FEBRUARY 2021

Trusted Partners

Human-Machine Teaming and the Future of Military AI

CSET Issue Brief

AUTHORS
Margarita Konaev
Tina Huang
Husanjot Chahal

Executive Summary

The Department of Defense wants to harness AI-enabled tools and systems to support and protect U.S. servicemembers, defend U.S. allies, and improve the affordability, effectiveness, and speed of U.S. military operations.[1] Ultimately, all AI systems that are being developed to complement and augment human intelligence and capabilities will have an element of human-AI interaction.[2] The U.S. military's vision for human-machine teaming, however, entails using intelligent machines not only as tools that facilitate human action but as *trusted partners* to human operators.

By pairing humans with machines, the U.S. military aims to both mitigate the risks from unchecked machine autonomy and capitalize on inherent human strengths such as contextualized judgement and creative problem solving.[3] There are, however, open questions about human trust and intelligent technologies in high-risk settings: What drives trust in human-machine teams? What are the risks from breakdowns in trust between humans and machines or alternatively, from uncritical and excessive trust? And how should AI systems be designed to ensure that humans can rely on them, especially in safety-critical situations?

This issue brief summarizes different perspectives on the role of trust in human-machine teams, analyzes efforts and challenges to building trustworthy AI systems, and assesses trends and gaps in relevant U.S. military research. Trust is a complex and multi-dimensional concept, but in essence, it refers to the human's confidence in the reliability of the system's conclusions and its ability to accomplish defined tasks and goals. Research on trust in technology cuts across many fields and academic disciplines. But for the defense research community, understanding the nature and effects of trust in human-machine teams is necessary for ensuring that the autonomous and AI-enabled systems the U.S. military develops are used in a safe, secure, effective, and ethical way.

While the outstanding questions regarding trust apply to a broad set of AI technologies, we pay particularly close attention to machine learning systems, which are capable not only of detecting patterns but also learning and making predictions from data

without being explicitly programmed to do so.[4] Over the past two decades, advances in ML have vastly expanded the realm of what is possible in human-machine teaming. But the increasing complexity and unique vulnerabilities of ML systems, as well as their ability to learn and adapt to changing environments, also raise new concerns about ensuring appropriate trust in human-machine teams.

With that, our key takeaways are:

- Human trust in technology is an attitude shaped by a confluence of rational and emotional factors, demographic attributes and personality traits, past experiences, and the situation at hand. Different organizational, political, and social systems and cultures also impact how people interact with technology, including their trust and reliance on intelligent systems.
    - That said, trust is a complex, multidimensional concept that can be abstract, subjective, and difficult to measure.
    - Much of the research on human-machine trust examines human interactions with automated systems or more traditional expert systems; there is notably less work on trust in autonomous systems and/or AI.
- Defense research has focused less on studying trust in human-machine teams directly and more on technological solutions that "build trust into the system" by enhancing system functions and features like transparency, explainability, auditability, reliability, robustness, and responsiveness.
    - Such technological advances are necessary, but not sufficient, for the development and proper calibration of trust in human-machine teams.
    - Systems engineering solutions should be complemented by research on human attitudes toward technology, accounting for the differences in people's perceptions and experiences, as well as the dynamic and changing environments where human-machine teams may be employed.

- To advance the U.S. military vision of using intelligent machines as trusted partners to human operators, future research directions should continue and expand on:
    - Research and experimentation under operational conditions,
    - Collaborative research with allied countries,
    - Research on trust and various aspects of transparency,
    - Research on the intersection of explainability and reliability,
    - Research on trust and cognitive workloads,
    - Research on trust and uncertainty, and
    - Research on trust, reliability, and robustness.

Human-machine teaming is, most basically, a relationship. And like with any other relationship, understanding human-machine teaming requires us to pay attention to three sets of factors—those focused on the human, the machine, and the interactions—all of which are inherently intertwined, affecting each other and shaping trust. For the defense research community, insights from research on human attitudes toward technology and the interactions and interdependencies between humans and technology can strengthen and refine systems engineering approaches to building trustworthy AI systems. Ultimately, human-machine teaming is key to realizing the full promise of AI for strengthening U.S. military capabilities and furthering America's strategic objectives. But the key to effective human-machine teaming is a comprehensive and holistic understanding of trust.

Table of Contents

Introduction

The U.S. military has a long history of developing and deploying AI systems that have the ability to perform tasks that generally require human intelligence, including aircraft autopilots, missile guidance technology, and highly-automated missile defense systems.[5] Humans, of course, have maintained a level of supervisory control—setting and monitoring tasks and goals, making safety critical decisions, and authorizing the use of lethal force. Over the past two decades, significant technological breakthroughs in the field of AI and most notably, advances in machine learning techniques, have expanded and diversified the ways in which humans can interact and collaborate with unmanned systems, robots, virtual assistants, algorithms, and other non-human intelligent agents. The Department of Defense, in turn, sees great potential in leveraging AI to redefine what is possible in the realm of human-machine teaming.

The U.S. Army, for instance, is interested in autonomous vehicle technology to reduce the number of service members needed to run resupply convoys in combat environments.[6] While the technology for fully autonomous vehicles does not yet exist, RAND researchers estimate that even a partially unmanned convoy—where the lead truck with soldiers is followed by unmanned vehicles in a convoy—would put 37 percent fewer soldiers at risk compared to current practices.[7]

The Air Force's Skyborg program, meanwhile, envisions autonomous, low-cost drones with a suite of AI capabilities as partners for fighter jets. Here, the focus on human-machine teaming helps solve one of the key challenges in aerial combat: the fact that sensors and shooters are collocated on a single platform with a human operator in it. In the future, teaming up manned fighter jets with AI-enabled autonomous drones could allow the Air Force to put sensors ahead of shooters, put unmanned systems ahead of human-operated fighter jets, take greater risks or tolerate the loss of some systems to protect others.[8]

That said, beyond certain information processing functions, current AI technologies (and more specifically, ML-based systems) are

largely not ready for operational deployment, in part due to their brittleness. These systems perform well in stable training and test environments but cannot yet reliably handle uncertain and novel situations. For instance, investigations into the 2018 incident in which one of Uber's self-driving cars killed a woman in Arizona revealed that while the automated driving system was able to recognize pedestrians with a high degree of accuracy in simulations, it wasn't very good at detecting, classifying or responding to other objects on the road or to pedestrians behaving unexpectedly, such as jaywalking or walking alongside their bike.[9]

ML-based systems are also vulnerable to adversarial manipulation and attacks that can pollute the training data or trick the machine, causing it to malfunction or otherwise fail in unpredictable ways. One popular example of adversarial manipulation involves an image of a turtle that an algorithm was fooled into believing was an image of a gun through pixel changes not visible to the human eye.[10] These challenges and risks are even greater in a military context where the environment is inherently adversarial, uncertain, and lethal.

While today's intelligent systems are still largely tools and not true teammates, human-machine teaming technology is progressing. The Department of Defense is looking to build machines that can adapt to the environment and the different states of their human teammates, anticipate the human teammates' capabilities and intentions, and generalize from learned experiences to operate in new situations.[11] But for the U.S. military to fully capitalize on the advantages in speed, precision, coordination, reach, persistence, lethality, and endurance promised by such advances, soldiers will need to trust these intelligent machines.

In the context of human-machine teaming, trust speaks to the human's confidence in the reliability of the system's conclusions and its ability to accomplish defined tasks and goals. Trust affects how people feel about and interact with technology, informing whether they choose to use, collaborate with, and rely on intelligent systems, and accept and follow the technology's recommendations. National security leaders, military professionals,

and academics therefore tend to agree that trust is essential for effective human-machine teaming.

Despite this apparent consensus, CSET research has found that few U.S. military research programs related to autonomy or AI focus directly on studying trust in human-machine teams.[12] To an extent, this gap reflects the broader state of the field, where research, thus far, has been more focused on trust in automation and less on trust in advanced autonomy and AI.[13] Moreover, considering that trust is an abstract concept that is difficult to measure directly, the defense research community seems to prioritize technology-centric approaches that seek to "build trust into the system." Alongside assurance, such efforts entail developing and enhancing system features and capabilities closely related to trust, including transparency, explainability, auditability, reliability, robustness, and responsiveness.

Technological advances in AI and robotics that extend the capabilities of machines, including the aforementioned trust-related system features, are of course necessary for progress toward advanced human-machine teaming. But without a better understanding of what it takes for military personnel to develop the kind of trust in their AI partners that they currently place in their fellow soldiers, sailors, airmen, and Marines, technology-centric solutions of this nature may not be sufficient. Rather than advocating for one approach or another, we simply suggest that insights from cognitive science, neuroscience, psychology, communications, and social sciences on human attitudes toward technology and the interactions and interdependencies between humans and intelligent machines can augment and refine systems engineering approaches to building trustworthy AI systems.

This issue brief reviews research on the drivers and effects of trust in human-machine teams, assesses the risks posed by deficits in trust and uncritical trust, examines efforts to build trustworthy AI systems, and offers future directions for research on trust in human-machine teams relevant to the U.S. military. We focus on trust not as an end in itself. Rather, our goal is to help the defense and national security community develop a more holistic understanding of trust in human-machine teams to ensure that

DoD is able to implement its vision of using AI systems as trusted partners to human operators in a safe, secure, effective and ethical way.

