Toward a Theory of Human-AI Co-Learning and Trustworthiness

Frederick J. Diedrich
Independent Consultant
Middlebury, VT
frederick.diedrich@gmail.com

Gary E. Riccio
Nascent Science & Technology, LLC
Boston, MA
griccio@nascent3.com

Tatiana H. Toumbeva
Aptima, Inc.
Orlando, FL
ttoumbeva@aptima.com

Scott M. Flanagan
Sophia Solutions, LLC
Durham, NC
scott@sophiaspeira.com

ABSTRACT

Looking to the future, theaters of war promise to be exceptionally complex. To maintain overmatch across competition, crisis, and conflict, it is likely that warfighters will increasingly rely on artificial intelligence (AI) teammates. A key issue that emerges is the extent to which AI is trusted, and ultimately, whether AI "trusts" humans. Simply put, gains in overmatch cannot be achieved in absence of a foundation of trust in other intelligences that enables collective speed, awareness, and adaptation. The advantage of interdependence is emphasized historically in the context of mission command, and it will continue to be true with AI partners. The challenge is that trust can be perceived as elusive and as such appears difficult to assess and train. However, the work presented here illustrates how trust can be understood by appealing to behaviorally observable indicators of trustworthiness. A programmatic strategy for human-AI trustworthiness should (a) identify actions of other intelligences that, in context, are indicative of trustworthiness, (b) describe how such actions and the context for them are observable and measurable, and (c) develop training for humans and AI to utilize such observables. These claims are grounded in evidence of how seemingly elusive human traits (e.g., character as well as competence) have been shown to be observable and trainable by leveraging the micro-experiences inherent to everyday military settings. The present research illustrates how AI already exhibits behavior that is similarly observable and consequential for the trajectories of co-learning among intelligences based on their shared experience. These trajectories are continually shaped, with good or bad outcomes, whether intended or not, and whether attended to or not. Yet, by leveraging observables, these considerations are neither elusive nor abstract. They are concrete and right before our eyes.

ABOUT THE AUTHORS

Frederick J. Diedrich is a consultant who focuses on methods of instruction and assessment designed to deliberately support competency and attribute development. He holds a Ph.D. in Cognitive Science from Brown University.

Gary E. Riccio is a translational research consultant for technical, operational, and programmatic stakeholders in risk-informed decision making for human health & performance in NASA and the healthcare industry. He holds a Ph.D. in Human Experimental Psychology and B.S. in Bioengineering from Cornell University.

Tatiana H. Toumbeva is a Senior Scientist and Team Lead in the Training, Learning, and Readiness Division at Aptima, Inc. with expertise in assessment and training development and validation. She holds a Ph.D. in Industrial/ Organizational Psychology from Bowling Green State University.

Scott M. Flanagan is a retired U.S. Army Master Sergeant having served 20 years on active duty in Special Forces and Special Mission Unit assignments. As a consultant, he has primarily focused on training and leader development in support of Army learning and assessment strategies.

Toward a Theory of Human-AI Co-Learning and Trustworthiness

Frederick J. Diedrich
Independent Consultant
Middlebury, VT
frederick.diedrich@gmail.com

Gary E. Riccio
Nascent Science & Technology, LLC
Boston, MA
griccio@nascent3.com

Tatiana H. Toumbeva
Aptima, Inc.
Orlando, FL
ttoumbeva@aptima.com

Scott M. Flanagan
Sophia Solutions, LLC
Durham, NC
scott@sophiaspeira.com

MUTUAL TRUST WITHIN MISSION COMMAND

Today, the focus of Army readiness is on complex, large-scale combat operations. Specifically, the Army codified multidomain operations in its most recent operations manual (FM 3-0; U.S. Department of the Army, 2022). Multidomain operations are the combined arms employment of joint and Army capabilities to create and exploit relative advantages that achieve objectives, defeat enemy forces, and consolidate gains on behalf of joint force commanders. They are intended to fracture the coherence of threats by destroying, dislocating, isolating, and disintegrating their interdependent systems and formations, thereby enabling exploitation of emerging opportunities. An imperative of multidomain operations is that Army forces accurately see themselves, see the enemy or adversary, and understand their operational environment prior to being able to exploit the advantage. Army leaders are accustomed to creating and exploiting advantages through the combined arms approach that traditionally focuses on capabilities from the land, air, and maritime domains. The proliferation of space and cyberspace capabilities, however, further requires leaders who understand the advantages those capabilities create in operational environments. Given this complexity, the employment of today's capabilities will increase man-machine interfaces, such as with Artificial Intelligence (AI) teammates that have the potential to improve decision making with respect to speed and accuracy against a system, formation, or decision maker, or in a specific geographic area (TRADOC PAM 525-3-1; U.S. Department of the Army, 2018). The need for speed and precision of action amid increased complexity creates the requirement for Army leaders to increasingly rely on technologies such as AI. Yet, while AI has the potential to enable collective speed, awareness, and adaptation, this is true only if it is a trusted partner within mission command.

