Research Report

JONATHAN P. WONG, ALEXANDER C. HOU, MICHAEL MILLER,
KATIE A. WILSON, EMILY LATHROP, SYDNEY KESSLER,
SAM WALLACE, EMILY YODER

One Team, One Fight

Volume I, Insights on Human-Machine Integration for the U.S. Army

For more information on this publication, visit **www.rand.org/t/RRA2764-1**.

Published by the RAND Corporation, Santa Monica, Calif.

Library of Congress Cataloging-in-Publication Data is available for this publication.

ISBN: 978-1-9774-1515-8

Cover: Samarion Hicks / U.S. Army.

About This Report

This report documents research and analysis conducted as part of a project entitled *Human Machine Integration and Artificial Intelligence*, sponsored by U.S. Army Futures Command. The purpose of the project was to investigate the kinds of difficulties that the Army might encounter as it attempts to pair humans with artificial intelligence (AI) algorithms to accomplish specific warfighting tasks. We sought to make recommendations for how these potential obstacles can best be overcome and ensure that the Army effectively creates AI systems that will integrate well with the Army soldiers who must interact with them.

This research was conducted within RAND Arroyo Center's Strategy, Doctrine, and Resources Program. RAND Arroyo Center, part of RAND, is a federally funded research and development center sponsored by the United States Army.

RAND operates under a "Federal-Wide Assurance" (FWA00003425) and complies with the *Code of Federal Regulations for the Protection of Human Subjects Under United States Law* (45 CFR 46), also known as "the Common Rule," as well as with the implementation guidance set forth in U.S. Department of Defense (DoD) Instruction 3216.02. As applicable, this compliance includes reviews and approvals by RAND's Institutional Review Board (the Human Subjects Protection Committee) and by the U.S. Army. The views of sources utilized in this study are solely their own and do not represent the official policy or position of DoD or the U.S. Government.

Acknowledgments

We thank our sponsor, MG Stephanie Ahern, for the guidance she provided over the course of this project. We also thank our action officer, LTC Timothy Cheung, for providing day-to-day support. At RAND, we thank Molly Dunigan and Jeff Martini for providing program oversight. Buzz Phillips and Zev Winkelman provided invaluable feedback as peer reviewers.

Summary

Issue

Recent advances in artificial intelligence (AI), machine learning, and robotics have raised the possibility that the profession of arms will soon include integrating human soldiers with AI-enabled machines and applications as part of the collective whole. Machines and software applications enabled by AI are starting to demonstrate capabilities that are relevant to military settings, such as moving autonomously through complex urban traffic and creating startlingly humanlike and interesting derivative works through large language models.

However, this does not mean that such developments can be implemented in military settings smoothly. The practice of building cohesive small units is no easy endeavor. The best small units cohere to the point where one soldier recognizes the silhouette and gait of another in the dark of a patrol base in an instant.[1] The best staffs internalize their commander's style and specific needs over time. Integrating humans and machines in military contexts will likely draw from civilian parallels but will also require substantial contextualization.

To that end, Army Futures Command asked us to consider the implications of human-machine integration. In particular, our objective was to investigate the kinds of difficulties the Army might encounter as it attempts to pair humans with AI algorithms to accomplish specific warfighting tasks. Our goal was to make recommendations addressing how these potential obstacles can best be overcome and ensure that the Army effectively creates AI systems that will integrate well with the soldiers who must interact with them.

---

1 The intricacies of small unit dynamics have been best described in memoirs such as James R. McDonough, *Platoon Leader: A Memoir of Command in Combat*, Presidio Press, 1985; and E. B. Sledge, *With the Old Breed: At Peleliu and Okinawa,* Presidio Press, 2007.

Approach

Our approach was largely qualitative and consisted of two elements. The first was an exploration of the Army's thinking on human-machine integration at the time of writing. We conducted this exploration by reviewing literature and refined those insights into two exploratory vignettes on movement to contact and urban warfare. From these vignettes, we identified key themes that characterize the Army's thinking.

The second element was an extensive literature review on emerging principles of AI, including the "starting conditions" of human and machine cognition, challenges to the approaches used at the time of writing, and more-promising approaches to human-machine integration.

Key Findings

- **Explainable AI and cognitive forcing functions are less effective than expected.** The first and most interesting finding from the academic literature was that, at the time of writing, efforts to increase human trust in AI through engineered solutions leave something to be desired. Attempts to increase trust through greater transparency are offset by increased deference to AI outputs even in cases when such deference is not warranted. This inability to increase human trust is especially critical because trust in AI is a key concern for the Army.
- **Enduring ideas of design, signaling, and mental models remain promising.** Integrating humans and machines will likely require a more deliberate approach of building shared mental models and transactive memory between humans and machines more deliberately over time. This suggests that integration should be a much longer process in which the human parts of the team must share more of their mental model explicitly and make the implicit transactive memory systems that form in units more explicit to enable greater machine integration. Human-centered design and human-machine signaling can positively influence the interface between humans and machines, but they might only be helpful if they are supported by a deeper and more deliberate effort for the *human* to explain their actions so that the *machine* can learn and adapt to them.

- **The integration of humans and machines in the Army in the realm of ground combat planning and execution is likely to be slower than that in the commercial sector.** This is largely because of the Army's outsize need for trust, the imperative of maintaining operational security, and policy decisions related to keeping humans in the loop of consequential decisions. This is observable in our narrative exploration and obvious when we consider the role of trust in soldiering as a profession; these factors are the foundation on which the Army philosophy of mission command rests. This is not to say that human-machine integration in the Army is impossible, only that the outsize need for trust will require more time to gain enough exposure to AI-enabled systems and applications such that the human elements of a team are comfortable with them.
- **Efforts to integrate humans and machines in the Army are predominantly focused on making the machine fit the human.** However, there are efforts to change the design of human organizations to work better with machines. From a tactical level, ideas of changing formations to accommodate humans and machines (e.g., creating "machine minder" billets) are worth pursuing. Ideas that affect the Army at a more basic organizational level (such as modifying the RDT&E [research, development, test, and evaluation] and acquisition enterprise to make it more responsive to the frequent iteration and refinement AI requires) are also worth investigating. Greater human-machine integration will also require soldiers and commanders to be better trained and educated about their machines. We asserted previously that human-machine integration is a sociotechnological system for which experience and familiarity with specific machines in specific contexts are necessary. Nevertheless, soldiers must have an underlying understanding of how AI works, which can only be accomplished through professional military education. This is especially necessary for commanders and small unit leaders, who must contextualize machine outputs and behaviors in the course of commanding an integrated human-machine unit. Relatedly, soldiers must gain mastery over specific machines themselves. This is especially true for soldiers performing duties and filling roles related to managing AI systems

within units. This will require progressively more advanced training, much as the Army does for other weapon systems.

Recommendations

- **Once machines are fielded to units, integrate them like new soldiers.** Once the Army is prepared to integrate machines in operational units, it should not treat that integration simply as it would for a weapon system. Rather, integration should feature elements that are more akin to integrating new soldiers into their units after basic training. That would entail integrating machines for less complex or lower-level mission-essential tasks with the unit before proceeding to more complex ones. It would mean taking the time for humans to make implicit details of the way they operate explicit to the machine to help build shared mental models. This might include more detailed articulation of the Commander's Critical Information Requirements (CCIRs), boundaries for decisionmaking, and standard operating procedures. It would also mean aligning updates to the machine to improve its AI with the unit training cycle; here, preserving whatever transactive memory and shared mental model that exists between the humans and machines in the unit is more valuable than any incremental gains in AI accuracy.
- **Prepare for variation.** If the Army does integrate machines like new soldiers, it should be prepared to accept the likelihood that the characteristics, performance, and value of AI will vary even across units of the same type. Our research reinforces the notion that human-machine integration is a sociotechnological construct. Therefore, variation should be expected because the *humans* vary. This is already the case in terms of the different standard operating procedures, performance, and culture in different units. This will have inevitable consequences on the way the AI-enhanced unit builds transactive memory and a shared mental model. The Army should be prepared for this.
- **Monitor how machines gain and maintain human trust.** Although there are numerous technical and procedural challenges that must be addressed as the Army integrates machines into formations,

most can conceivably be overcome with technical refinements to the machines themselves. However, one critical challenge is solely in the hands of humans: gaining and maintaining trust between humans and machines. Therefore, the Army should carefully monitor human-machine trust as machines become more integrated in formations and staffs. Just as the U.S. Department of Defense (DoD) used monitoring frameworks to (1) understand the progress of integrating women in combat arms roles and (2) highlight especially challenging issues before they developed, it might consider a similar process to monitor the integration of humans and machines.[2] Such an approach implies a relatively slow pace of integration between humans and machines. Although this might be in tension with tactical imperatives to field machines rapidly, we conclude that humans' trust in machines can only be built at a deliberate pace because it takes time for humans to build trust. We also believe that such monitoring should be carried out by Training and Doctrine Command to emphasize that monitoring human-machine integration is as much a behavioral issue as it is a technical one.

- **The human must remain the dominant partner in integrated units.** The integration of humans and machines in the U.S. Army is likely to be one filled with promise, challenges, progress, and setbacks as DoD continues to develop AI-enabled applications and weapon systems. Many of these developments will be derived from commercial advances. However, one constant should be that the role of machines should remain one that supports humans in the profession of arms. Without human preeminence in combat, tactical operations lose their connection to the political purposes of war and merely become sophisticated violence devoid of meaning.

[2] The issues of integrating women in combat arms unit and integrating machines and humans in units generally are not substantively comparable. However, this should not stop the Army from adapting basic methods of monitoring implementation of the former to assist in the latter.

Contents

Figure and Tables

Figure

Tables

CHAPTER 1

Introduction

The profession of arms is fundamentally a collective endeavor. From Baron von Steuben drilling the Continental Army at Valley Forge into a cohesive whole, to "little groups of paratroopers" forming up across the bocages of Normandy, to soldiers manning the outposts in the Korengal Valley in Afghanistan, soldiering has been about creating cohesive teams of individuals working together to accomplish tasks and missions while campaigning, during crises, and in conflict.

Recent advances in artificial intelligence (AI), machine learning (ML), and robotics have raised the possibility that the profession of arms will soon include integrating human soldiers with AI-enabled machines and applications as part of the collective whole. Machines and software applications enabled by AI are starting to demonstrate capabilities that are relevant to military settings, such as moving autonomously through complex urban traffic and creating startlingly humanlike and interesting derivative works through large language models.

However, this does not mean that such developments can be implemented in military settings smoothly. The practice of building cohesive small units is no easy endeavor. The best small units cohere to the point where one soldier recognizes the silhouette and gait of another in the dark of a patrol base in an instant.[1] The best staffs internalize their commander's style and specific needs over time. Integrating humans and machines in military contexts will likely draw from civilian parallels but will also require substantial

[1] The intricacies of small unit dynamics have been best described in such memoirs as James R. McDonough, *Platoon Leader: A Memoir of Command in Combat*, Presidio Press, 1985; and E. B. Sledge, *With the Old Breed: At Peleliu and Okinawa,* Presidio Press, 2007.

contextualization to understand what forms of integration are most useful to the Army and how they might be achieved.

To that end, Army Futures Command asked us to consider the implications of human-machine integration. In particular, our objective was to investigate the kinds of difficulties the Army might encounter as it attempts to pair humans with AI algorithms to accomplish specific warfighting tasks. Our goal was to make recommendations addressing how these potential obstacles can best be overcome and ensure that the Army effectively creates AI systems that will integrate well with the soldiers who must interact with them.

Methods and Scope

Our approach was largely qualitative and consisted of two elements. The first was an exploration of the Army's thinking on human-machine integration at the time of writing. We conducted this exploration this by reviewing literature and refined those insights into two exploratory vignettes, on movement to contact and urban warfare. From these vignettes, we identified key themes that characterize the Army's thinking.

The second element was an extensive literature review on emerging principles of AI, including the "starting conditions" of human and machine cognition, challenges to current approaches, and more-promising approaches to human-machine integration.

For the purposes of this research, we define *machines* as autonomous or semi-autonomous ground robots, uncrewed aerial vehicles, or software applications that perform decision support tasks using AI-enabled features. Examples we use in this report appear in Table 2.1 in Chapter 2.