Trust in Human-Machine Teams

Research on human trust in technology encompasses many fields, including engineering, computer science, cognitive sciences, organizational behavior, and philosophy, each with different ways to define and measure this complex and multidimensional concept. For the purposes of this report, trust in the context of human-machine teaming refers to the human's confidence in the reliability of the system's conclusions, and its ability to perform specified tasks and accomplish defined goals.[14]

By emphasizing both a system's conclusions and its ability to perform tasks, the above definition of trust applies to human interactions with different types of intelligent technologies—robots capable of taking action in the physical world, virtual agents or bots (i.e. a virtual assistant with a visual presence or a distinguished identity), or embedded AI that is invisible to the user (i.e. an algorithmic decision-support software).[15] This distinction is important considering that the U.S. military's vision for human-machine teaming includes all of these different interactions. Moreover, there is evidence that the trajectory of human trust, as well as the factors that influence it, vary depending on the type of technology representation—namely, robotic, virtual, or embedded.[16]

Trust affects the willingness of humans to use, collaborate with, and rely on intelligent technologies and accept their outcomes or recommendations. Trust is particularly relevant to human-machine interactions in military settings because of both the promise and the perils of autonomous and AI-enabled technology. Pairing humans with intelligent machines can help reduce the risk to U.S. service personnel, lighten the warfighters' cognitive and physical load to improve performance and endurance, and increase accuracy and speed in decision-making and operations. Yet current AI systems (and more specifically, ML-based systems) are largely unprepared for operational deployment; they are vulnerable to adversarial manipulation and attack, and cannot reliably handle uncertain and new situations. Their misuse, malfunction or failure can cause unacceptable levels of damage.

Ultimately, as DoD's AI ethics principles dictate, humans are responsible for the development, use, and outcomes of AI systems in both combat and non-combat situations.[17] Thus, as the U.S. military moves to employ intelligent agents and systems as trusted partners to human operators, one of the most important questions it faces is how to ensure appropriate trust, contingent on machine capabilities and the context of the task at hand.[18] This level of correspondence between the user's trust and the technology's capabilities, known as calibration, can influence the actual outcomes of technology use and the overall effectiveness of human-machine teaming.[19] Too little trust in highly capable technology can lead to underutilization or disuse of autonomous systems, as well as lost time and efficiency; too much trust in limited or untested technology can lead to overreliance or abuse of autonomous systems. As we discuss later in the report, both pose significant risks and could undermine the effective use of human-machine teams in military settings.

Researchers measure trust in different ways. Some studies use psychophysiological measurements of trust. Examples include the use of electroencephalography (EEG) to capture the cortical activity of the brain and track changes in levels of anxiety, excitability, and vigilance, or facial expression analysis to classify negative and positive emotions and approximate trust in automated vehicles, for instance.[20] Others rely on behavioral measures of trust, such as a user's willingness to take the system's advice and act on it or comply with requests. Surveys requiring participants to report their level of trust is another common measurement method that assesses people's attitudes or sentiments toward technology using different scales.[21]

Because trust can be an abstract, subjective, and relative concept, it is difficult to measure directly. Therefore, trust measurement often involves indirect assessments—measuring behaviors and actions influenced by trust or factors that influence trust. Overall, despite the progress in developing different scales and behavioral measures for trust, a recent review of empirical research on human trust in AI has found that "there is an urgent need for addressing variance in measures used to assess human trust in AI."[22]

Systems engineering plays a pivotal role in engendering trust in human-machine teams. The underlying logic is that for humans to trust and use automated, autonomous and AI-enabled systems, trust can and should be "deeply embedded in the fabric of the system."[23] At every step of the technology lifecycle, developers (through close consultation with end-users) need to identify, specify, and integrate into the system the appropriate attributes, capabilities, and features that instill confidence and allow for proper trust calibration. For example, a soldier driving one of the trucks in a partially unmanned convoy needs to know what action the system will take if it encounters an obstacle. To support trust, this type of information extraction would then need to be built into the system as it is being developed—both the capability to extract a key single piece of information via what-if type queries and the capability to explain it in the operator's language. In other words, this technological approach to engendering trust in human-machine teams posits that such trust-cultivating capabilities can and should be specified in the original requirements, implemented in the design, and then certified through testing, evaluation, validation, and verification. Indeed, it is hard to imagine achieving the trust of operators without such system engineering.

That said, as a 2017 Center for Naval Analyses report on AI-based technologies and DoD explains, "trust is not an innate trait of the system."[24] Rather, trust is best thought of as a "relative measure of how a human operator (or operators)—whose own performance depends, in part, on collaborating in some way with the system—experiences…and perceives the behavioral pattern of a system."[25] An inquiry into the nature and implications of trust in human-machine teams then first requires us to better understand what drives trust. As such, we must assess and synthesize insights not only from research on building trustworthy AI systems and research on the interactions and interdependencies between humans and technology, but also research on human attitudes toward technology. Human-machine teaming is, in essence, a relationship. While discussed separately throughout the next sections of the report, these three sets of factors—whether focused on the human, the machine, or the interaction—are inherently intertwined, affecting each other and shaping trust.

Understanding Human Attitudes Toward Technology

Human trust as it pertains to technology, and more specifically, automated, autonomous and/or AI-enabled systems, can be organized in three categories: dispositional, situational, and learned.[26] Dispositional trust refers to a human's inherent tendency to trust automation, which varies based on a multitude of factors such as a person's age, personality, or culture.[27] For example, a global survey of 18,000 adults aged 16–64 found that younger generations trust AI more than older generations.[28] Such discrepancies in dispositional trust could have significant implications for the future of human-machine teaming. If the effect is generational, as Millennials and Generation Z come to represent the majority of those serving in the U.S. military, we may see fewer barriers to trust in human-machine teams as AI-enabled systems are deployed and fielded. But if the effect is related to age rather than generation, there may be important discrepancies in how younger servicemembers relate to AI-enabled technologies compared to how those who are older (and are therefore more likely to be in higher positions of command) view AI.

This survey also revealed that dispositional trust in AI varies by country of origin: 70 percent of respondents in China, for instance, said they trust AI compared to the 25 percent of those in the United States. Previous research on negative attitudes towards robots also reveals that cultural background plays an important role; yet in this study, U.S. participants exhibited the most positive perceptions.[29] Cultural influences on trust in AI are relevant when thinking about how quickly and effectively U.S. competitors and adversaries could integrate AI into their military systems. Cross-national variation in trust in AI technologies could also affect coordination in multinational coalitions like NATO. If commanders from some allied countries are more reluctant to trust and use AI-enabled systems during multinational operations, such divergence could undermine coordination, interoperability, and overall effectiveness.[30]

Situational trust refers to human attitudes towards technology and automation as influenced by different environmental factors, a person's mental state, or the nature of the task. In high stress and

emergency situations, as well as when multi-tasking, research has found that people tend to overtrust recommendations made by machines even when other indicators suggest the system's conclusions are wrong.[31]

Lastly, learned trust is based on a person's past experience with automation. Several studies show that even skilled pilots and air traffic controllers who have experience with highly reliable automated technologies exhibit automation complacency, meaning that they are worse at detecting system malfunctions under automation control compared with manual control.[32] Training is another form of experience that speaks to users' learned trust, which affects how individuals relate to technology. Yet evidence suggests that training can lead to both better performance due to lower complacency levels as users become more familiar with the baseline reliability of the system and over-reliance on automation due to familiarity and desensitization effects.[33]

Taken together, people's trust in technology is shaped by a myriad of dispositional, situational, and learned factors and experiences. Notably, these factors do not operate in isolation, but overlap and interact with one another. The 2003 friendly-fire incident involving the U.S. Patriot system—a highly automated missile defense system tasked with shooting down enemy missiles—is an instructive example.

On April 2, 2003, after completing a mission over Baghdad, two U.S. Navy F/A-18 aircraft approached the area in central Iraq where Patriot batteries were positioned. The Patriot system misclassified the lead aircraft as a ballistic missile, issuing an (false) alert of an attack. The tactical director at the battalion command and control then ordered the subordinate battery fire units to "bring your launchers to ready."[34] With the system in automatic engagement mode, turning the launchers to ready resulted in an automatic engagement a few seconds later—killing the pilot of the F/A-18 and destroying the aircraft.

Subsequent investigations partially attributed this friendly-fire incident (alongside the preceding fratricide involving a British Tornado aircraft) to operators' "unwarranted and uncritical trust in

automation."[35] When considering the dispositional factors affecting trust, it is relevant that Patriot operators are relatively junior in both age and rank. On an organizational level, the U.S. Army air defense culture at the time "encouraged a posture of over-trust in technology."[36] Training and exercises, which cultivate learned trust, reinforced this culture of over-reliance on technology and automation.[37] Specifically, Patriot operators were not sufficiently trained in scenarios emphasizing careful discrimination between hostile aircraft and missiles and friendly aircraft. In terms of situational factors, the high-stakes mission of ballistic missile defense against the explicit Iraqi threat of chemical attacks on advancing U.S. troops upped the stress and pressure that tends to lead to uncritical trust in technology and automation. Dispositional, situational, and learned factors therefore converged to cultivate excessive trust in the highly automated Patriot system, with tragic results.