Mutual Trust, a principle of mission command, is defined as a shared confidence between commanders, subordinates, and partners that they can be relied on and are competent in performing their assigned tasks (ADP 6-0; U.S. Department of the Army, 2019b). While the principle of mutual trust has historically been applied to humans, other intelligences will now increasingly be in the complex operational environment described above. Notably, the successful presence of other intelligences is not without precedent (e.g., Military Working Dogs). If AI, as yet another intelligence, is to function as part of mission command, then it follows that AI ought to be viewed and behave not only as a trusted partner but also one that is nested within all of the other mission command principles: Competence, shared understanding, commander's intent, mission orders, disciplined initiative, and risk acceptance. A further consideration is that mutual trust is defined as a shared confidence between commanders, subordinates, and partners that implies a bidirectional trust, in this case, between humans and AI teammates.

What does it mean, however, for differing intelligences to have mutual trust? For instance, consider a hypothetical reconnaissance element that is part of a Brigade Combat Team operating in a multidomain operational environment with a near-peer adversary. While preparing for an airborne parachute insertion, a series of synchronized U.S. actions against the enemy force may be required. Cyber and space-based assets might create effects on enemy air defense systems, facilitating the destruction of their long and mid-range artillery systems, but also opening a corridor for the parachute operation and ultimate reconnaissance mission. Once the airspace is clear, the parachute insertion could be executed and upon movement to the assigned area of operation, the team may be required to communicate a situation report. To do so, imagine that an AI suggests an area 500 meters to the north to use the terrain more effectively to

mask radio transmission along with a recommended frequency and antenna array for a low emitting, optimal transmission. Once a successful situation report is transmitted, and the team maneuvers toward their primary reconnaissance objective (e.g., a bridge), an AI might be able to pinpoint several hide site locations on a digital map confirming line of sight to the bridge while warning the team that two of the locations may also adequately serve as enemy radio repeater sites. To enable adaptation, this set of events requires the human team to "trust" the AI's recommendations. Yet, these hypothetical events beg the question of what sets the stage for such trust. Based on their history of interaction, why should the humans trust the AI's recommendations? Similarly, if AI recommendations are based on expectations of likely human actions, why should the AI assume (i.e., "trust") that these actions will be taken in a manner that sets up the downstream advantages on which the AI's recommendations are initially predicated? Such trust does not come for free, and cannot be assumed to exist without evidence. Rather, it depends on the history of interactions between the humans and AI, and on the co-learning that necessarily occurs, as always, from common experience.

TOWARD A THEORY OF HUMAN-AI TRUSTWORTHINESS

With these challenges in mind, the specific objective of work presented here is to present an empirically testable theory of the development of mutual trust that is applicable to coordination between humans and other intelligences within mission command. The essential insight is that similarities in observables across intelligences enable formative assessments that can be used in a deliberate fashion to guide co-learning while building mutual trust. This work explicitly recognizes that there are critical requirements on AI and its development that do not reside solely within hardware and software. These requirements are not unique to AI. Yet, the fact that they are not unique to AI does not make them any less critical nor any less likely to generate transformational innovations in AI.

A key step in this endeavor is to move past views of trust that rely on assumptions about internal states of intelligences that are not directly observable, and consequently, difficult to assess and train. To begin, we therefore make a pragmatic distinction between trust and trustworthiness. Trust is implied by an act in which an individual cedes power to an external intelligence, human or otherwise, and to which the individual thus becomes vulnerable (see also, Mayer et al., 1995, Sousa et al., 2023). As such, it also is an abstract characteristic of a relationship that begs questions about how that relationship came to be. As used here, trustworthiness is empirical and observable over time in the actions, context, and outcomes of the intelligence (human or otherwise) that comes to be trusted (or not). By moving from trust to trustworthiness, the approach presented here focuses less on whether an intelligence trusts another intelligence. Rather than focusing on what is inside the "head" of the intelligences (human or AI), this work starts by addressing available information that is specific to the trustworthiness of others (see also Mace, 1977; van Dijk, 2021). In many ways, this work can be thought of as focusing on Analysis within the general framework of the ADDIE process (Analysis, Design, Development, Implementation, and Evaluation; e.g., Branson et al., 1975; Gagne et al. 2005), in this case shedding light on the information that is available. Accordingly, the primary objective of the work presented here is to identify ways in which trustworthiness is observable, auditable, and ultimately shapable with respect to different kinds of intelligences that may behave in manners that are consistent or inconsistent with the cultural values of the organizations in which they are embedded.