It is important to note that that this research was scoped to focus on the tactical level to ensure tractability. We did not consider the operational or strategic implications of human-machine integration or how those implications might in turn affect the tactical level.[2] We also structured the vignettes to consider the role of several different types of machines to accomplish tac-

[2] One example of analysis focusing on the strategic implications of AI is Yuna Huh Wong, John Yurchak, Robert W. Button, Aaron B. Frank, Burgess Laird, Osonde A.

tical tasks as they exist today; new tasks that might arise from the integration of humans and machines were not considered.

A Note on Terminology

The Army has focused on using the term *human-machine integration* over *human-machine teaming* to highlight the primacy of the human over machines. Our research takes the same stance but uses a variety of terms (e.g., *teams*, *interactions*) to explain certain points or arguments more clearly.

Organization of This Report

Chapter 2 outlines two tactical vignettes we developed to explore the implications of these topics on the Army context and identifies salient features of human-machine integration to tactical operations. Chapter 3 describes the findings of an extensive literature review that examined a variety of topics relevant to understanding human-machine integration, such as mechanisms of human cognition, explainable AI, limitations of robotics, and human-centered design. We conclude in Chapter 4 by identifying overarching findings and recommendations for the Army.

Osoba, Randall Steeb, Benjamin N. Harris, and Sebastian Joon Bae, *Deterrence in the Age of Thinking Machines*, RAND Corporation, RR-2797-RC, 2020.

CHAPTER 2

How Does the Army Think Humans and Machines Will Integrate?

To understand the difficulties the Army might encounter with integrating human-machine teams, we first must understand the Army's mental model for how human-machine integration will play out throughout its echelons. In this chapter, we explain our process for using fictional narrative to explore the mental model and then characterize five elements of that model.

Exploring Army Perspectives on Human-Machine Integration Through Fictional Narrative

Although we initially intended to explore this topic through a literature review, we realized that the Army is still forming its opinions and policy stances on the subject and that our understanding of the state of human-machine integration in the Army required a more immersive approach. To that end, we used fictional narrative to explore the topic.[1] The process of writing narrative required us to mentally fill in the blanks about details and considerations related to how humans and machines might interact that might not be explicitly stated in the literature. Once the initial drafts of the narratives were written, the editing process also allowed us to examine rough hypotheses in a more unconstrained way.

[1] We adapted a method described in Lisbeth A. Pino Gavidia and Joseph Adu, "Critical Narrative Inquiry: An Examination of a Methodological Approach," *International Journal of Qualitative Methods*, Vol. 21, April 2022.

The narratives are not meant to be prescriptive or predictive; rather, they are an exercise in imagining how the benefits of challenges of effective human-machine integration might arise in future, Army-specific scenarios. Although these narratives are written to adhere to a standard of realism relating to Army operations, and the AI tools described in these scenarios are based in current and projected capabilities, realism is tabled in some instances in favor of creating a well-crafted situation to illustrate human-machine interactions.

The different scenarios drive each narrative: a movement to contact scenario and an attack in urban terrain scenario. Both are company reinforced-size operations that include several AI-enabled machines and applications that are used to varying effect. The scenarios were selected in consultation with Army Futures Command.[2]

Emerging Themes

The development of the narratives revealed multiple themes that we analyzed through an anthropological method of pile sorting.[3] In this process, our research team members (a different group than the ones who wrote the narratives) reviewed the narratives and identified key observations or themes from the text. These observations or themes were then written down on slips of paper; themes that were most like each other were sorted into piles. Our team examined each pile and described the overarching theme of each. Five themes were evident and are described next.

[2] See Sydney Kessler, Sam Wallace, Emily Yoder, and Jonathan P. Wong, *One Team, One Fight: Ground Combat Narratives of Human-Machine Integration,* RAND Corporation, RR-A2764-2, forthcoming.

[3] Hung-Wen Yeh, Byron J. Gajewski, David G. Perdue, Angel Cully, Lance Cully, K. Allen Greiner, Won S. Choi, and Christine Makosky Daley, "Sorting It Out: Pile Sorting as a Mixed Methodology for Exploring Barriers to Cancer Screening," *Quality & Quantity*, Vol. 48, No. 5, September 2014.

The Army's Goal Is That Human-Machine Integration Will Reduce Risk and Improve Awareness

The most prominent theme in our exploration was in understanding that there are three main benefits the Army sees in the integration of humans and machines. The first is that machines can reduce risk to soldiers by taking on some tasks that are dangerous for humans to do on the battlefield. Many of those tasks need to be conducted while under fire (casualty evacuation, for example). The task can be accomplished with less risk to the force overall and might free human soldiers to take on other tasks.

A second benefit is in increasing human awareness and aiding in deliberate decisionmaking on staffs and in tactical operations centers (TOCs). Here, AI-enabled intelligence, surveillance, and reconnaissance platforms allow for more proliferated and finer-grain situational awareness on the battlefield. In the TOC, AI applications help fuse data into a more complete and more coherent common operational picture to aid staffs and the commanders they serve in making decisions. The Army envisions that these machines will aid in planning and might actually be necessary to keep up with the speed of operations in the future. An important distinction to this type of benefit is that decisions are left to humans; the machines merely present information to assist.

Finally, the Army also sees benefits in using machines to increase awareness and aid human decisionmaking on the battlefield. Experience that is often gained through painful firsthand experience could be replicated through the use of AI-enabled decision aids for small unit leaders faced with discrete tactical problems during the course of operations.

Human-Machine Integration Will Take Several Forms

The literature is dominated by three types of systems that will instantiate human-machine integration: AI applications that support deliberate decisionmaking, applications that will support tactical decisionmaking, and machines that work with humans on the battlefield. We examined a variety of sources, including Army documents, leader statements, trade press, and examples provided by Army Futures Command, to synthesize a list of machines and applications. Table 2.1 describes the ones used in the vignettes.

TABLE 2.1

Autonomous Machines and AI Applications Considered in Vignettes

Warfighting Functions	Type	Full Name	Description
Intelligence	AI application	Depiction of Enemy Disposition and Strength	This AI tool provides a predictive representation of enemy positions and capabilities using several inputs: enemy doctrine, terrain analysis, sensor data (IR, thermal, EO, audio), and contact reports from units.
Intelligence	AI application	Automated Red Team Partner	This is an AI system designed to red-team problems and identify potential enemy courses of action. It takes the form of a chat interface and is networked with several other AI applications.
Mission Command, Intelligence, Protection	AI application	Improved Small Unit Common Operating Picture	This is a suite of visual tools incorporated among multiple displays—screens in the unit's vehicles, on soldier's goggles, and/or on forearm tablets. Various overlays can be incorporated to visualize terrain, friendly and enemy positions, movement, capability, routes, fires planning, control measures, battlespace geometries, etc.
Mission Command	AI application	Automated Report Consolidation System	This AI tool is like a large language model. It takes contacts, situation reports, or other forms of reports from lower echelons and combines them into a report for the receiving unit to transmit up the chain of command.
Movement and Maneuver, Intelligence, Protection	AI application	Red Teaming of Operations Orders	This AI-enabled model and simulation tool, available at the company level, would run simulations of a company's scheme of maneuver against a variety of enemy courses of action and offer suggested refinements using the commander's desired end state.

Table 2.2—Continued

Warfighting Functions	Type	Full Name	Description
Movement and Maneuver, Intelligence, Sustainment, Protection	AI application	Route Analysis and Reconnaissance	This is an AI-enabled route-planning tool that takes routes of movement from an operations order and draws from available sources of information to analyze route conditions for route trafficability. The tool identifies and recommends alternate routes to meet stated end objectives (e.g., “get to X location within four hours”). The tool can also analyze known or suspected enemy positions, weapon capabilities, and doctrine to represent the likelihood of threats from weapons and nonkinetic effects. The tool presents this information to a commander via display.
Movement and Maneuver	AI application	Rapid Breach Support Tool	This is an AI system designed to recommend breaching techniques given a scenario. It networks with other AI applications (RAR, PACT, and DEDS), but is poorly designed.
Protection, Movement and Maneuver, Intelligence	AI application	Autonomous Counter Air Tool	This tool consists of small EO/IR sensors on vehicles that autonomously sense for UAS threats and orient and employ electronic warfare or directed energy weapons. These sensors would be networked within a unit, able to share threat information automatically and deconflict targeting either autonomously or in response to an order. This system can engage air threats without human approval.
Sustainment	AI application	Projected Ammunition Consumption Tool	This is an AI-enabled tool that assists in sustainment planning. It takes mission timelines as given in an operations order and determines the amount of ammunition needed, down to the fire team or single crew-served weapons level. The tool provides recommendations to the command team for ammo supply requests and how to distribute ammo.

Table 2.2—Continued

Warfighting Functions	Type	Full Name	Description
Sustainment	AI application	Rapid Support Tool	This AI tool would be available at the company level and identify battalion- and brigade-level assets that could be useful to support a mission and automatically link commanders or controlling units across units to rapidly establish new support relationships.
Mission Command, Movement and Maneuver	Other	Augmented Reality Integrated Display	These augmented reality goggles, similar to an Integrated Visual Augmentation System, display overlays of different AI applications to the user.
Intelligence, Movement and Maneuver, Protection	UAS	Group 1 UAS	This Group 1 UAS is organic to the company level (allocated 90 total systems to the armored company in MTC scenario). They are unarmed but have EO/IR and thermal sensors. These UASs can be networked with one another and with other systems at higher echelons to conduct reconnaissance, providing targeting data.
Intelligence, Movement and Maneuver, Protection	UAS	Group 2 UAS	This Group 2 UAS is organic to a battalion. Its characteristics are similar to existing group 2 UASs in use but have more-capable sensors and swarming ability.
Sustainment	UAS	Aerial Mule	These are large UASs designed to carry a payload and be used for resupply.
Intelligence, Movement and Maneuver, Protection	Uncrewed ground vehicle	Durable Autonomous Ground Vehicle	These small ground vehicles are capable of being thrown with EO and LIDAR sensors.

Table 2.2—Continued

Warfighting Functions	Type	Full Name	Description
Movement and Maneuver	Uncrewed ground vehicle	Optionally Manned Fighting Vehicle	This is a future version of the XM-30 (previously called the optionally manned fighting vehicle) currently under development. This system fulfills the enduring role of an infantry fighting vehicle, providing transportation and fire support for dismounted infantry.
Movement and Maneuver, Intelligence, Sustainment	Uncrewed ground vehicle	Company Ground Combat Vehicle	These are tracked vehicles with a modular design that allows them to be tailored to a variety of roles, including logistics, reconnaissance, hazard clearance, direct fire, and air defense. They are autonomous, with the ability to connect to and receive tasking from the units' manned vehicles.

NOTE: EO = electro-optical; IR = infrared; LIDAR = Light Detection and Ranging; MTC = movement to contact; UAS = uncrewed aerial system.

The Army's Main Concerns Are Related to Cognitive Burden and Trust

The benefits of human-machine integration do not come without challenges. We identified two main challenges that the Army perceives as it considers the integration of humans and machines for operations across all types of human-machine integration. The first is that integration comes with a degree of cognitive burden. This burden might land on a staff officer in a TOC trying to understand why an algorithm is recommending a particular course of action. It might also instantiate itself on the battlefield with a small unit leader trying to manage a balky autonomous system while leading their unit. In any case, there is a concern that the machine will require an inordinate amount of attention from the human, making it more of a liability than a benefit.

The second main challenge concerning the Army is about trusting the machine to perform as expected more generally. Trust might be eroded by the machine failing under adverse conditions posed by the enemy (e.g., communications denied or degraded environments that hinder machines that require connectivity) or hallucinating an obviously incorrect tactical choice. It might also be that working with the machine will require the human to calibrate their trust in the machine's behaviors such that the human is not unduly surprised during operations. Finally, there is a general concern that the machine might behave or come to a decision in a way that is not easily explainable, not intuitively correct, or some combination of both.

Predominant Mitigation Is to Make AI Work Better for the Army Through Human-Centered Design and Explainability

To mitigate some of these challenges, the Army has predominantly focused on changing the machine to work better with the human. The Army's policy as of this writing is to design machines and algorithms around the soldier. This might be accomplished by making algorithmic decisions explainable to the soldier. It might involve making the interface and outputs look and feel more like interfaces and outputs that soldiers are used to interacting with. Another mitigation the Army is considering is making the level of autonomy adjustable such that the human can modify it to fit the risk tolerance that they are comfortable with in a given situation. In any case, the general thrust of these mitigations is to seek design solutions in the algorithm or machine itself.