Much of this issue brief is focused on the relationship between individual operators and intelligent technology. Yet as the Patriot fratricide incident illustrates, different organizational, political, and social systems and cultures impact individuals' attitudes, decisions, and behavior, including their trust and reliance on technology. As we turn to the discussion of trust calibration, it is important to recognize these broader structures are always at play.

Calibrating Trust: Trust Gap and Automation Bias

As with most of life's questions, the answer to how much trust is needed for effective human-machine teams is, "it depends." Proper calibration of trust means that the amount or level of trust humans place in machines is appropriate given the machine's capabilities at that particular time and context. Having too little trust is a poor calibration which results in what researchers have called a "trust gap;" having too much trust is often referred to as "automation bias." Both present unique risks and obstacles for the application of human-machine teams in military settings.

A trust gap can develop due to dispositional factors, such as age or culture. Situational factors such as the task at hand can also play a role. For instance, research shows that humans are averse to

machines making morally relevant decisions when it comes to driving, legal matters, medical situations, and military operations.[38] But even when not dealing with life and death decisions, there is evidence that while algorithms generally outperform humans in forecasting and prediction tasks, people nonetheless trust and prefer human forecasts.[39]

Part of the challenge of ensuring good trust calibration stems from a misalignment between human expectations and machine capabilities. Research shows that users tend to approach intelligent technologies, particularly virtual AI agents or bots and embedded AI such as an algorithmic decision-support software that is invisible to the user, with high expectations of their performance and high levels of initial trust.[40] But when an error occurs, the contrast between what the system can do and what the human operator expects it can do can cause the human to overcorrect their expectations and assess the reliability of the system lower than warranted.[41]

People seem quick to lose trust when the technology makes mistakes, especially early on in the interaction or mission, which speaks to learned trust, or more accurately, learned mistrust.[42] Breakdowns in trust, as some researchers suggest, could be repaired by providing situation-specific training to operators (i.e. enhancing learned trust) or by increasing the transparency of the system.[43] Others argue that a system must offer enough value that humans feel as though it is worth forgiving when it fails.[44] Forgiveness may be contingent on dispositional factors and human judgment. But professional organizations such as the military have risk assessment and safety protocols that ultimately determine if a faulty system can be used again after malfunctioning. This brings attention to how the process of calibrating trust in human-machine teams is contingent not only on dispositional, situational, and learned factors at the level of the individual, but also on institutional and organizational procedures.[45]

Regardless of the approach one takes to repairing trust, bridging the 'trust gap' may be a necessary prerequisite for deploying some of the AI technologies the U.S. military is currently researching. An instructive example is the Defense Advanced Research Projects

Agency's (DARPA) experimentation program, Squad X, which partners infantry squads with AI and autonomous systems. In its most recent experiment, autonomous ground and aerial systems were used for sensing and surveillance to provide reconnaissance and improve situational awareness for infantry units moving through natural desert and mock city blocks. AI is used to synthesize the information accumulated through a network of warfighter and unmanned nodes, cutting through the noise to provide the squad members with actionable intelligence directly to their handheld devices.[46] With advances in real-time analytics and recommender systems technologies, such computational support could help warfighters gain the initiative in dynamic operational settings.[47] But if human operators do not trust the system, they would be reluctant to follow its recommendations.[48] Thus, without trust in human-machine teams, the U.S. military may not be able to capitalize on the advantages in speed, coordination, and precision AI promises to deliver.

While distrust is a form of poor calibration where human trust falls short of the technology's capabilities, over-trust or uncritical trust is another form of inappropriate reliance on technology, often described as 'automation bias.' Automation bias typically manifests itself in two types of errors: errors of omission, when people do not notice problems because the machine did not alert them, and errors of commission, where people follow automated commands or suggestions that are incorrect or inappropriate. Much of the research on automation bias comes from studies in aviation, including the analyses of incident reports citing overreliance on automated flight management systems as well as simulations and experiments.[49]

For instance, one study tested for both types of errors on a group of 25 pilots in a simulated flight experiment.[50] To test for commission errors, pilots were presented with an automated alert warning that the engine was on fire, though other engine parameters were normal and no other indicators of trouble appeared. The test for omission errors included misloading data such as altitude clearance and frequency change. While the pilots missed incorrect information and other automation failures 55 percent of the time (omission error rate), all of the participants shut

down the engine in response to the fake fire alert, indicating a 100 percent commission error rate.[51]

Interestingly, the study also found that more experienced pilots were less likely to detect automation failures, pointing to how learned trust can exhibit itself through complacency and dependence on technology that builds up over time, especially when the system has proven itself reliable. Overall, the results, supported by other studies in aviation and health care, demonstrate that when automated decision aids are available, people tend to follow their cues.[52]

To an extent, both errors of commission and errors of omission stem from the fact that humans tend to be "cognitive misers," meaning they choose the option requiring the least cognitive effort and are not likely to seek alternative options.[53] Evidence from other studies in human factors literature show that as it becomes more difficult for human operators to disaggregate the factors that influenced the machine's decision, they become more likely to accept these solutions without question.[54]

Notably, much of this research examines automated systems or more traditional expert systems that perform scripted tasks based on specified rules. These systems are less advanced than today's machine learning systems, especially those using deep neural network approaches that reason, reach conclusions, provide recommendations, and take action in ways not evident or easily explained to humans.[55] It is difficult to predict how such technological advances could affect trust in human-machine teams. The increased sophistication of intelligent technologies could amplify the inclination to over-trust complex systems. On the other hand, people may be reluctant to trust systems they do not understand. Such uncertainty only highlights the significant role broader social, organizational, and institutional structures and practices have in helping individual operators to properly calibrate trust depending on the capabilities and limitations of the system and the task at hand—serving as a bulwark against the risks that stem from both uncritical trust in technology and deficits in trust.[56]

Thus far, the discussion has focused on research studying human attitudes toward technology in order to better understand the drivers and effects of trust in human-machine teams. Yet a holistic understanding of trust in human-machine teaming accounts for all three—the human, the machine, and the team—each as its own unit of analysis. The following section therefore centers on AI system features, as well as the interactions and interdependencies between humans and intelligent technologies that facilitate the development of trust.

From Intelligent Tools to Trusted Partners

For humans to properly calibrate trust—that is, to accurately gauge the extent to which an intelligent system can be relied upon given its capabilities, limitations, and the context at hand—they need to understand what a system can and cannot do in a given mission or environment, and why it makes the decisions it does. The transparency of the system, the capacity of the system to explain its decisions, the quality of communications between human and machine, and the reliability of the system in the present and future are all critical factors for calibrating trust and enabling effective human-machine teaming. Research and innovation has therefore focused on ways to 'build in' trust into autonomous and AI-enabled systems through features and functions that make these systems more transparent, explainable, auditable, reliable, robust, and responsive.[57]

While the discussion below focuses predominantly on operators or end-users, technology-centric solutions to enabling trust in intelligent systems increasingly involve collaborative design between scientists, technicians, soldiers, and commanders, as well as efforts to test AI systems earlier in the development process to gather feedback. The capabilities and limitations of AI systems, however, change depending on the context at hand. Moreover, advanced AI systems are designed to continuously learn and alter themselves, even after they have been in operation. The trustworthiness of AI systems should therefore not be treated as static or permanent.[58] Indeed, this is partly why there are substantial challenges to testing, evaluation, validation, and verification of AI systems. While a comprehensive account of these issues is beyond the scope of this report, suffice it to say that senior defense leaders both outside and within the Pentagon recognize these challenges and are beginning to take steps toward reforming processes and practices that can accommodate a collaborative, holistic, and continuously evolving approach to building and deploying trustworthy AI systems.[59]

Transparency, Explainability, and Auditability

Transparency is a critical aspect of trustworthy technologies and is essential for calibrating operator trust in a machine. But defining the type and level of information the intelligent system needs to convey to the human operator, as well as how to communicate said information so that it is understandable to humans, remain areas of open inquiry.

When approaching the question of what information is important for transparency and trust in human-machine teams, researchers have looked into factors such as the intelligent agent's current actions and plans, reasoning process, projected outcomes, and uncertainty. The Department of Defense's Autonomy Research Pilot Initiative (ARPI), for example, conducted a series of studies exploring human interactions with an autonomous squad member—a robotic mule that accompanies a dismounted soldier squad within a simulated military environment. To determine how different configurations of information influence human perceptions of the autonomous squad member, the robot shared information about its current goal (e.g., to return to base), current priority (e.g., to save time), and its projected resource expenditure (e.g., how much extra fuel it needed to use to meet its goal given said priority). The study showed that participants' situational awareness and understanding of the robot peaked when the robot displayed information about its intent, logic, and possible outcome while the addition of uncertainty information did not further enhance trust.[60]

The finding regarding uncertainty information is notable, considering other research that shows humans perceive agents as more trustworthy when they convey uncertainty estimates and that doing so can also improve joint human-machine team performance.[61] In other words, while information about uncertainty can be beneficial, it may also cause confusion and prove less useful depending on the mission environment or individual differences.[62] More research is therefore needed to better understand how uncertainty affects trust (as well as performance) in human-machine teams.