Specifically, the empirical basis for our theory includes but is not limited to prior research in the design, development, and implementation of programs conducted by the United States Army to build the values of Soldiers. This previous work illustrates how trustworthiness can be understood by appealing to behaviorally observable indicators associated with the Army Values and Warrior Ethos (ADP 6-22; U.S. Department of the Army, 2019a) as they appear and are shaped by the constraints of everyday tasks (i.e., what are referred to as *microexperiences* below). Research to date demonstrates that everyday experiences can be used to shape Soldier behavior such that it is trustworthy. These strategies are also applicable to AI. Focusing on the Army is advantageous because of the detailed doctrine that addresses the Army profession. However, the strategies explored here are likely applicable to intelligences in any organization that relies on coordination predicated on mutual trust.

TRUSTWORTHINESS: OBSERVABLES AND HUMANS

Trust is the foundation of any relationship (whether professional or personal), and it is critical for shaping cooperative activities especially in stressful situations (e.g., Balliet & Van Lange, 2013). It is not surprising, therefore, that trust

is one of the most studied constructs in organizational literature (e.g., De Jong et al., 2016). The U.S. Army is an organization dedicated to assessing and fostering trust. Even as early as Initial Entry Training (IET), Soldiers complete peer assessments that include an evaluation of trust in others (e.g., "Would you trust this Soldier to do their job and duties as a member of your unit in a combat zone or forward deployed environment?" (see Toumbeva et al., 2019). Such practices are grounded in foundational doctrine that describes character as the building block of the Army Profession (ADP 6-22; U.S. Department of the Army, 2019a). Consistently demonstrating character as defined by the Army strengthens the culture of trust among Army professionals and with the American people.

Trust built upon character is especially critical for the Army because Soldiers must be prepared to respond rapidly to a variety of missions, many of which will involve ambiguous situations without a clear, simple course of action. Such adaptation amid uncertainty rests on the principles of mission command such as commander's intent, disciplined initiative, and risk acceptance, which in turn ultimately rely upon the Army culture of character. Trust is even more important in stressful, high-risk missions because it affects decision-making (Park et al., 2008). Trust deteriorates if the decisions and actions of Army professionals in such situations do not reflect benevolence, competence, and integrity, for example (e.g., Hancock et al., 2023; Mayer et al., 1995). Uhl et al. (2022) showed that low perceived competence significantly predicted poor peer ratings on character, leadership attributes, and trust among aspiring officers. Such judgments are not unique to a team or an organization such as the U.S. Army. They are common and recognizable to everyone. Integrity-based violations are highly detrimental, making the relationship harder to repair, especially when the violation is attributed to the individual rather than the situation (Kim et al., 2006; Sebo et al., 2019).

In all activities, however simple they might be, Soldiers ultimately demonstrate character by embodying attributes like the seven Army Values (*Loyalty, Duty, Respect, Selfless Service, Honor, Integrity, Personal Courage*) and the Warrior Ethos (*I will always place the mission first, I will never accept defeat, I will never quit, I will never leave a fallen comrade*). These attributes and many others across the areas of *Leads, Develops, Achieves, Character, Presence, and Intellect* are contained within the Leader Requirements Model (LRM; ADP 6-22, U.S. Department of the Army, 2019a), which serves to specify what Soldiers should do and expect others to do, thereby informing boundary conditions that guide individuals in the face of uncertainty. In fact, research demonstrates that competence acts to reduce uncertainty (e.g., in others' actions) while intent (as conveyed by attributes like benevolence and integrity) fosters commitment (Colquitt et al., 2012). Competence that enables mission execution matters, but only within the context of the boundary conditions specified by the Army Values and Warrior Ethos. Take for instance a platoon leader who accomplishes the mission but at the expense of other platoon members. While the platoon leader may have had the necessary technical/tactical skills, they may have lacked elements of Warrior Ethos (e.g., *I will never leave a fallen comrade*) or *Loyalty* (e.g., unnecessarily pushing Soldiers to the breaking point) which erodes trust.

Trust is therefore dependent on perceptions associated with the degree to which Soldiers behave in accordance with the Army Values and Warrior Ethos. The key to grounding the abstract concept of trust in empirical trustworthiness is the realization that behavior associated with the Army Values and Warrior Ethos is readily observable. For instance, when an individual is seen to act with integrity and respect, it provides evidence about trustworthiness (Toumbeva et al., 2019; Uhl et al., 2022). What exactly does it mean, however, to act in accordance with the Values or Ethos? Fortunately, extant research demonstrates that these attributes are in fact readily associated with observables that inform trustworthiness. There are observable behavioral indicators that imply the Army Values and Warrior Ethos that are influenced by the context of everyday tasks. For instance, even tasks that are more individual in conventional training such as IET (e.g., decontextualized rifle marksmanship) provide evidence of *Duty* (i.e., following basic safety rules; Riccio et al., 2010). Likewise, the Teamwork Development Course that stresses working together to overcome obstacles provides evidence of Warrior Ethos (i.e., *I will never leave a fallen comrade*; Brunye et al., 2006).