Less Attention Is Paid to Redesigning the Army for AI

Conversely, our exploration of the narratives also revealed a less dominant theme of making the humans work better with the machine. This theme usually surfaces through changes in formations that happen over time, as was the case when the Army transitioned to mechanized forces or introduced helicopters for vertical lift. For instance, the inclusion of a human who tends to the machines in small units was a commonly discussed idea. Finally, there is also the recognition that experimentation and iteration will be essential to creating optimal AI-enhanced units. This will require more changes to the way the Army develops, acquires, and sustains machines such that it facilitates iteration and experimentation in a more fluid way across different institutional and tactical echelons.

War Is Fundamentally Human

We close this chapter with a final theme: War is an inescapably human endeavor. Aside from the fact that U.S. Department of Defense (DoD) policy emphasizes the role of human control over autonomous systems, the role of humans is clearly evident in our analysis of the narratives.[4] Planning and executing combat operations without humans as the predominant actors renders such operations meaningless; without humans in decisive roles, the links between the strategic and geopolitical purposes of combat are severed, and the foundation of the Army's doctrine and understanding of its strategic roles dissolves.[5]

---

4 DoD policy states that "autonomous and semi-autonomous weapon systems will be designed to allow commanders and operators to exercise appropriate levels of human judgment over the use of force." See DoD Directive 3000.09, *Autonomy in Weapon Systems*, Office of the Under Secretary of Defense for Policy, January 25, 2023, p. 3.

5 For a foundational view of the Army's role in war, see Army Doctrinal Publication 1, *The Army*, Headquarters, Department of the Army, July 2019.

CHAPTER 3

Insights on Human-Machine Integration from Nonmilitary Settings

Having considered the Army's vision for human-machine integration in Chapter 2, we now turn to broader developments of the integration drawn from academic literature. That literature is vast and sprawling but can be organized into four categories that we will examine in this chapter. The first two categories are concise articulations of the "starting conditions" of humans and machines that are relevant to the discussion on integration. The third category is an examination of two prominent efforts to improve integration. The first effort revolves around making AI decisions more explainable, known as explainable AI (XAI). The second effort relates to choices in the user interface that require humans to intervene at points that are otherwise automated; these are known as cognitive forcing functions. The last category involves an examination of the broader theoretical basis for XAI and cognitive forcing functions.

Human Starting Conditions

Recognition Primed Decisionmaking

When considering how AI/ML tools or autonomous systems might affect the performance of teams of soldiers, one critical factor to take into account is how the commanders of Army formations typically make decisions. In the past, researchers studying decisionmaking often assumed that the process would take an analytical form, with the individuals in question carefully evaluating the strengths and weaknesses of their potential options before choosing a course of action.[1] Often, decisionmaking under operational conditions instead follows a process of *recognition-primed decisionmaking.* In this process, the commander subconsciously absorbs the context of their decision and generates a course of action using their training and experience. They then consciously evaluate whether that course of action satisfices their objectives. If it does, they execute it; if it does not, they subconsciously generate an alternative course of action that rectifies the deficiencies and determine whether that option satisfices the needs of the situation.

This mechanism of decisionmaking primarily relies on cognitive instinctive thinking (known in behavioral economics as *System 1*).[2] Because the soldier faces critical time pressure and delaying a decision might cause opportunities to be missed or for the situation to rapidly deteriorate, there is no time to consciously engage with the decision at hand. Instead, this process will be shaped by the soldier's instincts, experience, prejudices, and intuitions. One approach to effective human-machine integration, then, might be to ground human-machine interactions in the instinctive behavior of cognitive System 1.

Cognitive Burden

It is often assumed that maximizing the level of autonomy of machines in a human-machine team will maximize team performance. However, a review of the literature reveals that finding the level of autonomy that maximizes team performance or the performance of the unit overall might not be that straightforward. Thus, although it might be natural to push to maximize the autonomy of, for instance, robotic ground vehicles to reduce the burden on a human operator tasked with commanding or supervising them, it might not actually yield the best overall team performance.

The results of a computational simulation experiment illustrating this point are depicted in Figure 3.1. The experiment involved a simulation environment known as Blocks World for Teams. The environment consists of

1 Gary A. Klein, "A Recognition Primed Decision (RPD) Model of Rapid Decision Making," in Gary A. Klein, Judith Orasanu, Roberta Calderwood, and Caroline E. Zsambok, eds., *Decision Making in Action: Models and Methods*, Ablex, 1993.

2 Daniel Kahneman, *Thinking, Fast and Slow*, Farrar, Straus and Giroux, 2011.

nine rooms with a set of blocks of different colors randomly distributed among them and a goal area where the blocks are to be delivered. For this experiment, the team consisted of a human player and an agent that was under the command of the human. The human and agent players worked together to deliver the blocks to the goal area in a specified order. Both players were limited in their situational awareness. They could only see the blocks in the room they were in, and they could not see each other. The experiment was conducted with the agent employing four different autonomy "treatments," with each successive treatment capable of greater levels of autonomy.[3]

As shown in Figure 3.1, team performance improves at first with increasing levels of autonomy but later declines. The initial improvements are driven by reductions in the burden on the human player from commanding the agent with successive increases in the agent's level of autonomy. However, this comes at a cost of increasing "opacity" and decreasing situational awareness that eventually outweighs the benefit of burden reduction and hinders overall team performance.[4] Although this experiment was run under artificial conditions, it does give some indication of factoring in human-machine interdependence when developing autonomous systems for joint human-machine activities.

The opacity or situational awareness issues with more-autonomous AI can also serve to increase the cognitive burden on a human teammate. This can occur when humans are called on to step in when errors occur or for extreme cases the algorithm is not capable of handling. The human teammate might not have a good read on the situation when they are called on to intervene and might not be prepared to act accordingly. Consequently, the increased autonomy and automation can actually result in less effective teams.[5]

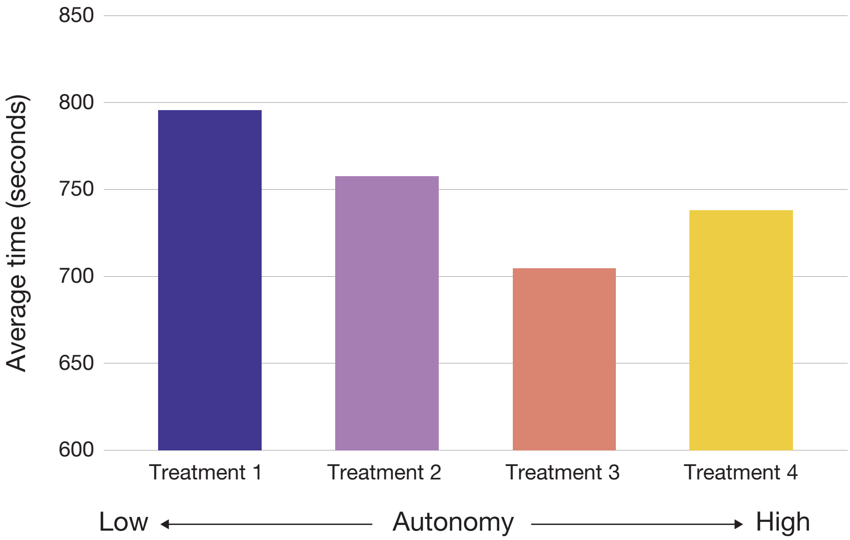
---

[3] Matthew Johnson, Jeffrey M. Bradshaw, Paul J. Feltovich, Catholijn M. Jonker, Birna van Riemsdijk, and Maarten Sierhuis, "Autonomy and Interdependence in Human-Agent-Robot Teams," *IEEE Intelligent Systems*, Vol. 27, No. 2, March 2012.

[4] Johnson et al., 2012. With higher levels of autonomy, the number of times the human player was blocked from entering a room by the agent player increased significantly.

[5] Raja Parasuraman and Victor Riley, "Humans and Automation: Use, Misuse, Disuse, Abuse," *Human Factors*, Vol. 39, No. 2, June 1997.

FIGURE 3.1
Blocks World for Teams Experiment Results



SOURCE: Adapted from Johnson et al., 2012, p. 48.

Given the potential for AI systems to increase the cognitive burden on a human teammate, care must be taken when introducing these systems to the key members of Army teams. These key members are either more central to team workflow or do more tasks. Increasing their cognitive burden or forcing them to context switch more rapidly can be particularly detrimental to team performance.[6] Increasing the cognitive load on these individuals will be disproportionately impactful.

Human Biases in Human-Machine Integration

A cognitive bias can be broadly defined as a human's deviation from rationality in judgement. This might lead to perceptual distortion, inaccurate

6 Stephen E. Humphrey, Frederick P. Morgeson, and Michael J. Mannor, "Developing a Theory of the Strategic Core of Teams: A Role Composition Model of Team Performance," *Journal of Applied Psychology*, Vol. 94, No. 1, January 2009.

judgement, illogical interpretation, and irrational conclusions and can influence decisionmaking that might not reflect an appropriate response from the perspective of an unbiased observer. Cognitive biases have been extensively researched in human-to-human interaction in the fields of cognitive science, social psychology, and behavioral economics. But as AI technology advances and becomes increasingly accessible and communication between humans and machines becomes commonplace, there is a need to further understand and analyze the cognitive biases that can arise in humans during human-machine integration (see Table 3.1).

One of the most common cognitive biases is *anthropocentric thinking*, in which users generalize an AI's or robot's competency in the same way the user would judge another human.[7] Within this domain, humans might exhibit either anthropomorphism of the body or anthropomorphism of the mind. In *anthropomorphism of the body*, the human user might treat the AI or robot differently depending on its appearance. For example, if the user interacts with a robot that resembles a dog, the user might interact with the machine through whistling and simple one-word commands. Similarly, if the robot resembles a human, the human user might interact with the machine the same way they communicate with another human, using complex speech, facial expressions, and body language. Even though the same AI might be imbedded in these two machines, the human user could judge the appearance (body) of the machine, which then influences the user's communication techniques.

In *anthropomorphism of the mind*, on the other hand, the user might inaccurately project their emotions onto a machine during interactive communication; humans might assume that the machine is able to "feel" the same way a human would. For example, when a Boston Dynamics robot was physically kicked by a human, viewers commented this was cruel behavior and took pity on the robot.[8] Related to this phenomenon is the *Computers as Social Actors Theory*, which proposes that when machines exhibit anthropomorphic characteristics when interacting with humans, such as empathy, interaction quality, and humor, the user is more likely to trust the AI,

7 Sangbae Kim, "AI and Robots Are a Minefield of Cognitive Biases," *IEEE Spectrum*, July 19, 2021.

8 Kim, 2021.

TABLE 3.1

Common Cognitive Biases, with Relevance

Cognitive Bias	Relevance in AI	Implications
Anthropomorphic thinking (body or mind)	Human tendency to generalize a robot or AI's competency the same way a human would judge a human's competency.	Users of the robot or AI might inaccurately assume that the machine should respond to the user in a way that humans do.
Automation bias	Tendency for human operators to over rely on AI automation. Users with the lowest levels of experience with AI are the most averse to its capabilities, whereas the users with the highest level of experience over rely on AI automation.	Some inexperienced users will distrust and therefore not use AI, whereas others will over-trust AI and rely on its capabilities when it should be managed by human judgment.
Hallucination	Tendency for AI to make false claims, including irrelevant, nonsensical, or factually incorrect answers, because of "learning" from information not verified as correct.	The user might ask AI to give a response to a question and assume that the machine gives the correct answer consistently, even though AI is providing the user with false information.
Human self-confidence	Low human self-confidence can result in an increased willingness to use AI; high self-confidence can cause humans to reject input from accurate automated systems.	Users without depth of knowledge in a particular subject will rely too heavily on AI for information, whereas experts in a particular field will remain critical of information from AI.
Computers as social actors theory	When AI exhibits anthropomorphic social characteristics, including empathy and interaction quality, humans are more likely to trust the machine.	If the AI mimics interaction between two humans, the user might treat the machine as if it were another human.
Immediacy behaviors	Socially oriented gestures intended to increase interpersonal closeness, such as proactivity, active listening, and responsiveness.	These behaviors are perceived as signs of machine intelligence and contribute to trust between the human and machine.
Gartner Hype Cycle	Most novel innovations will progress thorough a pattern of overenthusiasm, followed by disillusionment and eventual productivity.	Users might initially overestimate the utility of AI, followed by a period of suspicion and underuse. With time, the user will find a middle ground between the two.

and the human-machine interaction is more likely to resemble an interaction between two humans. If the AI or robot produces *immediacy behaviors*, socially oriented gestures and cues that include active listening, responsiveness, and attentiveness, an increase of trust can form between the human and the machine.

