profession depends on finding a balance between organizational stability and adaptation to change, whether the change is political or technological. Both types of change impact how the military adapts to fight.[2]

The application of AI-enabled systems can affect civil-military relations in a variety of ways. To the extent they allow for more precise and lethal targeting as well as lower risk to one's own soldiers, the systems can help to establish the expectation that wars should be relatively bloodless and quick. Moreover, to the extent the military has to rely on private-sector expertise to develop and employ the technology, this reliance can affect how society views and values military service as well as who joins. Finally, because short-term integration of AI technologies cannot be effectively centralized, even corps-level organizations must consider the initiative of other DoD entities to avoid unnecessary duplication and conflict.

Expectations

Making war less risky for soldiers and noncombatants while making it more lethal for the enemy suggests the political cost for waging wars

1. Lacquement, "Army Professional Expertise," 217.
2. Morris Janowitz, *The Professional Soldier* (New York: Free Press, 1988), 417.

in the future will be relatively low. As Christian Enemark argues, "Political leaders, having less cause to contemplate the prospect of deaths, injuries and grieving families, might accordingly feel less anxious about using force to solve political problems."[3] Of course, this concern is not unique to AI-enabled systems. Any technology that distances soldiers from the violence they do or decreases harm to civilians will lower the political risks associated with using the technology. The political concern is that when these expectations are not realized, public support for the military effort may fade.

The fading of support may not be a bad thing, especially if political and military leaders have allowed the use of force to proliferate to secure less-than-vital interests. But, to the extent interests are vital, public expectations about the actual costs of war relative these interests can affect the public's support, which is essential to maintaining the political will to continue the fight.

Having raised this concern, not overstating it is important. Precise, lethal, and discrete uses of force can be an important means to defend oneself or even limit escalation. But, as evidenced by public concern over collateral harms from unmanned aerial vehicles (UAVs) in the past, if these strikes are precise, then future strikes need to be even more so.[4] Moreover, as events in Ukraine aptly demonstrate, employing fires in LSCO imposes high costs, regardless of how the targets were identified and engaged. Nonetheless, commanders should remain sensitive to how the use of the technology affects public perceptions and take steps where appropriate to ensure they are accurate relative to any conflict.

Private-sector Expertise

As P. W. Singer observed in 2009, the employment of highly specialized technologies by contractors or civilians may be preferable to the employment of the technologies by the military.[5] Although he was specifically addressing remote technologies that enabled UAV operations, this concern certainly applies to AI and data technologies, some of which, as discussed previously, cannot even be owned by the military. As a result, even if the military could

3. Christian Enemark, *Armed Drones and the Ethics of War* (London: Routledge, 2014), 22–23.

4. Azmat Khan, "Hidden Pentagon Records Reveal Patterns of Failure in Deadly Airstrikes," *New York Times* (website), December 18, 2021, https://www.nytimes.com/interactive/2021/12/18/us/airstrikes-pentagon-records-civilian-deaths.html.

5. Singer, *Wired for War*, 372.

rapidly educate and train personnel to employ enabled systems, it would still have to rely on the private sector for the operation of the systems.

As noted previously, the private sector has developed much of the AI- and data-enabled technology (especially algorithms), and the private sector owns much of it too. The government owns the data that goes into the algorithms as well as the processed information the algorithms produce. But the government neither owns the algorithm nor has direct access to it. Not having this kind of access is not necessarily a problem, especially given that at the corps level, few, if any, personnel are able to understand and manipulate the algorithm. Moreover, this problem is not entirely new. The US military frequently hires contractors where it does not have the right personnel with the right expertise, skills, or abilities. For instance, the US Air Force hired private contractors to fly UAVs for reconnaissance missions because the current workload of active-duty pilots was causing burnout, stress, and other psychological problems.[6]

The issue with UAV pilots is the Air Force does not have enough of them. But the issue with AI and related skills—at the corps level, at least—is no one has them—at least, no one assigned to a position that requires such skills—because the positions do not exist. Moreover, even if the positions did exist, the Army does not have enough personnel to fill the positions. As these positions are created and the personnel trained, this problem should eventually be remedied. But if the military chooses to rely predominantly on contractors, then the concern that critical military expert knowledge lies outside the armed forces will persist.