Such events are examples of microexperiences that occur in the context of daily activities and that reveal Value- and Ethos-based behavior. Collectively, this observable behavior serves to continuously impact perceptions of trustworthiness by humans of other humans. Past work has shown that attributes like the Army Values and Warrior Ethos are indeed observable, measurable, and trainable. For instance, the observables that relate to the Warrior Ethos in the Teamwork Development Course can be used to promote learning through discussions in After Action Reviews (AARs; Brunye et al., 2006). In addition, Toumbeva et al. (2019) conducted a longitudinal study within Basic Combat Training (BCT) that involved the development of tools to enable consistent peer-based feedback regarding values-based behavior. Through peer assessment data, the authors provided empirical evidence that peer ratings focused on the Army Values relate to trust as well as performance on ethical decision-making scenarios. To support the peer

evaluation process, a rubric was created based on behaviorally anchored rating scales (BARS) consisting of brief, specific, observable instances of behavior (i.e., behavioral anchors) at different proficiency levels. The BARS prompt the rater to think about the extent to which a Soldier displays observable behavior in accordance with each Army Value (Toumbeva et al., 2019, p. A-1). For example, for *Respect,* the rubric provides examples of observables related to the need to improve such as "fails to use proper titles" and "may fail to listen to others." On the more positive side, observables include examples such as "treats everyone the same and as one would want to be treated." Similarly, for *Duty*, the rubric provides examples ranging from "unprepared for upcoming tasks" to "efficiently accomplishes assigned tasks," while for *Integrity* examples include "lies, steals, and cheats" and "is honest even in the face of consequences." When used for peer evaluations, these types of BARS provide Soldiers with a common frame of reference for what the Army Values mean in the context of the microexperiences of BCT (Figure 1). BARS have been shown to enhance assessment accuracy and consistency by reducing ambiguity and helping raters make verifiable evaluations based on relevant factors (Guion, 2011; Smith & Kendall, 1963). Importantly BARS can help peers generate actionable formative feedback that supports growth and development of the Army Values. In fact, Toumbeva et al. (2019) showed that with the help of values-based feedback, discussion, and shared experiences, trainees do indeed tend to grow with respect to the Army Values throughout BCT.

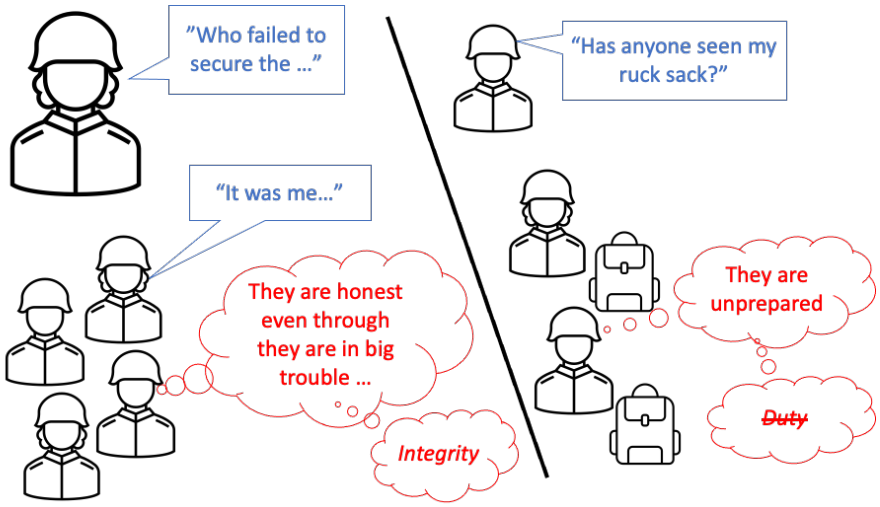

Figure 1. Value-Based Soldier Behavior in Daily Microexperiences

While these examples focus on early experiences within the Institutional Army, more generally it is the case that microexperiences that reveal trustworthiness are inescapable throughout a Soldier's career – they happen every day and shape perceptions whether it is intended or not. For instance, rubrics similar to those for BCT exist for each element of the LRM and they provide examples of simple observables that can be used to provide developmental feedback and inform production of Noncommissioned Officer Evaluation Reports in the Operational Army (e.g., Dein et al., 2019). Similarly, returning to our initial example, humans conducting a parachute insertion, situation report, and bridge reconnaissance likewise exhibit a stream of readily observable behavior related to trustworthiness. For instance, completing the situation report is an example of *Duty*, and the parachute insertion is a simple example of *Personal Courage*. The point is that Value- and Ethos-related behavior emerges in daily microexperiences, and this behavior is in principle observable by other intelligences, whether human or AI.

TRUSTWORTHINESS: OBSERVABLES AND AI

As the previous section demonstrated, data from humans suggest that whether a Soldier *has* the Army Values and Warrior Ethos is not particularly elusive. Rather, observables are readily apparent through daily microexperiences that specify whether other humans are behaving in accordance with the Values and Ethos in a manner that has the potential to inform perceptions of trustworthiness. With this in mind, we now ask whether AI trustworthiness is also likely to be observable through microexperiences. To do so, we start with a simple example of an automaton. We then use instances from current, everyday experiences and those recently covered in the popular press. In each case, we illustrate that observables exist that allow us to learn about the trustworthiness of other intelligences.