Another common cognitive bias is *automation bias*, in which "environmental factors such as task difficulty, background knowledge—familiarity, knowledge, experience—with AI, trust and confidence in AI, and self-confidence determine how much humans rely on machines—or other humans—when making decisions."[9] Furthermore, Horowitz and Kahn (2023) theorizes that a version of the Dunning-Kruger effect is at play, in which the user's aversion to the machine is highest when they know very little about the machine. Once a user gains an understanding of the machine, their aversion changes to an automation bias in favor of the machine. Finally, users who have high knowledge of the machine regain a level of aversion.

AI chatbots have also been shown to relay irrelevant, nonsensical, or factually incorrect answers when communicating with a user, known as *hallucination*. This phenomenon is thought to occur because the AI "learns" information from explicitly or implicitly biased text online that contains both factual and nonfactual sources to inform its decisions.[10] The result is that AI can answer questions with erroneous answers using convincing messages that mimic the text from which the information was collected.

Another common cognitive bias is known as *human self-confidence*, in which low human self-confidence can result in an increased willingness to use AI, whereas user high self-confidence can cause humans to reject input from accurate automated systems.[11] This might result in an over-reliance

---

9 Michael C. Horowitz and Lauren Kahn, "Bending the Automation Bias Curve: A Study of Human and AI-Based Decision Making in National Security Concerns," arXiv, arXiv: 2306.16507, June 30, 2023, p. 3.

10 Cade Metz, "What Makes A.I. Chatbots Go Wrong?" *New York Times,* March 29, 2023.

11 Leah Chong, Guanglu Zhang, Kosa Goucher-Lambert, Kenneth Kotovsky, and Jonathan Cagan, "Human Confidence in Artificial Intelligence and in Themselves: The Evolution and Impact of Confidence on Adoption of AI Advice," *Computers in Human Behavior*, Vol. 127, February 2022.

or under-reliance on AI technology depending on the user's background, educational level, and technical expertise in a particular field. For example, a senior airline pilot might prefer to land an airplane manually, whereas a novice pilot might prefer to engage autopilot to ensure a safe landing.

Finally, users of novel technology are subject to the *Gartner Hype Cycle*, which theorizes that technological innovations progress through an initial period of overenthusiasm, followed by disillusionment, and eventually plateau as the technology's applicability, relevance, and use is adopted by a broader market.

Machine Starting Conditions

We next discuss the limitations of autonomous systems at the time of this writing. Discussing the limitations of these systems is necessary when considering AI-enhanced units, because it enables users to make informed decisions when allocating tasks between humans and machines to maximize the performance of the team. In this section, we first categorize actions taken by autonomous systems into four basic categories, then discuss their limitations at the time of writing. Finally, we end by discussing which limitations will likely see progress by 2040 and which are longer-term limitations.

Four Categories of Autonomous Systems

Each task given to an autonomous system can be broken down into four categories,[12] which closely mirror the OODA (observe, orient, decide, act) loop process used for military decisionmaking:

- information acquisition (sensing, monitoring, and collecting information)
- information analysis (processing, prediction, and analysis)
- decision selection (a choice of action is made using the analysis)
- action implementation (the decision is carried out).

12 Raja Parasuraman, Thomas B. Sheridan, and Christopher D. Wickens, "A Model for Types and Levels of Human Interaction with Automation," *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, Vol. 30, No. 3, May 2000.

Robots and autonomous systems complete tasks across multiple categories—a more “fully autonomous” system will complete a larger portion of a task loop (including transitioning between task categories) with limited human input (human-in-the-loop), a human as a supervisor (human-on-the-loop), or without human input.

Key Limitations and Mitigation Measures

Using a literature review, we synthesized several key limitations of the autonomous systems that exist at the time of this writing. This was accomplished by examining the capabilities and limitations of commercially available robotic systems. Additionally, we examined recent publications in the academic research community to discover capabilities that are in the research stage and have not proliferated in the commercial robotics sector. The following are a few key limitations, organized by task category:

- *Information analysis*. Analysis failures might occur in novel situations or when the system encounters edge cases that were not included in the design and testing stages. Adversarial attacks can be used to alter data collection or data processing to force an AI system to produce a tampered output. This could look like incorrect predictions for predictive systems, untrustworthy outputs in generative AI, or other types of harmful output behavior.[13]
- *Decision selection*. Strategic-level decisions are difficult, especially when the correct decision is highly dependent on individual circumstances (context-dependent decisions). Current autonomous systems can reliably make decisions in highly constrained situations, for which the decision process can be fully mapped in advance and the environment is predictable.[14] In situations that fall outside this scope, autono-

[13] Apostol Vassilev, Alina Oprea, Alie Fordyce, and Hyrum Anderson, *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*, National Institute of Standards and Technology, NIST AI 100-2e2023, January 2024.

[14] For example, movement of goods inside warehouses, routine patrol in mostly static environments.

mous decision selection is a challenge, and human-in-the-loop is still necessary.

- *Action implementation*. Small ground robots lack the agility to reliably navigate through extreme environments.[15] Additionally, many ground robots cannot recover from failure without human intervention. This has restricted commercial ground robot use cases to simple environments (e.g., flat or gently sloped ground) with regularly sized obstacles (e.g., stairs).

Mitigation Measures and Projecting to 2040

Failures in information analysis have the potential to be remedied by extensive testing and evaluation in real-world scenarios and red-team testing to discover and mitigate adversarial attacks.[16]

Limitations for action implementation can be mitigated by preferring airborne platforms to ground robots, because air platforms can avoid interactions with difficult terrain. Enhanced ground robot mobility is an active area of research,[17] and advancements in robot control using AI might soon allow robots to robustly navigate difficult terrain.[18]

According to our analysis of the literature, limitations related to decision selection have less of a clear path forward and are likely to be a longer-term challenge. This gap highlights the need for AI-enhanced units, in which humans are responsible for strategic-level decision selection and autonomous systems complete actions in other task categories.

---

15 For example: large obstacles, rough or unstable terrain, environments with degraded perception.

16 Vassilev et al., 2024. See also John Launchbury, "A DARPA Perspective on Artificial Intelligence," Defense Advanced Research Projects Agency, March 19, 2017.

17 Guang-Zhong Yang, Jim Bellingham, Pierre E. Dupont, Peer Fischer, Luciano Floridi, Robert Full, Neil Jacobstein, Vijay Kumar, Marcia McNutt, Robert Merrifield, et al., "The Grand Challenges of *Science Robotics*," *Science Robotics*, Vol. 3, No. 14, January 2018.

18 See, for instance, Takahiro Miki, Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter, "Learning Robust Perceptive Locomotion for Quadrupedal Robots in the Wild," *Science Robotics*, Vol. 7, No. 62, January 2022.

When dividing tasks between a human and a robot in an AI-enhanced unit, it is important to consider the strengths and limitations of human and robot teammates so that each team member can be assigned tasks at which they excel. Considering tasks through the lens of which task category they fall into might help with this determination.

Reducing Manpower: Moving Toward Multi-Robot Control

Modern robots are often controlled on a one-to-one basis (one operator is responsible for one robot). Robots with fewer autonomous functions are often tele-operated, and robots with more autonomous features can complete a greater number of tasks without human intervention. Shifting to a one-to-many model (one operator is responsible for many robots) has the potential for enhancing operational efficiency and reducing manpower by enabling a single operator to manage multiple robots simultaneously.

Increasing Autonomy is Necessary for Multi-Robot Control

Increasing autonomy is necessary for this shift to take place, because it is not feasible for one operator to control multiple robots at the same time. Autonomy can be measured by the amount of human intervention that is necessary for operation.[19] Consider the case of one operator controlling two robots: The operator could only devote one-half of their time to each robot, so each robot would need to be operating autonomously at least 50 percent of the time (on average). One operator controlling five robots would require each robot to operate autonomously 80 percent of the time, with an operator dedicating 20 percent of their time to each robot.

Case Study: Autonomous Vehicle Operations in San Francisco

In practice, what does multi-robot control look like? Self-driving car companies might offer a few insights. Waymo, for example, is an autonomous driving company operating in the city of San Francisco that offers ride-

[19] Jenay M. Beer, Arthur D. Fisk, and Wendy A. Rogers, "Toward a Framework for Levels of Robot Autonomy in Human-Robot Interaction," *Journal of Human-Robot Interaction*, Vol. 3, No. 2, 2014.

hailing services with a fleet of driverless cars. These cars operate fully autonomously (without human intervention) in most cases, but Waymo has a remote team on hand ready to take control in cases in which the autonomous car runs into trouble.

Handoff of control to a human operator can occur either when the autonomous vehicle does not know what to do and requests an operator to take control or when an operator proactively takes control because they perceive that the vehicle is about to make a mistake or get stuck. When a human operator does take control, they are not tele-operating the vehicle but instead providing direct intervention.[20]

What Is the Process for Moving Toward Multi-Robot Control?

For autonomous car companies, the progression toward full autonomy begins with one-to-one operation, first with a driver physically in the vehicle ready to take over, then progressing to a remote operator outside the vehicle (still at a one-to-one ratio).[21] As a system becomes more autonomous, the operator-to-robot ratio can be scaled down, dictated by the ratio of time that the system operates with and without human intervention. Cruise (another self-driving car company) reported that it uses a ratio of about 15 to 20 driverless vehicles to one remote human operator. Cruise also reported that its vehicles need human intervention at a frequency of every four to five miles,[22] illustrating how far these vehicles are from operating without human input.

[20] Lora Kolodny, "Cruise Confirms Robotaxis Rely on Human Assistance Every Four to Five Miles," *CNBC,* November 6, 2023.

[21] The California Department of Motor Vehicles Autonomous Vehicle Tester Driverless Program requires autonomous test vehicles to be *continuously* supervised by a human test driver or remote operator during initial testing phases.

[22] Kolodny, 2023.

Challenges to Current Approaches

Current DoD efforts to foster human-machine integration center on two concepts: complementary skills and cognitive forcing functions. We explain challenges to these approaches in this section.

Explainable AI

One leading theory for how to create effective AI-enhanced units has centered around the pursuit of XAI. As the Defense Advanced Research Projects Agency (DARPA) defines it, XAI consists of "AI systems that can explain their rationale to a human user, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future."[23] XAI has many applications to aiding human decisionmaking.

Intuitively, it makes sense that these characteristics would be essential to allowing humans to understand where AI systems have complementary skills and helping them decipher when and why they should accept the recommendations from AI systems. As one study of XAI put it,

> Arguably, the most important contribution of explanations is not to convince users to adopt recommendations . . . , but to allow them to make more informed and accurate decisions about which recommendations to utilize . . . . If users are convinced to accept recommendations that are subsequently found to be lacking, their confidence in the system will rapidly deteriorate. A good explanation is one which accurately illuminates the reasons behind a recommendation and allows users to correctly differentiate between sound proposals and inadequately justified selections.[24]

[23] David Gunning and David W. Aha, "DARPA's Explainable Artificial Intelligence (XAI) Program," *AI Magazine*, Vol. 40, No. 2, 2019, p. 44.

[24] Mustafa Bilgic and Raymond Mooney, "Explaining Recommendations: Satisfaction vs. Promotion," paper presented at the Tenth International Conference on Intelligent User Interfaces, January 2005, p. 1.