Transparency about failures and errors is particularly important for trust in human-machine teams. As previously mentioned, humans are quick to lose confidence when a machine makes mistakes. The system should therefore be able to inform the user about the causes and the resulting impacts of the failure to the system and the mission, and ideally, also be able to present information on how to diagnose and mitigate the errors.[63]

Related to transparency is the issue of explainability, or the ability of an AI system to explain its rationale to human users, articulate its strengths and weaknesses, and convey how it will behave in the future. There seems to be a consensus that in order to properly calibrate trust and use AI systems effectively, people need to understand how these systems work and why they reach the conclusions that they do.[64] Research on explainability therefore tackles the black-box problem with AI (and more specifically ML-based systems)—namely, that many algorithms, including those based on deep learning, are opaque to users, with few mechanisms available for explaining their reasoning and results. Part of the challenge is that while AI techniques such as decision tree induction have built-in explanations, so far they have been generally less accurate compared to more complex deep learning algorithms, which perform better but are less explainable.[65] With the current state of the technology, developers therefore face a tradeoff in their choice of algorithm, whether to optimize for performance or for explainability—with both parameters being pertinent to achieving and maintaining trust in human-machine teams.

One of the key DoD research initiatives in this area is DARPA's "Explainable Artificial Intelligence (XAI)" program, focusing on three interrelated challenges: developing new ML techniques that produce more explainable models, designing new strategies and techniques for human-computer interaction and intelligent user interfaces for conveying effective explanations, and investigating the psychological requirements for effective explanations that help humans intuitively and quickly understand the system's rationale.[66]

XAI research raises the issue of communication in human-machine teams which speaks to both system design, (i.e. AI system

characteristics and functionalities) and the interactions between humans and intelligent technologies. As previously noted, different AI representations—physical robots, virtual bots, or embedded AI—evoke different cognitive and emotional responses that impact trust. The field of social robotics has shown that human-like design features, social behaviors, and implicit features and behaviors related to communication such as posture, head or eye movements, or changes in proximity can influence the human team member's understanding and trust.[67] The majority of fielded military systems today, however, have minimal anthropomorphic features. User displays, ubiquitous in human-machine teaming, are therefore particularly pertinent to communication in human-machine teams, and more specifically to how the system should convey information to best calibrate trust.

User displays vary significantly in their design and functionalities depending on the nature of the human-machine interaction, including factors such as task allocation, decision-making authority, and environmental constraints. Across these different configurations, however, decisions about interface design, layout, and graphics that visualize information all have an impact on the operator's perception of the system's current plans, comprehension of the system's behavior, and projection of future outcomes.[68]

An intuitive, easy-to-use interface can improve the operator's situational awareness and limit ambiguity by assuring human team members that the agent is aware of its environment. Such design can increase overall team performance by reducing communication times and minimizing errors. A smart, well-designed interface can even reduce the cognitive workload of human teammates by allowing the agent to perform tasks such as analysis, perception, or navigation best suited to its capabilities.[69] However, there is a delicate balance between ease of use and ensuring trustworthiness through transparency: the most detailed interfaces that provide information about the intricate inner workings of the system may be transparent, but are not necessarily the most user-friendly or conducive to optimal operator performance and effective human-machine teaming.

Currently, the technical challenges facing the development of displays and related audio or visual communication modalities are intertwined with the limits of AI technologies. Intelligent agents struggle with interpreting complex and ambiguous situations, understanding they have made a mistake and communicating the reasoning behind their decisions. With these underlying technological limitations, it is hard to know what type, how much, and how often information should be provided to the human operator. But as AI systems continue to evolve both technologically and socially, and human-machine teams proliferate across multiple tasks and domains, understanding the effects of displays on trust will become increasingly important.

Finally, while much of the discussion above has focused on the factors influencing trust between human operators and technology, there are other forms of transparency that may influence the trust of other audiences and publics. One such approach to ensuring transparency speaks to the need for traceable and auditable data sources, design procedures, and development processes of AI systems. In the commercial space, technology companies such as IBM have made trust and transparency a part of their operating principles, pushing for greater clarity on who trains AI systems, what data is used in training, and what goes into an algorithm's recommendations.

DoD AI ethics principles also call for traceable AI systems, stressing that technical experts within DoD need to possess an appropriate understanding of the technology, development processes, and operational methods of its AI systems, including "transparent and auditable methodologies, data sources, and design procedure and documentation."[70] Documentation practices will help ensure that AI systems are used appropriately, responsibly, and ethically, and that users are able to calibrate expectations and trust in what the system can and cannot do in a given context.[71] Moreover, auditability may prove useful for restoring trust in the event of machine malfunction or an accident, providing a track record of what happened and how such incidents can be avoided in the future.

Reliability, Robustness, and Responsiveness

Whether the operator can trust the AI system to function properly is perhaps the most fundamental question in human-machine teaming. Indeed, while transparency and explainability are important for calibrating trust, reliability may be even more critical. For instance, one of the aforementioned ARPI studies found that participants' trust in the autonomous squad member declined when the robotic mule made errors, and that displaying information to support transparency did not mitigate the impact of the errors on trust. On the other hand, when the autonomous squad member was reliable, participants anthropomorphized the agent more than when it was unreliable, ranking it as more likable, intelligent, and safer to work with.[72] Another study in which a human teammate worked with a robot in reconnaissance missions found a similar trend: when the robot's ability was high and it proved reliable, the explanations it provided about its decisions had no significant impact on trust.[73] While these studies suggest reliability may trump transparency and explainability for engendering and calibrating trust in human-machine teams, additional research on the interaction between reliability and explainability could offer further clarity and nuance.

Alongside reliability and robustness, advances in machine intelligence and capabilities that allow the technology to interact with the environment and be responsive to users also impact trust. Responsiveness, adaptability, cooperation, and pro-social behavior of intelligent technologies strengthen cognitive trust by raising expectations of high-quality performance and positive experience during mutual tasks or missions.[74] Machine behaviors that reflect social intelligence like active listening and personalization have also been linked to higher levels of emotional trust, with users reporting greater levels of engagement, likeability, and enjoyment.[75] Moreover, experimental research shows that cooperative behavior of intelligent agents can increase human-machine team performance as well as support resilience.[76]

There are non-negligible technical challenges to progress in research centered on ensuring that AI systems are reliable, robust, and responsive, especially in complex adversarial environments.

That said, experimentation and fielding of certain systems earlier in the development cycle can provide an opportunity for incorporating user feedback that helps the system learn and improve, as well as for building trust in human-machine teams. According to Mark Lewis, former Acting Deputy Under Secretary of Defense for Research and Engineering, one of the main goals is to figure out applications will have the biggest impact on the warfighter. "In some cases," Lewis explained, "that means getting the technologies in the hands of the warfighter and having them play with them, experiment with them, and figure out what makes their job more effective … [and] easier," as well as "to discard the things that don't buy their way into the war fight."[77]

One example of such efforts is DARPA's Air Combat Evolution (ACE) program which aims to increase warfighter trust in autonomous systems by using human-machine collaborative dogfighting (air-to-air combat) as its initial challenge scenario. As AI systems train in the rules of aerial dogfighting, their performance will be monitored by fighter instructor pilots which will help mature the technology. Once the human pilots feel the AI algorithms are trustworthy in handling the bounded and predictable environment, aerial engagement scenarios will grow more difficult and realistic, eventually going from virtual testing to demonstrating dogfighting algorithms on live, full-scale manned-unmanned teams.[78]

As a whole, building transparent, explainable, auditable, reliable, robust, and responsive intelligent systems will help foster appropriate trust in human-machine teams. Continual feedback between humans—developers, operators, commanders—and machines during the entire lifecycle of a system is another key element of the systems engineering approaches that seek to 'build in' trust into the intelligent machines. Such feedback is also instrumental to what some have referred to as a human-centric approach to AI development which seeks to integrate the needs, perceptions, and behaviors of the user into the design of AI systems.[79] That said, technological solutions alone cannot solve the trust problem in human-machine teams.