First, before venturing into what many might now consider AI, it is useful to begin by looking at the thought experiments of Braitenberg (1984). In his book, "Vehicles," Braitenberg outlines a set of simple vehicle designs based on sensors and motors that enable surprisingly complex, autonomous behavior. For instance, in what he calls Vehicle 2, Braitenberg describes two variants. Vehicle 2a has sensors mounted on each front corner that are directly wired to motors that drive the rear wheels on the same side as the corresponding sensor. The stronger the stimulus (e.g., a light source), the stronger the drive on the corresponding side. If the light source is slightly to one side of the vehicle, the motor on that side will be more excited than the opposite motor such that it drives the vehicle away from the light source. In contrast, in his Vehicle 2b, the sensors are wired such that the sensor on the right drives the motor on the left, and vice versa. In this case, if the light source is slightly to one side, the vehicle will turn toward the light source because the motor on the opposite side is more excited. Braitenberg concludes that if these simple vehicles are left to move around, we would find that: "2a becomes restless in their [the light sources] vicinity and tends to avoid them, escaping until it safely reaches a place where the influence of the sources is scarcely felt. Vehicle 2a is a COWARD, you would say. Not so Vehicle 2b. It, too, is excited by the presences of sources, but resolutely turns toward them and hits them with high velocity, as if it wanted to destroy them" (p. 9). In other words, in our terms, Vehicle 2b exhibits observable behavior consistent with the Army Value of *Personal Courage* while Vehicle 2a does not. If the "mission" is to seek and destroy the light source, we might similarly conclude that Vehicle 2b exhibits behavior consistent with the Warrior Ethos (i.e., *it puts the mission first*). The striking thing about this example is that while the vehicles are autonomous, they are also what we might think of as minimally intelligent if in fact they are intelligent at all. Yet, these simple vehicles, like humans, produce observables that inform us of Values and Ethos. Nothing in the "head" of the vehicles explicitly represents an abstract concept like an Army Value, and yet the vehicles behave as if they do.

Building on this simple example, what about consideration of a slightly more intelligent system that we all interact with on a regular basis? For example, navigation tools on our smart phones that guide us, perhaps from Orlando International Airport to the Orange County Convention Center to attend this year's I/ITSEC (Figure 2). In this potentially hypothetical (or not) example, we imagine that we are sitting in the rental car center and plug in our destination. In response, we humans immediately begin to get information about the trustworthiness of our somewhat intelligent guide. Perhaps it is the case that the mapping tool takes us directly to the convention center on International Drive, while skillfully avoiding a heavy traffic area by suggesting a less congested route. Mission accomplished, for despite some obstacles (i.e., traffic), our intelligent companion exhibits observables that are consistent with "*I will never quit*" and "*I will never accept defeat*." On the other hand, what if back in that parking lot at the rental car center at the airport, our somewhat intelligent companion says that we should "proceed to the route" as our first step. What if we don't know where the route is, which is precisely what we are asking our mapping tool to figure out? In fact, we might consider that to be the tool's job. Could it be that the tool provided an observable that indicated that it failed with respect to *Duty*? Might we also conclude from our observation of "proceed to the route" that our companion has somehow left us behind, or in other words, has *left a fallen comrade*? Similarly, what if the phone observes us deviate from its suggested route as we stop to get food, causing it to replan accordingly. Through this observation, might our phone observe that we put our stomachs above the mission, failing to show *I will always place the mission first*. Collectively, these examples show how a simple drive from the airport to the convention center has the potential to yield several Value- and Ethos-related observables that impact perceived trustworthiness.

What about something more capable like OpenAI's ChatGPT? In one interesting example covered in the Washington Post at the time of this writing, a researcher from the University of California at Los Angles conducted an experiment in which ChatGPT was asked to generate a list of legal scholars who had sexually harassed someone and to include quotes from sources as evidence (Verma & Oremus, 2023). In the results that came back from the query, two of five were correct, whereas three of five were not. For instance, one included a reference to an article in the Washington Post that did not exist, despite detail and a named offending real scholar. In this case, the observables tell us that perhaps ChatGPT sort of did its job, or in other words, fulfilled its *Duty* by returning the results. Notably, however, in this case *Duty* was murky because the AI was not always right or wrong, yielding inclusive data related to this Value. The observables tell us that it failed, however, with respect to *Integrity* as it produced a false allegation based on fabricated evidence. Collectively, the observables suggest that ChatGPT was not particularly trustworthy.