Such reliance could affect the professional status of the military. If the expert knowledge required to defend the nation is predominantly employed by civilians, then, arguably, the military will not retain its professional status. Rather, the military will likely devolve into a technocratic bureaucracy that manages civilian skills and capabilities while a relative few bear the burden of risk.[7]

Moreover, to the extent these technologies reduce risk to soldiers, the technologies create additional concerns as well. Although risk reduction is arguably the point of military innovation, risk reduction does impact the civilian-military relationship. Society rewards soldiers precisely because they

6. Alex Lockie, "The US Air Force Hired Contractors for Drone Operations, and It's an Ominous Sign," Business Insider (website), September 6, 2016, https://www.businessinsider.com/us-air-force-hired-contractors-for-drone-operations-2016-9.

7. Don M. Snider, "The US Army as Profession," in *Future of the Army Profession*, 2nd ed., 13.

expose themselves to risks and hardships on its behalf.[8] Thus, soldiers who experience neither risk nor sacrifice might be better thought of as technicians than warriors. Although enhancing soldier survivability and lethality will always make moral sense, enhancing them to the point of near-invulnerability (or even the perception of invulnerability) would profoundly alter the warrior identity—which would not necessarily be a bad thing, but militaries need to be prepared to manage it.

Stakeholder Management

In the book *Where Is My Flying Car?*, J. Storrs Hall argues that since the 1970s, the United States has seen a slowdown in innovation due to the "Machiavelli Effect." This effect occurs when entrenched interests defend the status quo against innovations because the resulting change might challenge their influence—and, consequently, power—within the institutional setting.[9] Setting aside whether this description of private-sector innovation is apt, the history of military innovation has plenty of examples. Indeed, the technology modern war depends on—the airplane, tank, and submarine, for example—were all resisted by the established leadership at the time.

As evidenced by the numerous Joint and service-specific organizations that have been established, making the case the Department of Defense has resisted the development and acquisition of AI the way it has perhaps done for other technologies would be difficult. Moreover, the budgets for these new organizations are increasing. For instance, the Joint Artificial Intelligence Center budget went from $89 million in 2019 to $278 million in 2021.[10] But pursuant to Hall's point, James Holmes, a US Naval War College professor, argues the Department of Defense still has an entrenched culture that can stifle innovation in the acquisition and integration of AI (and other emerging technologies, for that matter).[11] Although Holmes does not offer an example of DoD culture stifling innovation, the addition of new stakeholders to established ones can create confusion and inefficiency in the acquisition process.

8. Singer, *Wired for War*, 331–32.

9. J. Storrs Hall, *Where Is My Flying Car?* (San Francisco: Stripe Press, 2021).

10. Government Accountability Office (GAO), *Artificial Intelligence: Status of Developing and Acquiring Capabilities for Weapon Systems*, GAO-22-104765 (Washington, DC: GAO, 2022), 9, 21–24.

11. James Holmes, "Want to Innovate in the Military? Beware of the 'Machiavelli Effect,'" RealClearDefense (website), February 15, 2022, https://www.realcleardefense.com/articles/2022/02/15/want_to_innovate_in_the_military_beware_of_the_machiavelli_effect_816846.html.