U.S. Military Research: Gaps and Future Directions

In October 2020, CSET published a report on U.S. military investments in autonomy and AI, analyzing publicly available data from the FY2020 research, development, testing, and evaluation budget justification books of the Army, Air Force, Navy, and DARPA, focusing specifically on basic, applied, and advanced research.[80] The findings showed that human-machine collaboration and teaming is a crosscutting theme across autonomy and AI research and development programs related to unmanned systems, information processing, decision support, targeting functions, and other areas. That said, only 18 of the 789 research components related to autonomy and 11 out of the 287 research components related to AI mentioned the word "trust."[*81]

There are a number of possible explanations for this apparent gap. For one, while there is a rich literature on human-automation interactions, and the role of trust therein, there is far less research on human-autonomy and human-AI interactions, and specifically on trust in human-autonomy and human-AI teams.[82] Technology, it seems, has outpaced research on human-machine teaming. The U.S. military is developing autonomous systems capable of performing an ever-increasing range of tasks with limited, if any, human supervision and ML-based systems that learn and adapt to their environment. Yet much of what we know about trust in intelligent technologies still draws on research examining human interactions with automated systems and more traditional expert systems. The gap in research on trust in autonomy and AI in DoD's

---

[*] While we found relatively few instances where the word "trust" was mentioned, descriptions of different autonomy and AI research initiatives also included other keywords that signal research related to trust in human-machine teams, including but not limited to: assurance, reliability, robustness, resilience, predictability, explainability, interpretability, transparency, etc. These system features and characteristics are pertinent to trust, and can be thought of as elements of trust and components of effective human-machine teaming. But they are not synonymous with trust.

science and technology program reflects the broader state of the field.[83]

Furthermore, as previously noted, trust is a complex, abstract and hard to measure concept. Defense research therefore tends to favor technology-centric approaches geared more directly toward enhancing AI system attributes that are related to trust, including security, robustness, resilience, and reliability. DARPA leads in research focused on developing systems that behave reliably in operational settings and strengthening security in the face of adversarial attacks, with programs such as "Guaranteeing AI Robustness against Deception (GARD)," "Lifelong Learning Machines (L2M)," and "Assured Autonomy."[84] The Army also has several relevant initiatives. For instance, as part of its basic research portfolio, the "Army Collaborative Research and Tech Alliances" effort includes research on "AI-enabled cyber security that is robust to enemy deception," supporting "Army counter-AI against near-peer adversaries."[85]

These efforts represent systems engineering approaches that seek to "build trust into the system," and are indeed necessary for establishing and properly calibrating trust in human-machine teams. But they are not sufficient.

For intelligent machines to become true teammates, they need to be able to adapt to changing and new environments. The U.S. military has a number of research programs focused on assurance approaches for systems with advanced levels of autonomy that continue to learn and evolve after they are deployed. Yet the very ability to learn and adapt to the environment, as Heather Roff and David Danks argue, could undermine the human team members' trust.[86]

Trust, in both human and human-machine relationships, is built on repeated interactions that provide information about values, preferences, beliefs, and other factors that help develop shared goals and expectations, as well as allow people to evaluate risk, especially in high-stakes situations. But as the AI system learns and adapts, it may change, often in ways that are unexpected or not understandable to humans. As Roff and Danks assert, "the

battlespace is a dangerous place to be figuring out the preferences and values of a dynamically adapting weapon, so it is unsurprising that trust will be difficult to establish."[87]

Indeed, one of hardest challenges related to adaptive machine learning and trust in human-machine teams is how to ensure that the trust that has been "earned" by a system in a predictable, fixed environment translates not only to different, dynamic environments, but also to different machines as new team members and/or with new human team members.[88] Dispositional factors such as age, gender and cultural background impact people's attitudes and trust in technology. People behave differently and often unpredictably under stress and in high-stakes situations. Emotional factors, previous experiences with intelligent technologies as well as broader institutional and societal structures, and organizational culture all play a role in shaping the nature of trust in human-machine teams. Thus, while trust requirements built into a given system may cultivate appropriate trust in a particular human-machine team, there is no guarantee this "built in" trust holds for new human team members.[89]

For example, a recent study from the Army Research Lab examined soldiers' trust in their robotic teammates in autonomous driving scenarios by grouping individuals in four different categories based on "demographics, personality traits, responses to uncertainty, and initial perceptions about trust, stress, and workload associated with interaction with automation."[90] Based on a facial expressivity analysis, the researchers found that these groups had unique differences in their responses and attitudes toward the driving automation. The study therefore concluded that trust calibration metrics may not be the same for all groups of people and that trust-based interventions, such as changes in user display features or communication of intent, "may not be necessary for all individuals, or may vary depending on group dynamics."[91]

This report does not advocate for the study of trust as an end to itself. Rather, we suggest that research focused explicitly on the drivers and dynamics of trust in human-machine teams can augment technology-centric approaches to building trust into AI systems. With this in mind, we offer the following directions for

continued and additional research that could contribute to advances in human-machine teaming and the development of trustworthy AI systems.

- Multidisciplinary research on the drivers of trust in human-machine teams, specifically under operational conditions. Research on human-machine trust, including scholarship that applies sophisticated computation models of cognition to understand issues such as knowledge acquisition and problem solving, is predominantly conducted under closely controlled laboratory conditions.[92] More research is needed to assess whether these findings withstand complex real-world conditions and tasks.

- Collaborative research between U.S.-based researchers and defense research communities in allied countries to assess how cross-cultural variation in trust in human-machine teams may impact interoperability in multinational operations.

- Research to assess what aspects of transparency are most relevant for calibrating trust in human-machine teams, especially under operational conditions. For instance, how important is explainability vs. auditability, i.e., is the ability to understand how AI systems reach a particular conclusion more conducive to building, maintaining, and adjusting trust in human-machine teams than visibility into the data and models?

- Research on the interaction between explainability and reliability. While there seems to be a consensus that in order to trust their machine teammates, humans need to understand why autonomous and AI-enabled systems behave as they do, some research suggests that as long as the system is reliable, explainability is less important. Additional research could help connect and contextualize these seemingly contrasting views.

- Research on shifts in cognitive workloads and trust calibration across different types of human-machine

teaming. For example, research on autonomous vehicle technology for Army convoy operations shows that in a mix of manned and unmanned trucks, the soldiers who remain in the convoy would perform more tasks involving sensing and decision-making, resulting in a higher cognitive burden than their counterparts in a fully manned convoy (where cognitive burden can be shared across a larger number of soldiers).[93] This is significant considering there is evidence that users make more automation bias errors under higher workload conditions, when performing complex tasks or multitasking.[94] As such, there is a need for more research on how the distribution of tasks and decision-making responsibility in human-machine teams (and the resultant shifts in cognitive workloads) affect trust specifically in military settings.

- Research on uncertainty and trust calibration. What aspects of uncertainty are most critical for humans to understand in order to calibrate trust and use the system effectively, and how should this information be communicated?

- Research on reliability and trust calibration. Keeping in mind the growing urgency to field military AI systems, what are the minimum standards for AI system reliability, robustness, and resilience necessary for building and maintaining trust in human-machine teams? How do these standards vary based on operator characteristics, mission, environmental conditions, and the distribution of tasks and decision-making authority within the human-machine team?

While this is certainly not an exhaustive list, we believe additional research on these topics could help advance the U.S. military's vision of using intelligent machines as trusted partners to human operators as well as further the development of reliable, trustworthy, and safe AI systems that would cement U.S. military and technological advantages into the future.

Conclusion

The U.S. military sees many uses for human-machine teams, and with advances in AI technology, machines will be able to take on a greater variety of tasks and responsibilities, extending human-machine teaming to additional mission areas and functions. But progress toward advanced human-machine teaming will depend on advances in understanding human attitudes toward technology as well as breakthroughs in AI technologies, making these systems more transparent, explainable, auditable, reliable, robust, and responsive.

We offer a number of research directions that could help the U.S. military move forward with its vision of using intelligent machines as trusted partners to human operators: greater emphasis on research and experimentation under operational conditions; collaborative research with allied countries; research on trust and various aspects of transparency; research on the intersection of explainability and reliability; research on trust and cognitive workloads; research on trust and uncertainty; and research on trust and reliability.

As the U.S. military integrates AI technologies and capabilities into the force, resolving outstanding questions around the issue of trust in human-machine teams becomes increasingly imperative. There are no simple solutions and no single approach will suffice. But insights from research on the dispositional, situational, and learned factors that shape trust as well as the broader institutional and societal structures that influence people's attitudes and behaviors toward technology can inform and strengthen systems engineering approaches to building trustworthy AI.

Authors

Dr. Margarita Konaev is a research fellow with CSET. Tina Huang is a research analyst with CSET currently serving as a fellow in artificial intelligence policy for a member of Congress with a leadership role in AI issues. Husanjot Chahal is a research analyst with CSET.

Acknowledgements

This research has benefited a great deal from the insights of our CSET colleagues, Helen Toner, Tim Hwang, John Bansemer, Igor Mikolic-Torreira, and Melissa Flagg. We are especially grateful for the extensive and thoughtful feedback provided by Erin K. Chiou of Arizona State University and Larry Lewis of the Center for Naval Analysis. Many thanks as well to Matt Mahoney for fact-checking, Melissa Deng for formatting, and to Lynne Weil and the CSET external affairs team for editorial support.

Document Identifier: doi: 10.51593/20200024

Endnotes

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

[2] National Security Commission on Artificial Intelligence, *Key Considerations for Responsible Development and Fielding of Artificial Intelligence,* (Washington, DC: July 2020), 30, https://www.nscai.gov/reports.

[3] John D. Winkler, Timothy Marler, Marek N. Posard, Raphael S. Cohen, and Meagan L. Smith, "Reflections on the Future of Warfare and Implications for Personnel Policies of the U.S. Department of Defense" (RAND Corporation, 2019), https://www.rand.org/content/dam/rand/pubs/perspectives/PE300/PE324/RAND_PE324.pdf.