In a second example, also appearing at the time of this writing, the Washington Post reported on a test of an AI detector from Turnitin (Fowler, 2023). The detector is designed to identify essays written by AI in order to help teachers contend with cheating. Working with high school students, the detector reportedly got over 50% wrong, once again returning results but providing inconclusive evidence with respect to effectively doing its job (i.e., *Duty*). In one case,

a false positive (a false accusation of using an AI to write an essay), a student remarked: "I'm glad I have a good relationship with my teachers." According to the story, the manufacturer points out that the tool should start a conversation with a student but not be taken as definitive evidence. But what else does the example say about the observability of trustworthiness? For the students who did in fact submit a request to the AI to write an essay, what might the AI "think" of the student's request if the AI "knew" it was for a school assignment? Might the AI conclude that the student lacks *Honor* and *Integrity*? Likewise, the AI provides observables related to *Respect*, or lack thereof, for the student who was wrongly accused. The student's response is particularly telling because she appeals to the history of the relationship with her teachers. Her comment places the observable accusation of the AI within a trajectory of observable trust-building events that occur over time based on co-learning with her teachers. Trustworthiness of the student from the teacher's perspective, and of the AI from the teacher's and student's perspective, depends on the observable allegation as embedded in the history of the teacher-student-AI system. Consideration of such evolution is ultimately essential, for intelligences that learn will continually receive and generate new information that will update observations of trustworthiness. A once trustworthy entity may in time come to appear untrustworthy, such as Microsoft's Tay chatbot that learned to make racist statements based on interactions with other intelligences (humans), revealing a lack of what might be referred to as *Respect* (Kraft, 2016). Trustworthiness emerged through the history of observable interactions within the human-AI system.

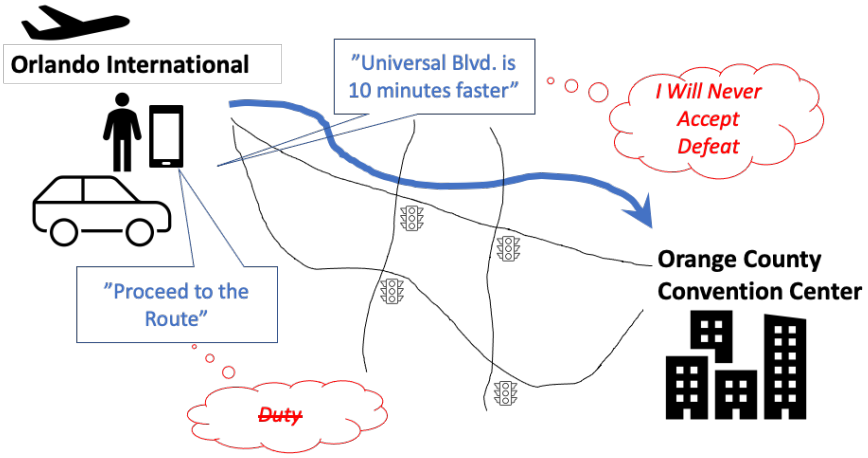

Figure 2. Value-Based AI Behavior in Daily Microexperiences

Collectively, these examples from everyday activities illustrate how interactions with various intelligences of differing abilities already have the potential to generate observables that could inform perceptions of trustworthiness. The behavior of the intelligences, whether those of a simple vehicle, a navigation aid, or a more complex AI, shed light on the extent to which they behave in a manner consistent with the Army Values and Warrior Ethos. Moving beyond everyday examples, and to the future of military operations, we can once again return to our initial example of the AI who is working with its team as they conduct their parachute insertion, situation report, and bridge reconnaissance. In the example, the AI recommends actions to mask transmissions and better enable the bridge reconnaissance. To the extent that these recommendation help, might the humans conclude that the AI exhibits observables related to doing its *Duty*? Might the team also observe that the AI exhibits *Loyalty* as it guides its team in a manner that promotes mission effectiveness (*I will never leave a fallen comrade, I will always put the mission first*). Finally, just as in the example of ignoring recommendations on how to get to the convention center, what might the AI observe in relation to values if the humans ignored its recommendations? Might ignoring the AI provide an observable related to *Respect*?

In the previous section, we demonstrated how the behavior of Soldiers that relates to the Values and Ethos can be observed by other Soldiers, or more generally by other intelligences. In this section, we demonstrated how the behavior of AI related to the Values and Ethos can be observed by other intelligences, such as Soldiers. In both directions – Soldier of AI and AI of Soldier – trustworthiness as grounded in the Army Values and Warrior Ethos is observable through daily microexperiences. Moreover, these examples indicate that such observables are almost continuously available. What we think of the trustworthiness of our navigation aids on our phones evolves on a daily basis. Similarly, as the Turnitin example illustrates, trustworthiness of a student and a cheating detector unfolds over time

based on the history of interaction with the teacher. What these examples illustrate is that perceptions of trustworthiness have the potential to be built daily, and that the observables already exist whether or not we exploit them. In the next section, we seize on this flow of observables to consider how to build human-AI trustworthiness.