In the military context, DARPA's XAI program concurs with this assessment, arguing that,

> Advances in machine learning . . . promise to produce AI systems that perceive, learn, decide, and act on their own. However, they will be unable to explain their decisions and actions to human users. This lack is especially important to the DoD, whose challenges require developing more intelligent, autonomous, and symbiotic systems. [XAI] will be essential if users are to understand, appropriately trust, and effectively manage these artificially intelligent partners.[25]

Consequently, many studies exploring how to make effective AI-enhanced units have focused on how AI/ML models can generate useful explanations and how humans react to explanations from AI.

Despite the expectation that XAI would be a key factor in making effective AI-enhanced units, studies exploring the effect of explanations on the performance of AI-enhanced units have come to very different conclusions. Instead, most studies have found that adding explanations for the AI's behavior result in AI-enhanced units that are less accurate and less effective compared with the AI alone. This result has occurred both when the AI is integrated with teams of experts—such as doctors or judges—or with teams of laypeople performing tasks that do not require specialized knowledge.

These experiments have confirmed that the primary effect of adding explanations to AI is a substantial increase in the likelihood that a human will defer to the output of the AI. Unfortunately, humans have not proven able to calibrate this deference to an understanding of whether the AI has generated a correct output or an incorrect one. In fact, the content of the AI's explanation seems to be nearly irrelevant in producing this phenomenon. Humans are just as likely to defer to an AI when it produces a *placebo explanation*—an explanation that contains no useful information about how or why the AI generated the result—as they are to defer to an AI that generates a meaningful explanation describing the factors it took into account when producing its answer. The simple fact that the AI has provided any

[25] Gunning and Aha, 2019, p. 44.

explanation for its answer seems to have a similar effect on how humans interact with the model.[26]

To understand why AI explanations fail to improve team effectiveness, we should consider two core concepts in terms of how AI models behave. The first concept is the task decision boundary for the problem to which AI has been applied. The *task decision boundary* describes the ground-truth reality of the dividing line between samples of one object class from another. For example, if the purpose of an AI model is to differentiate between photographs of cats and dogs, the task-decision boundary is a mathematical function that conceptually places all photos of cats on one side of the boundary and all photos of dogs on the other.

Next, the *model decision boundary* describes the dividing line between where the AI model will assign objects to one object class instead of another. Importantly, because most AI models have less than 100 percent accuracy, not all these assignments will be correct—the AI model will incorrectly place the label *dog* on some photographs of cats, and vice versa. Consequently, the *model error* is defined as the difference between the task decision boundary and the model decision boundary.

When AI models provide an explanation, the effect is to reinforce human confidence in the model decision boundary because the model is justifying why the classification decision it has made is correct. However, this explanation provides no context about the task decision boundary or the model error: The human gains no additional information about what the correct answer is in reality or the gap between the output of the model and the correct answer. AI explanations cannot provide such information: Neither AIs nor humans can ever "know what they don't know." Given this context, it is not surprising that AI explanations increase human deference to AI—but do not generate complementary team performance.

---

[26] A study that recruited 1,500 participants to test the accuracy of an AI-assisted support tool for simple tasks showed that explanations increased team performance when the AI was correct but decreased accuracy when the AI was incorrect. See Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besimra Nushi, Ece Kamar, Marco Tulio Ribeiro, Daniel S. Weld, "Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance," *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, May 2021.

Cognitive Forcing Functions

In human-human interaction, there is evidence that suggests that people are poor detectors of deception and have the tendency to infer that a message is honest independently of its veracity.[27] This *truth-default theory* is an example of a heuristic in which people are inherently gullible in communication and an individual has the tendency to default to the truth when discriminating between a truth and a lie. Similarly, when humans interact with XAI, users are also subject to the truth-default theory and assume that information received by the AI is correct.

The growing body of evidence that XAI results in automation deference has sparked research interest in identifying ways to counteract this effect. Because this deference appears to be an artifact of cognitive System 1 (in which humans are incorrectly relying on instinctive heuristics to trust the AI), one study explored the effectiveness of shifting the interactions with the AI from a System 1 interaction into a System 2 interaction.[28] Although this is in tension with the idea of grounding human-machine interactions in System 1 thinking to enable more timely decisionmaking on the battlefield, there might be benefits of System 2 interactions that are worth the risk.

To instantiate this, researchers used *cognitive forcing functions*—"interventions that are applied at the decision-making time to disrupt heuristic reasoning and thus cause the person to engage in analytical thinking."[29] Cognitive forcing functions can take a variety of forms. For example, asking a person to "think out loud" while making a decision or to make a pro versus con list evaluating potential alternatives can shift their thought process from System 1 into System 2. In this case, the experiment asked the research subjects to make an initial decision without any assis-

---

[27] Bella M. DePaulo, James J. Lindsay, Brian E. Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper, "Cues to Deception," *Psychological Bulletin,* Vol. 129, No. 1, 2003; Timothy J. Luke, "Lessons from Pinocchio: Cues to Deception May Be Highly Exaggerated," *Perspectives on Psychological Science*, Vol. 14, No. 4, July 2019.

[28] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Gajos, "To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making," *Proceedings of the ACM on Human-Computer Interaction*, Vol. 5, No. CSCW1, April 2021.

[29] Buçinca, Malaya, and Gajos, 2021, p. 2.

tance from the AI tool. After making their decision, the subject was shown the recommendation from the AI algorithm with an explanation of how the AI had reached its conclusion. Given this new context, the individual had to decide whether to change their answer to match the AI or whether they should stick with their initial result.

The result of this experiment demonstrates the limits of this approach. Adding cognitive forcing functions to the experimental setup did successfully reduce instances of *overtrust* (in which the human defers to an incorrect answer from the AI model). However, instances of *undertrust* (in which the human ignores a correct answer from the AI model and instead chooses their incorrect answer) increased by approximately the same amount. Thus, the net performance of the team remained the same. Ultimately, it does not seem to matter whether humans use their cognitive System 1 or System 2 to engage with the AI output. The AI-enhanced unit is unable to achieve complementary behavior because the humans lack the necessary context to understand what the right answer actually is (the task decision boundary), and this fact appears to prevent the human component of the team from understanding how to leverage the complementary skill sets found within the team.

Reconsidering Underlying Principles of Current Approaches Might Yield More Promise

Although current efforts, such as XAI and cognitive forcing functions, are likely to result in unintended and less optimal outcomes, they can also be considered to be examples or instantiations of more fundamental principles of human-machine integration. Revisiting those principles and gleaning their insights might yield more promise.

Complementary Skills

One potential benefit of forming AI-enhanced units is that humans and AI algorithms often have complementary strengths.[30] Although AI applications can accomplish astounding feats, humans often have tacit knowledge and context that is quite difficult to teach to a computer (common sense).[31] In some instances, the human might even have a superior ability to reason about an issue or sense parts of their environment compared with an AI algorithm or autonomous system.[32]

Correctly combining these divergent capabilities will create a team that can accomplish more than either element of the team could accomplish on its own. On the other hand, an AI-enhanced unit that is unable to recognize and leverage complementary skill sets often performs less effectively than if one of its components (either the human component or the machine component) were simply allowed to perform the task alone. Consequently, discovering how humans can recognize and appropriately use the unique abilities of AI applications and autonomous systems is an essential part of fully exploiting the potential of human-machine integration.

Human-Centered Design

Human-centered design prioritizes alignment between user needs and the design of objects, tools, or other products.[33] This approach matters across a spectrum of users, from individuals interacting with everyday things, to specialists operating in domains such as health care, aviation, and the

---

30 Dominik Dellermann, Philipp Ebel, Matthias Söllner, and Jan Marco Leimeister, "Hybrid Intelligence," *Business & Information Systems Engineering*, Vol. 61, No. 5, October 2019.

31 Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman, "Building Machines That Learn and Think Like People," *Behavioral and Brain Sciences,* Vol. 40, 2017. A more tangible example might be found in Katherine Tangalakis-Lippert, "Marines Fooled a DARPA Robot by Hiding in a Cardboard Box While Giggling and Pretending to Be Trees," *Business Insider,* January 29, 2023.

32 Bansal et al., 2021.

33 Don Norman, *The Design of Everyday Things*, rev. ed., Basic Books, 2013.

military. When designed appropriately, objects and devices are matched to users' capabilities and the environments in which they are used.[34]

Principles of Human-Centered Design

The process for human-centered design is multidisciplinary and requires contributions from individuals including (but not limited to) subject-matter experts, engineers, industrial designers, marketers, and interface designers. Experts across academia, government, and industry offer perspectives on how to best approach a human-centered design effort. Notably, behavioral science is important for understanding how user psychology can guide design, including how people see, read, remember, think, and focus attention and how people feel, make mistakes, and make decisions.[35] Over years of research, enduring principles for design have emerged from identified usability heuristics by the Nielson Norman Group and others.[36] These enduring principles can guide designers as they develop new systems, and we describe those principles next.

The first enduring principle is to design for *instinctual behavior*. Users interact best with systems when they match the real world and when information follows a logical order. These interactions are further strengthened when systems are designed to provide a level of consistency and standards that support instinctual use and reduce the need for additional training and learning. A second enduring principal centers on *user awareness*, such that users have visibility and feedback on system status, are provided with visible information rather than relying on memory, and can easily access support and documentation for any system difficulties they encounter. Next, systems should be designed with *flexibility* in mind, such that users are able to assert control over their actions and process, such as undoing actions and modifying the interface to match the degree of activity that is required. Finally, because *errors are expected*, systems should be designed to antici-

34 Andy Pratt and Jason Nunes, *Interactive Design: An Introduction to the Theory and Application of User-Centered Design*, Rockport Publishers, 2012.

35 Susan Weinschenk, *100 Things Every Designer Needs to Know About People*, New Riders Pub, 2011.

36 Jakob Nielsen, "10 Usability Heuristics for User Interface Design," Nielson Norman Group, updated January 30, 2024.

pate, prevent, and alert users to errors. When errors are not preventable, systems should then support users in recognizing when errors have been made and provide guidance on how to best recover from those errors. For systems designed to specifically support operational situation awareness, usability heuristics become even more nuanced.[37] For example, systems need to be designed to avoid such issues as attention tunneling and data overload. They also need to effectively represent and communicate uncertainties, tame complexity, and support team operators.

Designers of military systems are being encouraged to actively adopt these principles in designs, ensuring that products and systems are more usable, reducing cognitive load, and allowing for rapid and accurate decisions. Such designs also enable faster onboarding and training and gain the trust of users more quickly. One example for which human-centered design principles have been adopted in military systems is the NATO Joint Military Symbology, which provides operators with uncluttered displays and a common standard of joint military symbols that together enhance consistency, efficiency, and decisiveness in military operations. In another example, the most important capability between a system and commanders and officers is information communication.[38] To ensure optimal communication of the large amount of information generated during an operation, a human-centered design effort took into account the system user's behavioral characteristics, operating habits, and psychological characteristics. Outcomes of this effort led to a human-machine system that is flexible and customizable to user needs, upholds normative standards and compatibility with other systems, offers adjustable personalized and customized interfaces, and provides visual coding methods that minimize visual elements. This, of course, must be balanced with the need for standardization across different military users to ease the challenge of compatibility and interoperability.

---

[37] Mica R. Endsley and Debra G. Jones, *Designing for Situation Awareness: An Approach to User-Centered Design*, 2nd ed., CRC Press, 2011.

[38] Gang Liu, and Yi Su, "Study on Big Data Visualization of Joint Operation Command and Control System," in Francis Y. L. Chin, C. L. Philip Chen, Latifur Khan, Kisung Lee, and Liang-Jie Zhang, eds., *Big Data 2018: Lecture Notes in Computer Science*, Springer, June 2018.

Human-Centered Design Applied to Artificial Intelligence

The interaction between humans and machines is trending toward the inclusion of AI applications, bringing potential benefits to both humans and machines. Humans and machines have an opportunity to improve each other, such that humans can serve as input to make AI and ML applications better, while ML applications can support tasks and improve human performance.[39]

One example is the use of AI agents as decision support for military command and control systems, especially in bottlenecks within the OODA (observe, orient, decide, act) loop.[40] In an AI-enhanced unit, an AI agent can support analysis activities, including classifying, finding, and fusing information; detecting anomalies; and developing and updating the common operating picture. From a planning perspective, an AI agent can support human team members by developing courses of action and red teaming. At an execution level, an AI agent can estimate the position and purpose of adversaries, predict tactical movements, and improve team efficiency by completing tasks.