[4] Andrew Ilachinski, "AI, Robots, and Swarms" (Center for Naval Analysis, January 2017), 49, https://www.cna.org/cna_files/pdf/DRM-2017-U-014796-Final.pdf.

[5] The Department of Defense AI Strategy defines AI as "the ability of machines to perform tasks that normally require human intelligence." U.S. Department of Defense, *Summary of the 2018 Department of Defense Artificial Intelligence Strategy*; for a practical summary of AI and ML technology, see Greg Allen, "Understanding AI Technology" (Joint Artificial Intelligence Center, U.S. Department of Defense, April 2020), https://www.ai.mil/docs/Understanding%20AI%20Technology.pdf.

[6] RAND Office of Media Relations, "Autonomous Vehicle Technology May Improve Safety for U.S. Army Convoys," RAND Corporation, February 12, 2020, https://www.rand.org/news/press/2020/02/12.html.

[7] Shawn McKay, Matthew E. Boyer, Nahom M. Beyene, Michael Lerario, Matthew W. Lewis, Karlyn D. Stanley, Randall Steeb, Bradley Wilson, and Katheryn Giglio, "Automating Army Convoys: Technical and Tactical Risks and Opportunities" (RAND Corporation, 2020), https://www.rand.org/pubs/research_reports/RR2406.html.

[8] Rachel S. Cohen, "USAF Solicitation Offers Seed Money for Skyborg," *Air Force Magazine*, May 20, 2020, https://www.airforcemag.com/usaf-solicitation-offers-seed-money-for-skyborg/; Valerie Insinna, "More Than One Company Could Get Cash to Build the Air Force's AI-Equipped Skyborg Drone," *Defense News*, May

20, 2020, https://www.defensenews.com/air/2020/05/20/more-than-one-company-could-get-cash-to-build-the-air-forces-ai-equipped-skyborg-drone/.

[9] Andrew J. Hawkins, "Serious Safety Lapses Led to Uber's Fatal Self-Driving Crash, New Documents Suggest," *The Verge*, November 6, 2019, https://www.theverge.com/2019/11/6/20951385/uber-self-driving-crash-death-reason-ntsb-dcouments.

[10] Jackie Snow, "Computer Vision Algorithms Are Still Way Too Easy to Trick," *MIT Technology Review*, December 2020, https://www.technologyreview.com/2017/12/20/146692/computer-vision-algorithms-are-still-way-too-easy-to-trick/; "Black-box Adversarial Attacks with Limited Queries and Information," LabSix Press, https://www.labsix.org/press/.

[11] Office of Prepublication and Security Review, *Future Directions in Human Machine Teaming Workshop* (Washington, DC: Department of Defense, January 15, 2020), https://basicresearch.defense.gov/Portals/61/Future%20Directions%20in%20Human%20Machine%20Teaming%20Workshop%20report%20%20%28for%20public%20release%29.pdf.

[12] Margarita Konaev, Husanjot Chahal, Ryan Fedasiuk, Tina Huang, and Ilya Rahkovsky, "U.S. Military Investments in Autonomy and AI: A Budgetary Assessment" (Center for Security and Emerging Technology, October 2020), https://cset.georgetown.edu/research/u-s-military-investments-in-autonomy-and-ai-a-budgetary-assessment/; Margarita Konaev, Husanjot Chahal, Ryan Fedasiuk, Tina Huang, and Ilya Rahkovsky, "U.S. Military Investments in Autonomy and AI: A Strategic Assessment" (Center for Security and Emerging Technology, October 2020), https://cset.georgetown.edu/research/u-s-military-investments-in-autonomy-and-ai-a-strategic-assessment/.

[13] We are grateful to Erin K. Chiou for this important observation.

[14] There are many ways to define trust. One commonly used definition comes from Lee and See's 2004 review of research on trust in automation which defines trust as "the attitude that an agent will help achieve an individual's goal in a situation characterized by uncertainty and vulnerability." See John D. Lee and Katrina A. See, "Trust in Automation: Designing for Appropriate Reliance," *Human Factors* 46, no. 1 (2004): 50–80, http://user.engineering.uiowa.edu/~csl/publications/pdf/leesee04.pdf; Another study on measuring trust in human-robot collaboration offers that "trust in automation refers to the trust in the capability of a system to accomplish an individual's goals." See Amos Freedy, Ewart DeVisser, Gershon Weltman, and Nicole Coeyman, "Measurement of Trust in Human-Robot Collaboration," presented at *International Symposium on Collaborative Technologies and Systems* (Orlando, FL, June 2007), doi:10.1109/CTS.2007.4621745; Many

studies build on Mayer et al.'s ABI model which identifies ability, benevolence, and integrity as the foundational principles of trust, with later research by Dietz and Den Hartog adding predictability or reliability as another important factor for shaping trustworthiness. See Roger C. Mayer, James H. Davis, and F. David Schoorman, "An Integrative Model of Organizational Trust," *The Academy of Management Review* 20, no. 3 (July 1995): 709–734; Graham Dietz and Deanne N Den Hartog, "Measuring Trust Inside Organizations," *Personnel Review* 35, no. 5 (September 2006): 557–588; Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad van Moorsel, "The Relationship Between Trust in AI and Trustworthy Machine Learning Technologies," *arXiv [cs.CY]* (November 27, 2019), arXiv, https://arxiv.org/abs/1912.00782.

[15] Ella Glikson and Anita Williams Woolley, "Human Trust in Artificial Intelligence: Review of Empirical Research," *Academy of Management Annals* 14, no. 2 (2020): 627–660, https://doi.org/10.5465/annals.2018.0057.

[16] Glikson and Woolley, "Human Trust in Artificial Intelligence."

[17] Defense Innovation Board, *AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense*, Supporting Document (Washington, DC: October 2019), https://media.defense.gov/2019/Oct/31/2002204459/-1/-1/0/DIB_AI_PRINCIPLES_SUPPORTING_DOCUMENT.PDF.

[18] John D. Lee and Katrina A. See, "Trust in Automation: Designing for Appropriate Reliance," *Human Factors* 46, no. 1 (2004): 50–80, http://user.engineering.uiowa.edu/~csl/publications/pdf/leesee04.pdf.

[19] Glikson and Woolley, "Human Trust in Artificial Intelligence," 629.

[20] Wan-Lin Hu, Kumar Akash, Neera Jain, and Tahira Reid, "Real-Time Sensing of Trust in Human-Machine Interactions," *Cyber-Physical and Human-Systems* 49 no. 32 (2016): 48–53, https://doi.org/10.1016/j.ifacol.2016.12.188; Catherine Neubauer, Gregory Gremillion, Kristin E. Shaefer, Brandon S. Perelman, Claire La Fleur, and Jason S. Metcalfe, "Analysis of Facial Expressions: Explaining Affective State and Trust-Based Decisions during Interaction with Automation" (Aberdeen Providing Ground, MD: CCDC Army Research Laboratory, April 2020), https://apps.dtic.mil/sti/pdfs/AD1098113.pdf.

[21] For a review of different approaches to measuring trust, see Glikson and Woolley, "Human Trust in Artificial Intelligence"; Bing Cai Kok and Harold Soh, "Trust in Robots: Challenges and Opportunities," *Current Robotics Reports* 1 (September 2020): 297–309, https://doi.org/10.1007/s43154-020-00029-y.

[22] Glikson and Woolley, "Human Trust in Artificial Intelligence," 647.

[23] Gary Palmer, Anne Selwyn, and Dan Zwillinger, "The "Trust V": Building and Measuring Trust in Autonomous Systems," in *Robust Intelligence and Trust in Autonomous Systems*, edited by Ranjeev Mittu, Donald Sofge, Alan Wagner, and W.F. Lawless (New York, NY: Springer, 2016), 65, https://doi.org/10.1007/978-1-4899-7668-0.

[24] Ilachinski, "AI, Robots, and Swarms," 183.

[25] Ilachinski, "AI, Robots, and Swarms."

[26] Kimberly F. Jackson, Zahar Prasov, Emily C. Vincent, and Eric M. Jones, "A Heuristic Based Framework for Improving Design of Unmanned Systems by Quantifying and Assessing Operator Trust," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 60, no. 1 (2016): 1696–1700, https://journals.sagepub.com/doi/pdf/10.1177/1541931213601390.

[27] Keng Siau and Weiyu Wang, "Building Trust in Artificial Intelligence, Machine Learning, and Robotics," *Cutter Business Technology Journal* 31, no. 2 (March 2018): 47–53, https://www.cutter.com/article/building-trust-artificial-intelligence-machine-learning-and-robotics-498981.

[28] Ipsos Public Affairs, "Entrepreneurialism," *Ipsos*, November 12-18, 2018, https://www.ipsos.com/sites/default/files/ct/news/documents/2018-10/entrepreneurialism-2018-global-report.pdf; one of the challenges of interpreting the results of such surveys is that it is difficult to ascertain what respondents have in mind when asked about their trust in AI—while some individuals may be interpreting AI to mean an existing system like Siri others could be envisioning a futuristic, fictionalized representation such as the Terminator, which would in turn shape their responses regarding trust. We are grateful to Tim Hwang for this observation.