For example, according to a March 2022 Government Accountability Office report, long acquisition timelines, which can limit the usefulness of technology, are hampering the Department of Defense's ability to acquire AI. In addition, the department is having trouble acquiring useable data with which to train algorithms.[12] Complicating matters, the Department of Defense has difficulty tracking, coordinating, and assessing its AI-related programs. As of the writing of this report, the Department was able to identify 685 AI projects, but it could not provide their estimated costs. Part of the problem is AI is often part of a program rather than a program itself.[13] Even where AI is a program itself, the Department of Defense also has difficulty deploying useable applications to end users.[14] As defense reform advocate William C. Greenwalt points out, the department's centralized, predictive program budgeting processes discourage the kind of risk taking necessary to develop new technologies rapidly. As a result of this centralized process, Greenwalt states, producing "marginal incremental capability improvements to existing systems" takes 15 years.[15]

A centralized but uncoordinated acquisition process is obviously not a good foundation for trust, neither for soldiers who must use the technology nor for Americans who depend on the use of the technology for their defense. Fortunately, as defense scholars Margarita Konaev and Tate Nurkin point out, the Department of Defense is aware of these challenges and has set up a number of programs to address them. Yet the scholars still opine that DoD AI acquisition professionals are "working on a common goal but, unfortunately, on parallel tracks."[16]

Nevertheless, Konaev and Nurkin find some reasons to be optimistic. For instance, the Department of Defense recently consolidated the role of chief data and AI officer, who reports directly to the under secretary of defense, with the Joint Artificial Intelligence Center, the Defense Digital Service, and the Office of Advancing Analytics. This

12. GAO, *Artificial Intelligence*, 4.

13. GAO, *Artificial Intelligence*, 16.

14. GAO, *Artificial Intelligence*, 21.

15. William C. Greenwalt, "Competing in Time: How DoD Is Losing the Innovation Race to China," American Enterprise Institute (website), March 9, 2021, https://www.aei.org/op-eds/competing-in-time-how-dod-is-losing-the-innovation-race-to-china/.

16. Lauren C. Williams, "How 'Cultural Artifacts' Impede DOD's Ability to Go Big on AI," FCW (website), May 26, 2022, https://fcw.com/defense/2022/05/how-cultural-artifacts-impede-dods-ability-go-big-ai/367459/; and Margarita Konaev and Tate Nurkin, *Eye to Eye in AI: Developing Artificial Intelligence for National Security and Defense* (Washington, DC: Atlantic Council, May 2022), 19–21.

consolidation was an effort to "deconflict overlapping authorities" that make planning and executing AI development programs difficult. Collectively, these organizations were responsible for data management and coordination, finding digital solutions for internal data and security issues, aggregating data, and conducting data analytics. These functions are now all under one roof. Moreover, the Department of Defense is expanding the use of alternative acquisition methods, such as the Defense Innovation Unit and the Air Force's AFWERX, to bridge the gap between the private sector and the defense sector more effectively.[17]

Although any efficiencies such consolidation may bring will be welcome, not stifling bottom-up initiatives like Project Ridgway will be important. Consolidating top-down processes will be important to the development of new technologies and applications. But even with process reform, these efforts will most likely take a long time. Bottom-up efforts can make use of current technology, and doing so can accelerate soldier familiarity with AI technologies as well as enhance soldiers' AI skills beyond that which more centralized programs can accomplish. Taking advantage of these efforts could be critical to maintaining advantage over adversaries like China, which has undertaken its own bottom-up efforts and has fewer barriers to centralized control.

Conclusion

Political-cultural knowledge requires knowing how the use of an emerging technology will affect public expectations about the use of force, how these expectations affect the way society perceives military service, and how other DoD efforts to employ the technology may affect one's own efforts. Moreover, political-cultural knowledge requires senior military leaders to understand how these shifting expectations will affect civil-military relations and military culture because they will affect who joins the military in addition to how they serve.

To the extent the use of AI technology reduces risks to soldiers and noncombatants, it reduces the political risks associated with using force. Thus, senior military leaders will need to manage senior civilian leaders' expectations to ensure the use of the technology does not risk escalation

17. Konaev and Nurkin, *Eye to Eye*, 4–5; and Michael C. Horowitz and Lauren Kahn, "Why DoD's New Approach to Data and Artificial Intelligence Should Enhance National Defense," Council on Foreign Relations (website), March 11, 2022, https://www.cfr.org/blog/why-dods-new-approach-data-and-artificial-intelligence-should-enhance-national-defense.

to a wider conflict. Senior military leaders must also manage public expectations about collateral harms to ensure continued support. Perhaps most importantly, senior military leaders will need to manage expectations about the effectiveness of AI technology so civilian leaders do not rely on it too much and the public does not become frustrated by a lack of results. The public is not likely to trust a military that cannot deliver results and that imposes risks on soldiers and noncombatants alike.