Human-Machine Signaling

Effective communication is crucial for successful interactions between robots and humans. To date, effective human-robot communication remains an unsolved problem. For example, in San Francisco, driverless autonomous cars regularly interfere with fire department operations. Firefighters report that they have no way of communicating with autonomous vehicles when said vehicles impede emergency responders by blocking fire station driveways, stopping on top of fire department hoses, and parking themselves between fire trucks and burning cars, to name a few of the many documented incidents.[41]

---

[39] Robert Monarch, *Human-in-the-Loop Machine Learning,* Manning Publications, 2021.

[40] Johan Schubert, Joel Brynielsson, Mattias Nilsson, and Peter Svenmarck, "Artificial Intelligence for Decisions Support in Command and Control Systems," *23rd International Command and Control Research and Technology Symposium*, November 2018.

[41] Joe Eskenazi, and Will Jarrett, "Explore: See the 55 Reports—So Far—of Robot Cars Interfering with SF Fire Dept.," *Mission Local*, August 9, 2023.

Effective communication involves two directions of communication: human to robot and robot to human. Both directions need to be considered to create an effective team.[42]

When examining recent literature in the field of human to robot interaction, it is evident that recent advances in AI have created new ways for humans to communicate with robots. For example, advances in computer vision could allow robots to recognize hand and arm signals given by human teammates.[43] The proliferation of large language models could allow humans to give robots directions via natural language (voice) commands.[44] In our literature review, we found that human-robot communication paradigms are trending toward mimicking person-to-person communication methods that are used by human teams today. One reason for this is that language commands are intuitive for human users, transforming human-robot interactions "from rigid commands to natural conversations."[45] Language commands also allow for increased communication of context and reasoning in a way that more rigid commands might not.

When it comes to robot to human communication, it is important for robots to communicate actions and intentions to their human teammates. This can help achieve better teaming by reducing criticism of robot behaviors that appear unexplainable to humans but are actually useful. Literature shows that effective signaling is composed of three parts: action (I will do X), explanation (I will do X to achieve Y), and timing (I will do X before Y).[46] Researchers note that it is important for robots to not attempt to

---

[42] Ze Gong and Yu Zhang, "Behavior Explanation as Intention Signaling In Human-Robot Teaming," *27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2018.

[43] Munir Oudah, Ali Al-Naji, and Javaan Chahl, "Hand Gesture Recognition Based on Computer Vision: A Review of Techniques," *Journal of Imaging*, Vol. 6, No. 8, 2020.

[44] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang, "Drive As You Speak: Enabling Human-Like Interaction with Large Language Models in Autonomous Vehicles," *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.

[45] Cui et al., 2024, p. 2.

[46] Jennifer M. Riley, Laura D. Strater, Sheryl L. Chappell, Erik S. Connors, and Mica R. Endsley, "Situation Awareness in Human-Robot Interaction: Challenges and User

explain every action, because this could increase cognitive load for human teammates, but only explain actions that the robot deems would be incomprehensible to a human teammate.[47]

A final option might be to vary the level of autonomy when communication breaks down. Returning to the example of firefighters having difficulty communicating with autonomous vehicles that must be moved, one autonomous vehicle company created a first responder emergency access function to allow firefighters and others to access the vehicle and move it manually.[48]

Robots that can offer these types of task explanations to human teammates are less likely to be misunderstood. Large language models and natural language commands are once again a key technology enabler, allowing robots to communicate using plain English.

Shared Mental Models

Shared mental models are mental representations held by team members and are important for team cognition and overall success. Shared mental models evolve from team members who are working on a shared task and include information about their goals, environment, and strategies for accomplishing a task.[49] Although an individual's mental models might not initially match those of their team members, over time they converge through training, practice, and team interaction.[50] For teams that include an AI agent, both human and machine team members must develop mental models for how each other operates before a shared mental model can be established.

Interface Requirements," in Michael Barnes and Florian Jentsch, eds., *Human-Robot Interactions in Future Military Operations*, CRC Press, 2010.

[47] Gong and Zhang, 2018.

[48] Cruise, "Improving Emergency Vehicle and First Responder Interactions," blog post, October 12, 2023.

[49] Jessica Mesmer-Magnus, Ashley A. Niler, Gabriel Plummer, Lindsay E. Larson, and Leslie A. DeChurch, "The Cognitive Underpinnings of Effective Teamwork: A Continuation," *Career Development International*, Vol. 22, No. 5, August 2017.

[50] Robert W. Andrews, J. Mason Lilly, Divya Srivastava, and Karen M. Feigh, "The Role of Shared Mental Models in Human-AI Teams: A Theoretical Review," *Theoretical Issues in Ergonomics Science*, Vol. 24, No. 2, March 2023.

Developing a shared mental model within a human team is a highly social activity and requires team members to invest in one another and in their team's goals. Therefore, AI-enhanced units face challenges because they are constrained in the ways that they can engage. This challenge makes it more difficult for humans to develop trust in AI agents and for AI agents to accurately represent the evolving mental models of their human team members.[51] A mismatch in human and machine mental models can lead to suboptimal decisions, lower productivity, and decreased accuracy.[52] Therefore, AI agents can be developed so that their systems are easier for humans to build a mental model around. Guidance includes

- developing systems that have parsimonious error boundaries and are simple to represent
- minimizing the randomness of system errors to distinguish success from errors
- reducing task dimensionality with fewer features
- minimizing model updates that introduce new errors or instances.[53]

Following this guidance, humans will more easily understand the AI agent, develop and maintain trust in the AI agent, and work better with the AI agent to ultimately improve team performance.

Relatedly, this approach can be supplemented with neuro-symbolic AI approaches. In *neuro-symbolic AI*, symbolic reasoning is applied in concert with ML to help AI make deliberate decisions when pattern recognition is not sufficient.[54]

51 Andrews et al., 2023.

52 Renée J. Stout, Janis A. Cannon-Bowers, and Eduardo Salas, "The Role of Shared Mental Models in Developing Team Situational Awareness: Implications for Training," in Eduardo Salas and Aaron S. Dietz, eds., *Situational Awareness*, Routledge, 2017.

53 Stout et al., 2017.

54 Md Kamruzzaman Sarker, Lu Zhou, Aaron Eberhart, and Pascal Hitzler, "Neuro-Symbolic Artificial Intelligence: Current Trends," arXiv, arXiv: 2105.05330v2, May 14, 2021.

Transactive Memory Systems

Team performance depends on collective behavior through shared cognition, such as transactive memory systems. A *transactive memory system* describes team members' mental representations of the information and knowledge that is held by each team member within a group. The representations are a function of perceptual cognition (i.e., beliefs, attitudes, and expectations) and structural cognition (i.e., team knowledge patterns and the extent to which cognition is shared among team members).[55] Being aware of what other team members know and transferring information and ideas across team members is an important enabler for effective and efficient operations.[56]

Growing trends in the abundance and accessibility of information, and a growing need to ascertain the credibility of that information, present numerous opportunities and challenges to Army operations. More than ever, the ability to select and process the most accurate and most relevant information is essential for enabling decisionmaking in a complex operational environment. For example, the Commander's Critical Information Requirements (CCIRs) require personnel to sift through massive amounts of data to identify critical information that provides knowledge, understanding, and support across time horizons and for different courses of action.[57] Transactive memory systems can help teams appropriately filter this information while minimizing the potential for cognitive overload. Furthermore, a human or machine team member can enhance a team's transactive memory system by acting as an information specialist and expediting the coordination of activities across the vertical and lateral structure of teams and units.[58]

---

[55] Mesmer-Magnus et al., 2017.

[56] Verlin B. Hinsz, R. Scott Tindale, and David A. Vollrath, "The Emerging Conceptualization of Groups as Information Processors," *Psychological Bulletin*, Vol. 121, No. 1, January 1997.

[57] Deployable Training Division, Joint Force Development Directorate, "Commander's Critical Information Requirements (CCIRs)," Insights and Best Practices Focus Paper, 4th ed., Joint Staff, January 2020.

[58] Eoin Whelan and Robin Teigland, "Transactive Memory Systems as a Collective Filter for Mitigating Information Overload in Digitally Enabled Organizational Groups," *Information and Organization*, Vol. 23, No. 3, July 2013.

Incorporating AI agents as part of an AI-enhanced unit can help support Army transactive memory systems and enable more rapid adaptations to decisions that better reflect real-world conditions. There are a variety of roles in which an AI agent could serve to enhance collective intelligence, reduce cognitive overload, and expand the variety of options considered within a team.[59] First, AI agents can *assist* teams to augment human processes, which can include reminders to redirect attention or effort, for example. Second, AI agents can *coach* teams to collaborate better by guiding team members to those who might have required information and initiating or suggesting action in anticipation of possible issues. A team's collective effort could also be monitored through a *diagnostic* AI-agent, which could suggest approaches to tasks and identify team members who are best matched to different task elements. Finally, the *absence of social group dynamics* for AI agents means that they can also offer teams unique information that has not been influenced by behavioral factors. A resulting positive outcome is that teams are better able to generate novel ideas, and team members can more comfortably share AI-originating conflicting information without placing blame or initiating disagreement with others.[60]

---

59 Pranav Gupta and Anita Williams Woolley, "Articulating the Role of Artificial Intelligence in Collective Intelligence: A Transactive Systems Framework," *Proceedings of the Human Factors and Ergonomics Society International Annual Meeting*, Vol. 65, No. 1, 2021.

60 Nadine Bienefeld, Michaela Kolbe, Giovanni Camen, Dominic Huser, and Philipp Karl Buehler, "Human-AI Teaming: Leveraging Transactive Memory and Speaking Up for Enhanced Team Effectiveness," *Frontiers in Psychology,* Vol. 14, August 2023.

CHAPTER 4

Implications for the U.S. Army

> Treat [soldiers] . . . like professionals. Challenge them to meet the standard of the highest common denominator. Sometimes you'll be disappointed, but not that often because you'll be dealing with professional young [soldiers] who are earnestly seeking self-improvement. Treat them like untrustworthy adolescents who can only be counted on to produce the minimum requirement and you'll have an unmotivated mob that will require your direct and extensive supervision to accomplish the simplest tasks.[1]

We return now to the initial observation in the introduction of this report: The profession of arms is fundamentally a collective endeavor. As the excerpt from a seasoned combat leader notes, there is a distinctly didactic element to leadership in the profession of arms. It is through these lenses that we now reflect on the totality of the research we have conducted to identify findings and recommendations on human-machine integration in the U.S. Army.

Findings: Trust and Cognitive Burdens Are Key Difficulties to Human-Machine Integration

Returning to the objective of this research, we find that the key difficulties the Army is likely to encounter are related to cognitive burden and trust in

[1] U.S. Marine Corps Infantry Officer Course instructor, personal correspondence with the authors, August 7, 2005.

the machines with which soldiers will integrate. Our findings expand on this overarching theme by first exploring the inability of addressing these difficulties using XAI and cognitive forcing functions before turning to the deeper insight that building trust and overcoming cognitive burdens will likely require more time and more exposure to reach integration, just as human soldiers require between each other.

Efforts to Use Explainable AI and Cognitive Forcing Functions to Overcome Trust and Cognitive Burden Difficulties Are Less Effective Than Expected

The first and most interesting finding from the academic literature was that current efforts to increase human trust in AI leave something to be desired. Attempts to increase trust through greater transparency are offset by increased deference to AI outputs even in cases in which such deference is not warranted. Efforts to use cognitive forcing functions to counteract those effects reduce overtrust in AI but also result in an overcorrection in which humans trust algorithms *less* than they should.

Although these findings are in the context of experiments and use in medical and legal fields, similar outcomes are likely to be found in military contexts. This will be particularly true in environments where the human-machine dynamics are most similar to those in the civilian contexts, such as using AI to support staff work. There are more contextual differences when considering the issues of AI trust in tactical settings (e.g., executing a movement to contact or operating in urban terrain). However, we can imagine that either trust will be harder to gain or deference will be given; in either case, XAI as it is being developed is not likely to add value to the Army. This inability to increase human trust through XAI or cognitive forcing functions is especially critical because trust in AI is a key concern for the Army.