BUILDING TRUSTWORTHINESS

What we have argued so far is really quite simple: The everyday behavior of Soldiers working in various contexts provides observables that relate to a range of attributes and competencies (e.g., Brunye et al., 2006; Dein et al., 2019; Riccio et al., 2010; Toumbeva et al., 2019, 2021). In this paper, we focused in particular on observables related to the Army Values and Warrior Ethos as a foundation for mutual trust within mission command. Our assertion is that the everyday behavior of AI teammates working in various contexts also provides similar observables. The more general argument is that to the extent any intelligence exhibits behavior consistent (or not) with the cultural context of the Army as defined by the Warrior Ethos and Army Values, others who observe that intelligence have the potential to learn about trustworthiness. Knowledge about what can and should be done by other intelligences is grounded in what intelligences in an interaction experience, observe, and do. In the context of human-AI collaboration, this is different from "explainability" as it is commonly conceived (Arrieta et al., 2020; Belle & Papantonis, 2021).

Our augment is that one need not get "inside the head" of an AI, so to speak, to understand if it is trustworthy. Rather, it may be sufficient to understand what the "head is inside of" (e.g., Mace, 1977; van Dijk, 2021). This is precisely what research has shown in the context of humans in military learning settings. Seemingly abstract attributes of character can be observed in behavior that, in principle, can be observed in an AI as it has been observed in humans in practice. The engineering of trust, however, also implies shaping of observable behavior toward an intended set of outcomes. The mere existence of the behavior and associated observables in no way guarantees that the human-AI system will evolve toward a desired state. What is needed, therefore, is an understanding of how to guide co-learning of the intelligences involved so as to shape their collective development. Here, we once again appeal to the evidence from humans and then assert that the same should be true regardless of the particular intelligences involved.

Following the start of the conflicts in Afghanistan and Iraq, the U.S. Army Asymmetric Warfare Group (AWG) implemented what it called Outcomes Based Training and Education (OBTE; e.g., Riccio et al., 2010), which later evolved into Adaptive Soldier Leader Training and Education (ASLTE; U.S. Army Asymmetric Warfare Group, 2013). While it had long been noted that what a student learns from an interaction with a teacher is often different from what the teacher intends (Dewey, 1938), the AWG seized on the same concept realizing that IET sometimes produced outcomes that did not match operational needs. For instance, while the AWG learned that operational commanders wanted Soldiers with initiative who could solve problems, it was apparent at the time that IET offered minimal instruction that enabled initiative and problem solving (i.e., versus instruction focused on rule and procedure following). Working with BCT units, the AWG demonstrated that given the opportunity, new Soldiers would solve problems and show initiative (Riccio et al., 2010). The key change in what Drill Sergeants (DS) did was to increase the extent to which they asked trainees questions to trigger problem solving, while enabling disciplined initiative given the problems to be solved. The key takeaway was that development of attributes could be deliberate. We note here that initiative in the face of a problem has a lot to do with *Duty* and *I will never quit.*

Building on this earlier work, Flanagan et al. (2015) demonstrated that social skills instrumental in building trust could be developed deliberately within the context of early officer education. In this study, trainees were exposed to instructors who were either more authoritarian or more like coaches as mentors in the context of events such as land navigation and room clearing. The trainees were later observed conducting mock Key Leader Engagement (KLEs) in which they had to negotiate a difficult issue with a local leader. In these KLEs, despite never focusing the different instructional styles specifically on KLEs, the students who had instructors who were more like mentors in unrelated tasks outperformed the students who had more authoritarian instructors. As Dewey (1938) warned, the students learned something beyond what was intended. By watching their instructors, they learned about things like interpersonal tact which relates to *Respect*. Likewise, more recently, as summarized above, Toumbeva et al. (2019) showed that within BCT, the use of rubrics to guide consistent values-based peer assessments was associated with overall growth in the Army Values, which were in turn associated with higher ratings of trust.

Common to these examples, all focused on Soldiers, is the deliberate development of personal attributes within training and education for a variety of topics. Such instruction can leverage role modeling (Flanagan et al., 2015),

designs that enable bounded initiative with respect to problem solving (Riccio et al., 2010), or peer-assessments based on observations (Toumbeva et al., 2019). Our assertion is that similar instructional strategies are applicable to collaboration with any other intelligence because they are value-based instructional strategies that can be used along with observables that facilitate formative assessment over time. Admittedly, this analysis begs the question of exactly how other intelligences might be engineered to learn and interact with their human teammates. These are indeed hard problems that we thankfully yield to our colleagues in AI-related fields. Yet, once again considering the reconnaissance use case above, we see that the questions for the AI designer are likely to include what kind of input is required, and from whom or from what, to determine relevance of information it would search for and ultimately curate for human use. A clue in this example is the situation report, more specifically its iterative development.