— 6 —

Conclusion

Developing and employing new military technologies is a part of being a military professional. Indeed, military history is very much a story of technological innovation and the need for soldiers to learn how to operate new systems. So much about integrating AI is not new. As with the tank, the airplane, and even the crossbow, soldiers learn to use and employ technology over time, industry learns to produce it in sufficient quantity and quality, and senior leaders learn to employ it to strategic effect. As mentioned earlier, the difference between AI technologies and their disruptive predecessors is the former's capability to improve a wide range of military weapons, systems, and applications. Because of this potential pervasiveness, nearly all soldiers will have to become adept at some level to employ the technology effectively and ethically. As this technology expands in application, war will be as much about managing data as it is managing violence.

This pervasiveness also raises questions about human development and talent management. Although training programs will eventually produce more knowledgeable soldiers, and the personnel systems will improve their ability to manage the soldiers, limits to the knowledge and skills uniformed personnel can acquire will still exist, especially at the operational level. Although it is not intended to establish firm guidelines, this discussion has identified much of the knowledge soldiers will need to obtain. For example, soldiers will need to know how to curate and train databases so they are useful for the tasks the soldiers are performing. Doing so requires ensuring data is accurate, complete, consistent, and timely. Using this data requires proficiency in applying the conditions described in the recommended model card, and the proficiency helps to ensure the algorithm acting on the data performs in an effective and ethical manner.

Of course, trust cannot be ensured by policies and processes alone. Commanders, staffs, and operators need to know what they are being trusted to do as well as what they are trusting systems to do. Commanders, staffs, and operators are trusting AI systems to identify legitimate targets and to avoid the identification of illegitimate targets. The humans involved in the process must use this information in a way that balances the need to defeat the enemy with the equally important imperatives of avoiding friendly and noncombatant casualties. Finding this balance will require making judgments about how much risk persons in these categories should bear.

Systems enabled by AI can facilitate the finding of this balance as long as the humans involved in the process are able to interact with the system effectively. When integrating human control over machine processes, one is frequently forced to choose between control and speed: the more human control that is imposed, the slower the system will operate. But this study has found this dilemma to be false. Although a trade-off between human control and speed may be necessary in some circumstances, human input is necessary if the system is to function optimally.

Achieving optimal performance first requires commanders to ensure staffs and operators understand the competence of the model, the quality of data shaping the model's understanding, and the model's demonstrated performance in the operational environment. Although it may not make the system more precise or accurate, achieving these tasks should make the system better able to assign probabilities to the output. Second, commanders need to determine how much risk to the mission, friendly combatants, and enemy noncombatants is appropriate. This determination can be complex. A critical mission might require tolerating more friendly and noncombatant casualties. Similarly, a low density of noncombatants may enable higher risk tolerance, even if the mission is not as urgent. Finding this balance is the human's job.

But with the help of the fuzzy-logic controller described earlier, commanders can better determine when an AI-enabled system can be trusted to perform some targeting steps without human oversight. Moreover, the logic of the interaction can be constructed to find multiple different configurations of human-machine interaction to ensure optimal use of the system while avoiding unnecessary harms. Giving commanders the option to intelligently and responsibly expedite the targeting process when needed will be essential during LSCO, and the design proposed in this report achieves this objective. This achievement will be especially important in the future when, to protect the force and achieve mission objectives, commanders will be faced with many

time-sensitive targets and operating conditions that dictate the assumption of more risk.