Overcoming Trust and Cognitive Burden Difficulties Might Require a More Deliberate Approach over a Longer Period

In a broad sense, XAI and cognitive forcing functions are forms of human-centered design *and* shared mental models at the same time. However, they are instantiations that attempt to build a complete shared mental model instantly. Moreover, the mental model is not shared between machine and human; it is only the machine either presenting its mental model to the human (in the case of XAI) or cleaving a part of its mental model off to give to the human to complete (in the case of a cognitive forcing function). Essentially, the human and the machine are not integrated but working in parallel.

Integrating humans and machines, then, will likely require a more deliberate approach of building shared mental models and transactive memory over time. This suggests that integration should be a much longer process in which the human parts of the team must share more of their mental model explicitly and make the implicit transactive memory systems that form in units more explicit to enable greater machine integration. Human-centered design and human-machine signaling can positively influence the interface between humans and machines, but such integration might only be helpful if it is supported by a deeper and more deliberate effort for the *human* to explain its actions so that the *machine* can learn and adapt to it. This interplay makes human-machine integration an exemplar of a sociotechnical system.[2]

Gaining Trust Will Take Time

The upshot of the developments noted in Chapter 3 is that the integration of humans and machines in the Army in the realm of ground combat planning and execution is likely to be slower than that in the commercial sector. This is largely because of the Army's outsize need for trust. This is observable in our narrative exploration and obvious when we consider the role of trust in soldiering as a profession; it is the foundation on which the Army's philosophy of mission command rests.[3]

Although the Army is not the only institution in which trust and integrity are needed at a high level, it is one of the few institutions that must work and act in opposition to a thinking adversary that is actively working

---

[2] Jon R. Lindsay, "War Is from Mars, AI Is from Venus: Rediscovering the Institutional Context of Military Automation," *Texas National Security Review,* Vol. 7, No. 1, Winter 2023.

[3] Army Doctrinal Publication 1, 2019, pp. 2–10.

to undermine and destroy it. This makes the Army rightfully conservative in its adoption of AI in integrated teams of humans and machines. This is especially the case for integration of humans and machines in tactical settings, where human recognition–primed decisionmaking is even more applicable and relevant to success.

Although some in the Army have advocated for faster adoption *because* peer threats are pursuing the same, we believe that the integration of humans and machines as envisioned in the narratives should be slower than in any analogous civilian setting, except for applications for which AI-enabled robots and machines perform the same function in civil and military settings. This is because the key difficulty we have uncovered centers in part on humans trusting machines. Fast adoption that results in bad results is likely to erode trust and make it difficult to rebuild. This has been observed for simple weapons such as the M-16 rifle, for which broken trust in its reliability took decades to repair; the dynamic is likely to be even more true for AI-enabled robots and machines.[4] This is also likely to be true even after machines—such as those noted in Table 3.1—are fielded on a large scale across Army formations because the process of trusting and adopting machines into a human team is a social one that cannot be fully designed for before fielding.

The Army Is More Focused on Fitting the Machine to the Human

Another finding is that efforts to integrate humans and machines in the Army are predominantly focused on making the machine fit the human. Efforts such as XAI, cognitive forcing functions, and human-centric design are all efforts to influence and guide the design of machines so that humans find them easier or more efficient to work with.

This is not a bad thing; there is far more space and opportunity to influence the design of machines than there is to change the design of humans. However, there are efforts to change the design of human *organizations* to work better with machines. From a tactical level, ideas of changing for-

4 See Bob Orkand and Lyman Duryea, *Misfire: The Tragic Failure of the M16 in Vietnam*, Stackpole Books, 2019.

mations to accommodate humans and machines (e.g., creating "machine minder" billets) are worth pursuing.[5] So are ideas that affect the Army at a more basic organizational level, such as modifying the RDT&E (research, development, test, and evaluation) and acquisition enterprise to make it more responsive to the frequent iteration and refinement AI requires.

Lastly, greater human-machine integration will require soldiers and commanders to be better trained and better educated about their machines. We asserted previously that human-machine integration is a sociotechnological system in which experience and familiarity with specific machines in specific contexts are necessary. Nevertheless, soldiers must have an underlying understanding of how AI works, which can only be accomplished through professional military education. This is especially necessary for commanders and small unit leaders, who must contextualize machine outputs and behaviors in the course of commanding an integrated human-machine unit. Relatedly, soldiers must gain mastery over specific machines. This is especially true for soldiers performing duties and filling roles related to managing AI systems within units. This will require progressively more-advanced training, much as the Army does for other weapon systems.[6] The contextual details will vary depending on which instantiations of human-machine integration are ultimately fielded across the Army, but it is clear that the Army must prepare to invest significant training and education time for its human soldiers to use AI-enabled machines and robots.

---

5 Todd South and Shawn Snow, "15-Marine Rifle Squad: An Exclusive Look Inside the Future Infantry," *Marine Corps Times*, August 9, 2019.

6 For comparison, the Army devotes significant amounts of time to train its infantry soldiers on their main weapon systems. An indirect fire infantryman (MOS 11C), for example, receives 22 weeks of basic training to learn basic employment of 60mm, 81mm, and 120mm mortars. Noncommissioned officers and officers receive a further five weeks of advanced training later in their careers to learn manual and digital fire direction processes, gun-laying, and advanced tactics. One would expect the Army to invest similar (or more) time to train soldiers and leaders on AI-enabled machines.

Recommendations

With these findings in mind, we conclude by offering three conceptual recommendations for the Army to consider as it continues to seek the integration of humans and AI-enabled machines and applications. All three recommendations focus on sociotechnical aspects of human-machine integration; this is a purposeful acknowledgement of our finding that most efforts focus on fitting machines to humans and not the other way around.

Integrate Machines Like New Soldiers

As the Army integrates machines in operational units, it should not treat that integration simply like it would for a weapon system. Rather, integration should feature elements that are more akin to integrating new soldiers into their units after basic training. That would entail integrating machines in less-complex or lower level mission-essential tasks with the unit before proceeding to more-complex ones. It would mean taking the time for humans to make implicit details of the way they operate explicit to the machine to help build shared mental models. This might include more detailed articulation of CCIRs, boundaries for decisionmaking, and standard operating procedures. It would also mean aligning system updates for the machine to improve its AI with the unit training cycle; here, preserving whatever transactive memory and shared mental model that exists between the humans and machines in the unit is more valuable than any incremental gains in AI accuracy from model updates that are implemented in the middle of a unit training cycle.

Certainly, there will be aspects of the machine that must be treated like a weapon system rather than a new soldier joining a unit. The physical elements of the autonomous system will need to be maintained, for example. However, the autonomous qualities of the machines the Army envisions using on the battlefield in 2040 will require the Army to adopt a mindset and strategic approach that resembles that of the Army's approach to individual and collective training.

Prepare for Variation

The Army should also be prepared to accept the likelihood that the characteristics, performance, and value of AI will vary across units. Our research in Chapters 2 and 3 reinforces the notion that human-machine integration is a sociotechnological construct. Therefore, variation should be expected because the *humans* vary. This is already the case in terms of the different standard operating procedures, performance, and culture of different units. This will have inevitable consequences on the way the AI-enhanced unit builds transactive memory and a shared mental model. The Army should be prepared for this.

Articulate What Trust Means in Human-Machine Integration and Monitor How to Gain and Maintain It

Although there are numerous technical and procedural challenges that must be addressed as the Army integrates machines into formations, most can conceivably be overcome with technical refinements to the machine themselves. However, one critical one is solely in the hands of humans: gaining and maintain trust between humans and machines. As we observed in Chapter 2 and in our aforementioned findings, the profession of arms demands an outsize degree of trust between its members; for machines, it will be gained more slowly than the pace of technical refinements to the machines themselves.

First, the Army must develop a clear articulation of what trust means in the context of human-machine integration. Our exploration of this topic leads us to believe that the trust between humans and an AI-enabled robot or application will take a different form compared with other forms of trust within the profession of arms. It will not look exactly like the trust between human soldiers, as the machine will never possess the moral agency of another human. Nor will it be the simple trust that soldiers place in their equipment, in that they trust it to perform as designed, because AI-enabled machines and applications will behave in ways that are not as transparent as a vehicle or weapon system. The trust between humans and AI-enabled machines and applications might be a hybrid of the two. In any case, monitoring that trust will require a clearer articulation of the trust that is sought between humans and machines.

Second, we recommend that the Army carefully monitor human-machine trust in formations and staffs. Just as DoD used monitoring frameworks to understand the progress of integrating women in combat arms roles and highlight especially challenging issues before they developed, it might consider a similar process to monitor the integration of humans and machines.[7]

To be clear, the issues of integrating women in combat arms units and integrating machines and humans in units generally are not substantively comparable. However, this should not stop the Army from adapting basic methods of monitoring implementation in the former to assist in the latter. Such a plan should include the identification of metrics and measures that indicate trust (i.e., increased utilization rate of machines), broad surveys and analysis of administrative data across operational units to understand macro trends, and the targeted use of focus groups and interviews to gain more nuanced understanding of trust issues in units with outlying trends.

Such an approach implies a relatively slow pace of integration between humans and machines. Although this might be in tension with tactical imperatives to field machines rapidly, we conclude that humans gaining trust in machines can only be built at a deliberate pace because humans take time to build trust in anyone (or anything). We also believe that such monitoring should be carried out by Training and Doctrine Command in partnership with Army Futures Command, Army Research Institute, the Judge Advocate General Corps, and others to emphasize that monitoring human-machine integration is as much a behavioral issue as a technical one.

Final Thought: The Human Must Remain the Dominant Partner in Integrated Units

The integration of humans and machines in the U.S. Army is likely to be one filled with promise, challenges, progress, and setbacks as DoD continues to develop AI-enabled applications and weapon systems. Many of these devel-

7 See Agnes Gereben Schaefer, Jennie W. Wenger, Jennifer Kavanagh, Jonathan P. Wong, Gillian S. Oak, Thomas E. Trail, and Todd Nichols, *Implications of Integrating Women into the Marine Corps Infantry*, RAND Corporation, RR-1103-USMC, 2015, pp. 127–138 and pp. 156–171.

opments will be derived from commercial advances. However, one constant should be that the role of machines should remain in support of humans in the profession of arms. One prominent AI researcher notes humans today play three roles in targeting: as a weapon system operator, as a fail-safe, and as a moral agent who makes "value-based judgements about whether the use of force is appropriate."[8] However advanced machines get, humans should not cede the moral agent role.

---

8 Paul Scharre, *Army of None: Autonomous Weapons and the Future of War,* W. W. Norton & Co., 2018, p. 233.

Abbreviations

AI	artificial intelligence
CCIR	Commander’s Critical Information Requirement
DARPA	Defense Advanced Research Projects Agency
DoD	U.S. Department of Defense
ML	machine learning
TOC	tactical operations center
UAS	uncrewed aerial system
XAI	explainable artificial intelligence

References

Andrews, Robert W., J. Mason Lilly, Divya Srivastava, and Karen M. Feigh, "The Role of Shared Mental Models in Human-AI Teams: A Theoretical Review," *Theoretical Issues in Ergonomics Science*, Vol. 24, No. 2, March 2023.

Army Doctrinal Publication 1, *The Army*, Headquarters, Department of the Army, July 2019.

Bansal, Gagan, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S. Weld, "Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance," *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, May 2021.

Beer, Jenay M., Arthur D. Fisk, and Wendy A. Rogers, "Toward a Framework for Levels of Robot Autonomy in Human-Robot Interaction," *Journal of Human-Robot Interaction*, Vol. 3, No. 2, 2014.

Bienefeld, Nadine, Michaela Kolbe, Giovanni Camen, Dominic Huser, and Philipp Karl Buehler, "Human-AI Teaming: Leveraging Transactive Memory and Speaking Up for Enhanced Team Effectiveness," *Frontiers in Psychology,* Vol. 14, August 2023.

Bilgic, Mustafa, and Raymond Mooney, "Explaining Recommendations: Satisfaction vs. Promotion," paper presented at the Tenth International Conference on Intelligent User Interfaces, January 2005.

Buçinca, Zana, Maja Barbara Malaya, and Krzysztof Z. Gajos, "To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making," *Proceedings of the ACM on Human-Computer Interaction*, Vol. 5, No. CSCW1, April 2021.