[29] Christopher Bartneck, Tomohiro Suzuki, Takayuki Kanda, and Tatsuya Nomura, "The Influence of People's Culture and Prior Experiences with Aibo on Their Attitude Towards Robots." *AI and Society* 21 (2007): 217–230.

[30] Erik Lin-Greenberg, "Allies and Artificial Intelligence: Obstacles to Operations and Decision-Making," *Texas National Security Review* 3, no. 2 (Spring 2020): 56–76, https://tnsr.org/2020/03/allies-and-artificial-intelligence-obstacles-to-operations-and-decision-making/.

[31] M.L. Cummings, "Automation Bias in Intelligent Time Critical Decision Support Systems" (American Institute of Aeronautics and Astronautics, July 2012), http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=073721BD8AA0FA36

0C632E 4265F27507?doi=10.1.1.91.2634&rep=rep1&type=pdf; Jackson et al., "A Heuristic Based Framework."

[32] Raja Parasuraman and Dietrich H. Manzey, "Complacency and Bias in Human Use of Automation: An Attentional Integration," *Human Factors* 52, no. 3 (June 2010): 381–410, https://www.depositonce.tu-berlin.de/bitstream/11303/8923/1/Parasuraman_Manzey_2010.pdf.

[33] On training, trust, and better performance, see Javier Chagoya, "NPS Researchers, Marines Explore The Trust Factor in Human-Machine Teaming," United States Marine Corps, December 7, 2020, https://www.marines.mil/News/News-Display/Article/2437025/nps-researchers-marines-explore-the-trust-factor-in-human-machine-teaming/.

[34] John K. Hawley, "Patriot Wars: Automation and the Patriot Air and Missile Defense System" (Center for a New American Security, January 25, 2017), https://www.cnas.org/publications/reports/patriot-wars.

[35] Hawley, "Patriot Wars."

[36] We are grateful to Larry Lewis for providing an assessment of the dispositional, situational, and learned factors that played a role in the Patriot friendly-fire incident. Larry Lewis, "AI and Autonomy: Cultivating Appropriate Trust," working paper.

[37] As John Hawley of the U.S. Army Research Laboratory put it, "the Patriot crews did what they had been trained to do, which was reinforced by the prevailing command climate and widespread, but not generally accurate, beliefs about the system's engagement reliability." Hawley, "Patriot Wars."

[38] Yochanan E. Bigman and Kurt Gray, "People are Averse to Machines Making Moral Decisions," *Cognition* 181 (December 2018): 21–34, https://doi.org/10.1016/j.cognition.2018.08.003.

[39] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey, "Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err," *Journal of Experimental Psychology*: General 144, no. 1 (2015): 114–126.

[40] Dietrich Manzey, Juliane Reichenbach, and Linda Onnasch, "Human Performance Consequences of Automated Decision Aids: The Impact of Degree of Automation and System Experience," *Journal of Cognitive Engineering and Decision Making* 6, no.1 (January 2012): 57–87; Kevin Anthony Hoff and Masooda Bashir, "Trust in Automation: Integrating Empirical Evidence on Factors that Influence Trust," *Human Factors* 57, no. 3 (2014): 407–434, https://doi.org/10.1177/0018720814547570.

[41] Dietvorst, et al., "Algorithm Aversion."

[42] Dietvorst et al., "Algorithm Aversion"; Munjal Desai, Poornima Kaniarasu, Mikhail Medvedev, Aaron Steinfeld, and Holly Yanco, "Impact of Robot Failures and Feedback on Real-Time Trust," *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction*, (2013): 251–258, https://ieeexplore.ieee.org/document/6483596.

[43] Ewart J. de Visser, Richard Pak, and Mark A. Neerincx, "Trust Development and Repair in Human-Robot Teams," *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (March 2017): 103–104, https://dl-acm-org.proxy.library.georgetown.edu/doi/pdf/10.1145/3029798.3038409.

[44] DSIAC, "Warfighter Trust in Autonomy," *DSIAC Journal* 4, no. 4 (November 2017), https://issuu.com/dsiac/docs/dsiac-fall-2017-volume-4-number-4_1.

[45] We are grateful to Erin K. Chiou for raising this crucial point. See also Erin K. Chiou and John D. Lee, "Beyond Reliance and Compliance: Human-Automation Coordination and Cooperation," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 59, no. 1 (2015): 195–199, https://doi.org/10.1177/1541931215591040; Ben Shneiderman, "Human-Centered Artificial Intelligence: Reliance, Safe, and Trustworthy," *arXiv [cs.HC]* (February 23, 2020), arXiv, https://arxiv.org/abs/2002.04087v2.

[46] DARPA, *With Squad X, Dismounted Units Partner with AI to Dominate Battlespace* (Washington, DC: Defense Advanced Research Projects Agency, 2019), https://www.darpa.mil/news-events/2019-07-12.

[47] Vijay N. Gadepally, Braden J. Hancock, Kara B. Greenfield, Joseph P. Campbell, William M. Campbell, and Albert I. Reuther, "Recommender Systems for the Department of Defense and Intelligence Community," *Lincoln Laboratory Journal* 22, no. 1 (2016): 74–89, https://www.ll.mit.edu/sites/default/files/publication/doc/2019-04/recommender-systems-department-defense-intelligence-gadepally-108929.pdf; Margarita Konaev, "With AI, We'll See Faster Fights, But Longer Wars," War on the Rocks, October 29, 2019, https://warontherocks.com/2019/10/with-ai-well-see-faster-fights-but-longer-wars/.

[48] Whether human operators follow the system's recommendations depends in part on the content of these recommendations and the context of the mission. System recommendations alerting of potential danger may be heeled differently than less urgent outputs. That said, human operators are less likely to accept and follow a system's recommendations if they don't trust it. We are grateful to John Bansemer for this observation.

[49] For an overview of early studies, see Raja Parasuraman and Victor Riley, "Humans and Automation: Use, Misuse, Disuse, Abuse," *Human Factors* 39, no. 2 (1997): 230–253; For a more recent review of automation bias studies in health care, see David Lyell and Enrico Coiera, "Automation Bias and Verification Complexity: A Systematic Review," *Journal of the American Medical Informatics Association* 24, no. 2 (2017): 423–431, https://doi.org/10.1093/jamia/ocw105.

[50] Linda J. Skitka, Kathleen L. Mosier, and Mark Burdick, "Does Automation Bias Decision-Making?" *International Journal of Human-Computer Studies* 51, no. 5 (November 1999): 991–1006, https://lskitka.people.uic.edu/AutomationBias.pdf.

[51] It is also worth noting that in the experiment the simulator test entails pilots operating B-747, a large, multi-engine jet which can continue operating even after losing an engine. Thus, if alerted of a fire in one of the engines, the reasonable response is to shut down the engine since the risks from doing so are lower than the potentially catastrophic outcome of a fire. A more appropriate test would perhaps entail a single engine jet, such as an FA-18 or FA-35. That said, the fact that all of the experiment participants responded to the fake fire alert by shutting down the engine while all other indicators and engine parameters appeared normal lends validity to the automation bias argument. We are grateful to Igor Mikolic-Torreira for this observation. See Skitka et al., "Does Automation Bias Decision-Making?"

[52] Skitka et al., "Does Automation Bias Decision-Making?"; Cummings, "Automation Bias in Intelligent Time Critical Decision Support Systems"; Kate Goddard, Abdul Roudsari, and Jeremy C. Wyatt, "Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators," *Journal of the American Medical Informatics Association* 19, no. 1 (2012): 121–127, doi:10.1136/amiajnl-2011-000089.

[53] Skitka et al. "Does Automation Bias Decision-Making?"

[54] Cummings, "Automation Bias in Intelligent Time Critical Decision Support Systems"; Jackson et al., "A Heuristic Based Framework."

[55] Forrest E. Morgan, Benjamin Boudreaux, Andrew J. Lohn, Mark Ashby, Christian Curriden, Kelly Klima, and Derek Grossman, *Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World* (RAND Corporation, 2020), 11, 36, https://www.rand.org/pubs/research_reports/RR3139-1.html.

[56] On rethinking human control and calibrating trust, see Larry Lewis, "Redefining Human Control: Lessons From the Battlefield for Autonomous Weapons" (Center for Naval Analyses, March 2018), 7, https://www.cna.org/CNA_files/PDF/DOP-2018-U-017258-Final.pdf.

[57] Transparency, explainability, auditability, reliability, robustness, and responsiveness are central to trustworthiness for reasons discussed throughout the report. Our focus on these elements also builds on the discussion of considerations on ethical and trustworthy AI in NSCAI's interim report which mentions reliability, robustness, auditability, explainability and fairness. See National Security Commission on Artificial Intelligence, *Interim Report* (Washington, DC: November 2019), 48, https://www.nscai.gov/reports; For an overview of the different elements identified in a variety of international Principled AI policy and technology frameworks, see Ehsan Toreini et al., "The Relationship Between Trust in AI and Trustworthy Machine Learning Technologies."

[58] Toreini et al., "The Relationship Between Trust in AI and Trustworthy Machine Learning Technologies"; Defense Science Board, *Summer Study on Autonomy* (Washington, DC: Department of Defense, June 2016), 1, https://www.hsdl.org/?view&did=794641.