Abundant work remains in the development of enough soldiers with the right skills to take advantage of AI technologies fully. The current talent management program is not yet up to managing this challenge, though multiple promising programs are poised to meet the needs eventually. Yet, for the most part, these programs are designed to meet requirements at the institutional level, where decisions about the Army-wide acquisition of AI and related technologies are made. But how these skills will filter down to the operational Army, where educated and skilled personnel manage and maintain systems that rely on these technologies and play a critical role in innovating as these technologies advance, is less clear.

Although the use of AI in targeting does not violate current laws of war, it does raise some ethical concerns. In the context of the targeting systems under discussion, primary among these ethical concerns is the accountability gap and automation bias. The first concern is critical to answering the central question, "On what basis can commanders trust AI-enabled systems such that commanders may be held accountable for the use of these systems?" Automation bias and data hygiene are related to the accountability gap because where these concerns are present, they undermine measures for the meaningful human control commanders may want to emplace. Commanders can close the accountability gap by, first, ensuring personnel are properly educated, skilled, and trained to curate the relevant data and, second, by ensuring the risk the commanders allow accurately reflects the demands of balancing mission accomplishment with the protection of friendly soldiers and noncombatants. Commanders can also reduce the chance of automation bias and its potential effects by signaling to the humans involved in the process when the machine needs more oversight.

Being a professional means more than simply providing a service and being accountable when something goes wrong. Professionals must also understand how various stakeholders, including the public and government and private-sector entities, interact and compete with the profession. Given the potential of these technologies, military professionals must first learn to manage expectations as they evolve the technology and its applications. Because this evolution impacts the character of professional work, military professionals must also pay attention to how those outside the profession value, reward, and support the work. Thus, as the US military continues to integrate AI and data technologies into various operations, the test of its professionalism will lie in the ability both to have

expertise and to build the kinds of institutions that can continue to develop, maintain, and certify this expertise in ways that both meet the defense needs of the American people and reflect their values.

Select Bibliography

Abbott, Andrew. *The System of Professions: An Essay on the Division of Expert Labor.* Chicago: University of Chicago Press, 1988.

Independent High-level Expert Group on Artificial Intelligence. *The Assessment List for Trustworthy Artificial Intelligence*. Brussels: European Commission, 2020.

Lilly, John H. *Fuzzy Control and Identification*. New York: John Wiley and Sons, 2011.

Osoba, Osonde, and William Welser IV. *An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence*. Santa Monica, CA: RAND Corporation, 2017.

Ozar, Anne C. "The Plausibility of Client Trust of Professionals." *Business and Professional Ethics Journal* 33, no. 1 (2014).

Roberts, Huw et al. "The Chinese Approach to Artificial Intelligence: An Analysis of Policy, Ethics, and Regulation." *AI and Society* 36 (2020).

Schmidt, Eric, and Bob Work. *Final Report.* Washington, DC: National Security Commission on Artificial Intelligence, 2021.

West, Darrell M., and John R. Allen. *Turning Point: Policymaking in the Era of Artificial Intelligence*. Washington, DC: Brookings Institution Press, 2020.

About the Project Director

Dr. C. Anthony Pfaff (colonel, US Army retired) is the research professor for strategy, the military profession, and ethics at the US Army War College Strategic Studies Institute and a senior nonresident fellow at the Atlantic Council. He is the author of several articles on ethics and disruptive technologies, such as "The Ethics of Acquiring Disruptive Military Technologies," published in the *Texas National Security Review*. Pfaff holds a bachelor's degree in philosophy and economics from Washington and Lee University, a master's degree in philosophy from Stanford University (with a concentration in philosophy of science), a master's degree in national resource management from the Dwight D. Eisenhower School for National Security and Resource Strategy, and a doctorate degree in philosophy from Georgetown University.

About the Researchers

Lieutenant Colonel Christopher J. Lowrance is the chief autonomous systems engineer at the US Army Artificial Intelligence Integration Center. He holds a doctorate degree in computer science and engineering from the University of Louisville, a master's degree in electrical engineering from The George Washington University, a master's degree in strategic studies from the US Army War College, and a bachelor's degree in electrical engineering from the Virginia Military Institute.