Chong, Leah, Guanglu Zhang, Kosa Goucher-Lambert, Kenneth Kotovsky, and Jonathan Cagan, "Human Confidence in Artificial Intelligence and in Themselves: The Evolution and Impact of Confidence on Adoption of AI Advice," *Computers in Human Behavior*, Vol. 127, February 2022.

Cruise, "Improving Emergency Vehicle and First Responder Interactions," blog post, October 12, 2023. As of January 9, 2025:
https://www.getcruise.com/news/blog/2023/how-were-improving-emergency-vehicle-and-first-responder-interactions/

Cui, Can, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang, "Drive As You Speak: Enabling Human-Like Interaction with Large Language Models in Autonomous Vehicles," *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.

Dellermann, Dominik, Philipp Ebel, Matthias Söllner, and Jan Marco Leimeister, "Hybrid Intelligence," *Business & Information Systems Engineering*, Vol. 61, No. 5, October 2019.

DePaulo, Bella M., James J. Lindsay, Brian E. Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper, "Cues to Deception," *Psychological Bulletin,* Vol. 129, No. 1, 2003.

Deployable Training Division, Joint Force Development Directorate, "Commander's Critical Information Requirements (CCIRs)," Insights and Best Practices Focus Paper, 4th. ed., Joint Staff, January 2020.

DoD Directive 3000.09, *Autonomy in Weapon Systems*, Office of the Under Secretary of Defense for Policy, January 25, 2023.

Endsley, Mica R., and Debra G. Jones, *Designing for Situation Awareness: An Approach to User-Centered Design*, 2nd. ed., CRC Press, 2011.

Eskenazi, Joe, and Will Jarrett, "Explore: See the 55 Reports—So Far—of Robot Cars Interfering with SF Fire Dept.," *Mission Local*, August 9, 2023.

Gavidia, Lisbeth A. Pino, and Joseph Adu, "Critical Narrative Inquiry: An Examination of a Methodological Approach," *International Journal of Qualitative Methods*, Vol. 21, April 2022.

Gong, Ze, and Yu Zhang, "Behavior Explanation as Intention Signaling In Human-Robot Teaming," *27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2018.

Gunning, David, and David W. Aha, "DARPA's Explainable Artificial Intelligence (XAI) Program," *AI Magazine*, Vol. 40, No. 2, 2019.

Gupta, Pranav, and Anita Williams Woolley, "Articulating the Role of Artificial Intelligence in Collective Intelligence: A Transactive Systems Framework," *Proceedings of the Human Factors and Ergonomics Society International Annual Meeting*, Vol. 65, No. 1, 2021.

Hinsz, Verlin B., R. Scott Tindale, and David A. Vollrath, "The Emerging Conceptualization of Groups as Information Processors," *Psychological Bulletin*, Vol. 121, No. 1, January 1997.

Horowitz, Michael C., and Lauren Kahn, "Bending the Automation Bias Curve: A Study of Human and AI-Based Decision Making in National Security Contexts," arXiv, arXiv:2306.16507, June 30, 2023.

Humphrey, Stephen E., Frederick P. Morgeson, and Michael J. Mannor, "Developing a Theory of the Strategic Core of Teams: A Role Composition Model of Team Performance," *Journal of Applied Psychology*, Vol. 94, No. 1, January 2009.

Johnson, Matthew, Jeffrey M. Bradshaw, Paul J. Feltovich, Catholijn Jonker, Birna van Riemsdijk, and Maarten Sierhuis, "Autonomy and Interdependence in Human-Agent-Robot Teams," *IEEE Intelligent Systems*, Vol. 27, No. 2, March 2012.

Kahneman, Daniel, *Thinking, Fast and Slow*, Farrar, Straus and Giroux, 2011.

Kessler, Sydney, Sam Wallace, Emily Yoder, and Jonathan P. Wong, *One Team, One Fight: Ground Combat Narratives of Human-Machine Integration,* RAND Corporation, RR-A2764-2, forthcoming.

Kim, Sangbae, "AI and Robots Are a Minefield of Cognitive Biases," *IEEE Spectrum*, July 19, 2021.

Klein, Gary A., "A Recognition Primed Decision (RPD) Model of Rapid Decision Making," in Gary A. Klein, Judith Orasanu, Roberta Calderwood, and Caroline E. Zsambok, eds., *Decision Making in Action: Models and Methods*, Ablex, 1993.

Kolodny, Lora, "Cruise Confirms Robotaxis Rely on Human Assistance Every Four to Five Miles," *CNBC,* November 6, 2023.

Lake, Brenden M., Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman, "Building Machines That Learn and Think Like People," *Behavioral and Brain Sciences*, Vol. 40, 2017.

Launchbury, John, "A DARPA Perspective on Artificial Intelligence," Defense Advanced Research Projects Agency, March 19, 2017.

Lindsay, Jon R., "War Is from Mars, AI Is from Venus: Rediscovering the Institutional Context of Military Automation," *Texas National Security Review,* Vol. 7, No. 1, Winter 2023.

Liu, Gang, and Yi Su, "Study on Big Data Visualization of Joint Operation Command and Control System," in Francis Y. L. Chin, C. L. Philip Chen, Latifur Khan, Kisung Lee, and Liang-Jie Zhang, eds., *Big Data 2018: Lecture Notes in Computer Science*, Springer, June 2018.

Luke, Timothy J., "Lessons from Pinocchio: Cues to Deception May Be Highly Exaggerated," *Perspectives on Psychological Science*, Vol. 14, No. 4, July 2019.

McDonough, James R., *Platoon Leader: A Memoir of Command in Combat*, Presidio Press, 1985.

Mesmer-Magnus, Jessica, Ashley A. Niler, Gabriel Plummer, Lindsay E. Larson, and Leslie A. DeChurch, "The Cognitive Underpinnings of Effective Teamwork: A Continuation," *Career Development International*, Vol. 22, No. 5, August 2017.

Metz, Cade, "What Makes A.I. Chatbots Go Wrong?" *New York Times,* March 29, 2023.

Miki, Takahiro, Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter, "Learning Robust Perceptive Locomotion for Quadrupedal Robots in the Wild," *Science Robotics*, Vol. 7, No. 62, January 2022.

Monarch, Robert, *Human-in-the-Loop Machine Learning,* Manning Publications, 2021.

Nielsen, Jakob, "10 Usability Heuristics for User Interface Design," Nielson Norman Group, updated January 30, 2024. As of January 6, 2025: https://www.nngroup.com/articles/ten-usability-heuristics/

Norman, Don, *The Design of Everyday Things*, rev. ed., Basic Books, 2013.

Orkand, Bob, and Lyman Duryea, *Misfire: The Tragic Failure of the M16 in Vietnam*, Stackpole Books, 2019.

Oudah, Munir, Ali Al-Naji, and Javaan Chahl, "Hand Gesture Recognition Based on Computer Vision: A Review of Techniques," *Journal of Imaging*, Vol. 6, No. 8, 2020.

Parasuraman, Raja, and Victor Riley, "Humans and Automation: Use, Misuse, Disuse, Abuse," *Human Factors*, Vol. 39, No. 2, June 1997.

Parasuraman, Raja, Thomas B. Sheridan, and Christopher D. Wickens, "A Model for Types and Levels of Human Interaction with Automation," *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, Vol. 30, No. 3, May 2000.

Pratt, Andy, and Jason Nunes, *Interactive Design: An Introduction to the Theory and Application of User-Centered Design*, Rockport Publishers, 2012.

Riley, Jennifer M., Laura D. Strater, Sheryl L. Chappell, Erik S. Connors, and Mica R. Endsley, "Situation Awareness in Human-Robot Interaction: Challenges and User Interface Requirements," in Michael Barnes and Florian Jentsch, eds., *Human-Robot Interactions in Future Military Operations*, CRC Press, 2010.

Sarker, Md Kamruzzaman, Lu Zhou, Aaron Eberhart, and Pascal Hitzler, "Neuro-Symbolic Artificial Intelligence: Current Trends," arXiv, arXiv:2105.05330v2, May 14, 2021.

Schaefer, Agnes Gereben, Jennie W. Wenger, Jennifer Kavanagh, Jonathan P. Wong, Gillian S. Oak, Thomas E. Trail, and Todd Nichols, *Implications of Integrating Women into the Marine Corps Infantry*, RAND Corporation, RR-1103-USMC, 2015. As of February 3, 2025: https://www.rand.org/pubs/research_reports/RR1103.html

Scharre, Paul, *Army of None: Autonomous Weapons and the Future of War,* W. W. Norton & Co., 2018.

Schubert, Johan, Joel Brynielsson, Mattias Nilsson, and Peter Svenmarck, "Artificial Intelligence for Decision Support in Command and Control Systems," *23rd International Command and Control Research and Technology Symposium*, November 2018.

Sledge, E. B., *With the Old Breed: At Peleliu and Okinawa,* Presidio Press, 2007.

South, Todd, and Shawn Snow, "15-Marine Rifle Squad: An Exclusive Look Inside the Future Infantry," *Marine Corps Times*, August 9, 2019.

Stout, Renée J., Janis A. Cannon-Bowers, and Eduardo Salas, "The Role of Shared Mental Models in Developing Team Situational Awareness: Implications for Training," in Eduardo Salas and Aaron S. Dietz, eds., *Situational Awareness*, Routledge, 2011.

Tangalakis-Lippert, Katherine, "Marines Fooled a DARPA Robot by Hiding in a Cardboard Box While Giggling and Pretending to Be Trees," *Business Insider,* January 29, 2023.

Vassilev, Apostol, Alina Oprea, Alie Fordyce, and Hyrum Anderson, *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*, National Institute of Standards and Technology, NIST AI 100-2e2023, January 2024.

Weinschenk, Susan, *100 Things Every Designer Needs to Know About People*, New Riders Pub, 2011.

Whelan, Eoin, and Robin Teigland, "Transactive Memory Systems as a Collective Filter for Mitigating Information Overload in Digitally Enabled Organizational Groups," *Information and Organization*, Vol. 23, No. 3, July 2013.

Wong, Yuna Huh, John Yurchak, Robert W. Button, Aaron B. Frank, Burgess Laird, Osonde A. Osoba, Randall Steeb, Benjamin N. Harris, and Sebastian Joon Bae, *Deterrence in the Age of Thinking Machines*, RAND Corporation, RR-2797-RC, 2020. As of February 3, 2025:
https://www.rand.org/pubs/research_reports/RR2797.html

Yang, Guang-Zhong, Jim Bellingham, Pierre E. Dupont, Peer Fischer, Luciano Floridi, Robert Full, Neil Jacobstein, Vijay Kumar, Marcia McNutt, Robert Merrifield, et al., "The Grand Challenges of *Science Robotics*," *Science Robotics*, Vol. 3, No. 14, January 2018.

Yeh, Hung-Wen, Byron J. Gajewski, David G. Perdue, Angel Cully, Lance Cully, K. Allen Greiner, Won S. Choi, and Christine Makosy Daley, "Sorting It Out: Pile Sorting as a Mixed Methodology for Exploring Barriers to Cancer Screening," *Quality & Quantity*, Vol. 48, No. 5, September 2014.

RAND ARROYO CENTER

Advances in artificial intelligence (AI), machine learning, and robotics have raised the possibility that the profession of arms will soon include integrating human soldiers with AI-enabled machines and applications as part of the collective whole. Machines and software applications enabled by AI are starting to demonstrate capabilities that are relevant to military settings, such as moving autonomously through complex urban traffic and creating startlingly humanlike and interesting derivative works through large language models.

However, this does not mean that such developments can be implemented in military settings smoothly. The practice of building cohesive small units is no easy endeavor. The best small units cohere to the point where one soldier recognizes the silhouette and gait of another in the dark of a patrol base in an instant. The best staffs internalize their commander's style and specific needs over time. Integrating humans and machines in military contexts will likely draw from civilian parallels but will also require substantial contextualization.

In this report, the authors investigate the kinds of difficulties that the Army might encounter as it attempts to pair humans with AI algorithms to accomplish specific warfighting tasks. They make recommendations to address how these potential obstacles can best be overcome and ensure that the Army effectively creates AI systems that will integrate well with the soldiers who must interact with them.

$26.00

ISBN-10 1-9774-1515-6
ISBN-13 978-1-9774-1515-8

52600

9 781977 415158

www.rand.org

RR-A2764-1