[59] Heather M. Roff and David Danks, "Trust but Verify: The Difficulty of Trusting Autonomous Weapons Systems," *Journal of Military Ethics* 17, no.1 (2018): 2–20; Michelle A. Flournoy, Avril Haines, and Gabrielle Chefitz, "Building Trust through Testing" (Center for Security and Emerging Technology, October 2020), https://cset.georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf; Office of the Secretary of Defense, "2019 Advancements in Test and Evaluation of Autonomous Systems (ATEAS) Workshop Report" (Air Force Institute of Technology, August 31, 2020), https://www.afit.edu/stat/statcoe_files/1015AFIT2020ENS09119%201015true%202-2.pdf.

[60] Jessie Y.C. Chen, Michael J. Barnes, Anthony R. Selkowitz, and Kimberly Stowers, "Effects of Agent Transparency on Human-Autonomy Teaming Effectiveness," *2016 IEEE International Conference on Systems, Man, and Cybernetics* (SMC) (October 2016): 001838–001843, doi: 10.1109/SMC.2016.7844505.

[61] Jessie Y. C. Chen, Katelyn Procci, Michael Boyce, Julia Wright, Andre Garcia, and Michael Barnes, "Situation Awareness-Based Agent Transparency (Technical Report: ARL-TR-6905)" (Aberdeen Proving Ground, MD: US Army Research Laboratory, 2014), https://apps.dtic.mil/dtic/tr/fulltext/u2/a600351.pdf; J. Mercado, M. Rupp, J. Chen, D. Barber, K. Procci, and M. Barnes, "Intelligent Agent Transparency in Human-Agent Teaming for Multi-UxV Management," *Human Factors* 58 (2016): 401–415; J. M. McGuirl and N. B. Sarter, "Supporting Trust Calibration and the Effective Use of Decision Aids by Presenting Dynamic System Confidence Information," *Human Factors* 48 (2016): 656–665; J. Beller, M. Heesen, and M. Vollrath, "Improving the Driver-

Automation Interaction: An Approach Using Automation Uncertainty," Human Factors 55 (December 2013): 1130–1141.

[62] Chen et al., "Effects of Agent Transparency on Human-Autonomy Teaming Effectiveness."

[63] Jackson et al., "A Heuristic Based Framework," 1698. In military settings, transperancy about failures is important, but the question of when and how such information should be presented to the human operator depends on the mission and context at hand. Information about system failures must be shared in a way that doesn't distract or endanger the user. We are grateful to John Bansemer for this observation.

[64] Tim Miller, "Explanation in Artificial Intelligence: Insights from the Social Sciences," *Artificial Intelligence* 267 (February 2019): 1–38, https://doi.org/10.1016/j.artint.2018.07.007.

[65] National Science and Technology Council, *The National Artificial Intelligence Research and Development Strategic Plan* (Washington, DC: Executive Office of the President of the United States, October 2016), https://www.nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf.

[66] David Gunning and David W. Aha, "DARPA's Explainable Artificial Intelligence Program," *AI Magazine* 40, no. 2 (Summer 2019): 44–58, https://doi.org/10.1609/aimag.v40i2.2850. Notably, there are some reservations about whether the definition of explainability adopted by XAI is sufficiently informed by previous research, see Miller, "Explanation in Artificial Intelligence," 2019.

[67] Kristin E. Schaefer, Edward R. Straub, Jessie Y.C. Chen, Joe Putney, and A.W. Evans III, "Communicating Intent to Develop Shared Situation Awareness and Engender Trust in Human-Agent Teams," *Cognitive Systems Research* 46 (December 2017): 26–39, https://doi.org/10.1016/j.cogsys.2017.02.002.

[68] Lee and See, "Trust in Technology."

[69] Schaefer et al., "Communicating Intent to Develop Shared Situation Awareness."

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

[71] National Security Commission on Artificial Intelligence, *First Quarter Recommendations* (Washington, DC: March 2020), 71.

[72] J.L. Wright, J.Y.C. Chen, and S. G. Lakhmani, "Agent Transparency and Reliability in Human–Robot Interaction: The Influence on User Confidence and Perceived Reliability," *IEEE Transactions on Human-Machine Systems* 50, no. 3 (June 2020): 254–263, https://ieeexplore.ieee.org/document/8795544#full-text-header.

[73] Ning Wang, David V. Pynadath, and Susan. G. Hill, "Trust Calibration Within a Human-Robot Team: Comparing Automatically Generated Explanations," *11th ACM/IEEE International Conference on Human-Robot Interaction* (HRI) (Christchurch, 2016): 109–116, doi: 10.1109/HRI.2016.7451741.

[74] On collaborative behavior, see Erin K. Chiou and John D. Lee, "Cooperation in Human-Agent Systems to Support Resilience: A Microworld Experiment," *Human Factors* 58, no. 6 (September 2016): 846–863, https://journals-sagepub-com.proxy.library.georgetown.edu/doi/pdf/10.1177/0018720816649094.

[75] Glikson and Woolley, 632; Gurit E. Birnbaum, Moran Mizrahi, Guy Hoffman, Harry T. Reis, Eli J. Finkel, and Omri Sass. "Machines as a Source of Consolation: Robot Responsiveness Increases Human Approach Behavior and Desire for Companionship." *ACM/IEEE International Conference on Human-Robot Interaction* (April 2016): 165–171. Available at https://doi.org/10.1109/HRI.2016.7451748.

[76] Chiou and Lee, "Cooperation in Human-Agent Systems to Support Resilience."

[77] Aaron Mehta, "DOD Developing 'Best Practices for AI Programs," *C4ISRNET*, May 19, 2020, https://www.c4isrnet.com/artificial-intelligence/2020/05/19/dod-developing-best-practices-for-ai-programs/.

[78] DARPA, Training AI to Win a Dogfight (Washington, DC: Defense Advanced Research Projects Agency, May 8, 2019), https://www.darpa.mil/news-events/2019-05-08.

[79] Shneiderman, "Human-Centered Artificial Intelligence."

[80] Margarita Konaev, Husanjot Chahal, Ryan Fedasiuk, Tina Huang, and Ilya Rahkovksy, "U.S. Military Investments in Autonomy and AI: Costs, Benefits, and Strategic Effects" (Center for Security and Emerging Technology, October 2020). https://cset.georgetown.edu/research/u-s-military-investments-in-autonomy-and-ai-executive-summary/.

[81] These findings are based on an analysis of publicly available data. It is of course possible that there are classified efforts exploring these questions within

the different research bodies of DOD. That said, the research gaps we were able to ascertain from publicly facing data could have significant implications on U.S. military's ability to employ AI-enabled technologies as trusted partners to human operators, and more broadly, for DOD's vision for military AI, and are therefore important to address.

[82] Nathan J. McNeese, Mustafa Demir, Erin Chiou, Nancy Cooke, and Giovanni Yanikian, "Understanding the Role of Trust in Human-Autonomy Teaming," *Proceedings of the 52nd Hawaii International Conference on System Sciences 2019*, https://scholarspace.manoa.hawaii.edu/bitstream/10125/59466/0026.pdf.

[83] McNeese et al., "Understanding the Role of Trust."

[84] United States Department of Defense, Fiscal Year (FY) 2020 Budget Estimates, March 2019, *Defense Advanced Research Projects Agency, Defense-Wide Justification Book Volume 1 of 5*, Research, Development, Test & Evaluation, Defense-Wide (Washington, DC: Department of Defense, 2019), vol 1-61.

[85] United States Department of Defense, Fiscal Year (FY) 2020 Budget Estimates, March 2019, Defense Advanced Research Projects Agency, Defense-Wide Justification Book Volume 1 of 5, vol 1-82.

[86] Roff and Danks, "Trust but Verify," 11.

[87] Roff and Danks, "Trust but Verify," 11.

[88] Defense Science Board, *Summer Study on Autonomy*; Ilachinski, "AI, Robots, and Swarms," 186.

[89] Interestingly, some system design approaches recognize that differences between people can affect trust but focus on the developers rather than end users and/or commanders. See Carol J. Smith, "Designing Trustworthy AI: A Human-Machine Teaming Framework to Guide Development," arXiv [cs.AI] (October 8, 2019), arXiv, https://arxiv.org/ftp/arxiv/papers/1910/1910.03515.pdf.

[90] Neubauer et al., "Analysis of Facial Expressions," vi.

[91] Susan Miller, "How's My Driving? Researchers Measure Trust in Autonomous Vehicles," GCN, August 7, 2020, https://gcn.com/articles/2020/08/07/facial-recognition-trust-autonomous-systems.aspx.

[92] Office of Prepublication and Security Review, *Future Directions in Human Machine Teaming Workshop*, 14; Bing Cai Kok and Harold Soh, "Trust in Robots:

Challenges and Opportunities," *Current Robotics Reports* 1, (September 2020): 297–309, https://doi.org/10.1007/s43154-020-00029-y.

[93] McKay et al., *Automating Army Convoys*.

[94] For an overview of automation bias and task complexity, see Lyell and Coiera, "Automation Bias and Verification Complexity."