Lieutenant Colonel Bre M. Washburn is a US Army military intelligence officer with over 19 years serving in tactical, operational, and strategic units. Her interests include development and mentorship; diversity, equity, and inclusion; and the digital transformation of Army intelligence forces. Washburn is a 2003 graduate of the United States Military Academy and a Marshall and Harry S. Truman scholar. She holds master's degrees in international security studies, national security studies, and war studies.

Lieutenant Colonel Brett A. Carey, US Army, is a nuclear and counter weapons of mass destruction (functional area 52) officer with more than 33 years of service, including 15 years as an explosive ordnance disposal technician, both enlisted and officer. He is an action officer at the Office of the Under Secretary of Defense for Policy (homeland defense integration and defense support of civil authorities). He holds a master of science degree in mechanical engineering with a specialization in explosives engineering from the New Mexico Institute of Mining and Technology.

The United States Army War College educates and develops leaders for service at the strategic level while advancing knowledge in the global application of Landpower.

The purpose of the United States Army War College is to produce graduates who are skilled critical thinkers and complex problem solvers in the global application of Landpower. Concurrently, it is our duty to the Army to also act as a "think factory" for commanders and civilian leaders at the strategic level worldwide and routinely engage in discourse and debate on the role of ground forces in achieving national security objectives.

The Strategic Studies Institute publishes national security and strategic research and analysis to influence policy debate and bridge the gap between military and academia.

The SSI Live Podcast Series provides access to SSI analyses and scholars on issues related to national security and military strategy with an emphasis on geostrategic analysis. https://ssi.armywarcollege.edu/ssi-live-archive

The Center for Strategic Leadership provides strategic education, ideas, doctrine, and capabilities to the Army, the Joint Force, and the nation. The Army, Joint Force, and national partners recognize the Center for Strategic Leadership as a strategic laboratory that generates and cultivates strategic thought, tests strategic theories, sustains strategic doctrine, educates strategic leaders, and supports strategic decision making.

The School of Strategic Landpower provides support to the US Army War College purpose, mission, vision, and the academic teaching departments through the initiation, coordination, and management of academic-related policy, plans, programs, and procedures, with emphasis on curriculum development, execution, and evaluation; planning and execution of independent and/or interdepartmental academic programs; student and faculty development; and performance of academic-related functions as may be directed by the Commandant.

The US Army Heritage and Education Center makes available contemporary and historical materials related to strategic leadership, the global application of Landpower, and US Army Heritage to inform research, educate an international audience, and honor soldiers, past and present.

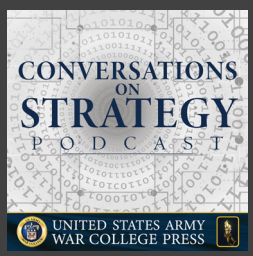
The Army Strategic Education Program executes General Officer professional military education for the entire population of Army General Officers across the total force and provides assessments to keep senior leaders informed and to support programmatic change through evidence-based decision making.

US ARMY WAR COLLEGE PRESS

The US Army War College Press supports the US Army War College by publishing monographs and a quarterly academic journal, *Parameters*, focused on geostrategic issues, national security, and Landpower. Press materials are distributed to key strategic leaders in the Army and Department of Defense, the military educational system, Congress, the media, other think tanks and defense institutes, and major colleges and universities. The US Army War College Press serves as a bridge to the wider strategic community.

The US Army War College Press produces two podcast series. Decisive Point, the podcast companion series to the US Army War College Press, features authors discussing the research presented in their articles and publications. Visit the website at: https://ssi.armywarcollege.edu/decisive.

Conversations on Strategy, a Decisive Point podcast subseries, features distinguished authors and contributors who explore timely issues in national security affairs. Visit the website at: https://ssi.armywarcollege.edu/cos.



ISBN 1-58487-846-0
9 781584 878469 90000 >

https://press.armywarcollege.edu