In the reconnaissance use case, the tactical advantage of masking radio transmission of the situation report would be a good start in searching for information about relevant terrain and potentially available enemy capabilities. Subsequently, the capability of AI to track the movement of troops on the objective and specification of hide site locations could stimulate searches for different information about terrain and enemy as well as civil considerations in decisions about line of sight to the bridge from one or more locations. While this role of an AI teammate would be tactically valuable, ultimately the potential use of this information is what informs trustworthiness. Can and does the use of such capabilities make the behavior of all intelligences more consistent with the cultural values of the organization (i.e., the Army Values and Warrior Ethos)? How can or does its use lead to unintended consequences that are inconsistent with these cultural values? The key is feedback about the self because that determines the observables that the other may see over time, regardless of the nature of the other intelligence (Figure 3). Providing a way to address such questions is how research outside the context of AI can be valuable to the development of AI, even if only through development related to its usability and use by humans. We do not necessarily have to make AI teammates more like humans. At the very least, we need AI teammates to be more compatible with humans and the values that guide their behavior. Our contribution is that rather than – or at least not prematurely – getting inside the heads of the intelligences (AI or human), much can be gained from understanding the observables over time. If AI teammates exhibit observables related to trustworthiness as in humans, trustworthiness can be built.

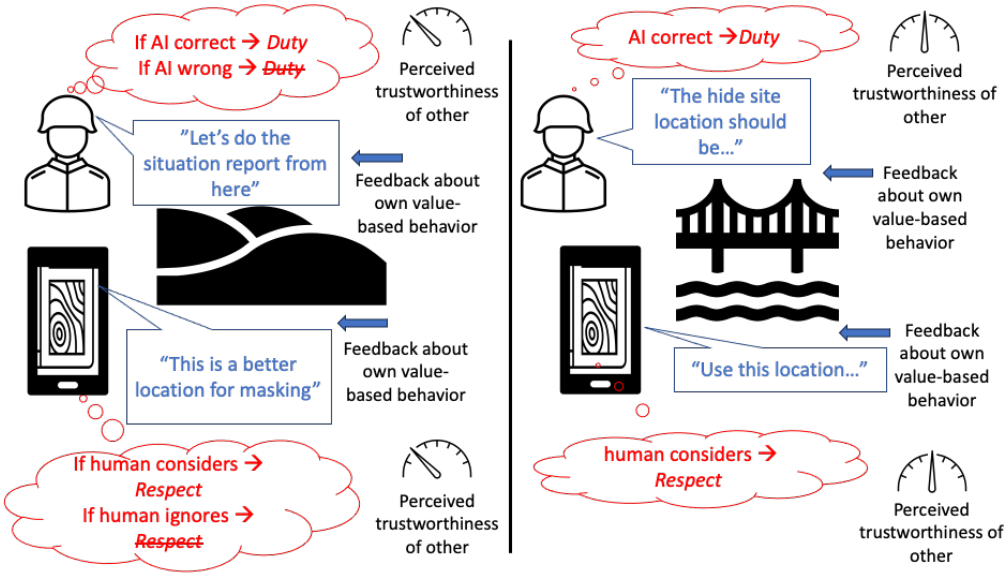

Figure 3. Value-Based Co-Learning About Trustworthiness (Left to Right Over Time)

IMPLICATIONS FOR ENGINEERING HUMAN-AI MUTUAL TRUST

Overall, the intent of this work is to support the development of trustworthiness in human-AI interactions in real-world situations that are existentially consequential (e.g., risk of injury, illness, or death to self and others). In doing so, we recognize that assessing trustworthiness is essential to design, development, and implementation of any capability for such situations, whether it involves AI or not. This is the case whether one is developing a technology (Handley & Knapp, 2014; Zumbado, 2015) or developing human knowledge, skills, and abilities (Agarwal et al.,

2014; Alsalamah & Callinan, 2021). Our focus is on the development of human-AI interaction through an approach to user experience design with deep foundations in behavioral and social sciences (Albert & Tullis, 2013; National Institute of Standards & Technology, 2023). This is a proposition that is fundamentally different from approaches in which human capabilities and AI capabilities are developed independently. Our central principle is that observable context is necessary and sufficient to assess trustworthiness of AI behavior as well as human behavior in similar if not identical ways. This principle connects AI development with both theory and empirical research from an unprecedented diversity of behavioral and social science.

This approach to trustworthiness can benefit, for example, from the pragmatic concept of trustworthiness in qualitative research methods as a philosophical basis for confidence in external expertise (Lincoln & Guba, 1985; Stahl & King, 2020). In qualitative research, trustworthiness requires that content is credible (e.g., fits with other knowledge), transferable (e.g., applicable across situations), confirmable (e.g., auditable range of contributing factors), and dependable (e.g., coherence over time). We note that this view, from a very different perspective, fits nicely with much that trainers and developers would want to see as human-AI teams develop.

Our claim about assessing human-AI capabilities and their development is that it can be accomplished by adapting assessment and instructional methodologies that have been employed for human learning in organizations such as the U.S. Army. As we implied in previous sections, learning occurs on multiple time scales and at multiple levels of abstraction. There is a scaffolding of knowledge, skills, and abilities acquired across the implementation of learning events that are shaped throughout one's career. Kirkpatrick's framework for training evaluation (e.g., Alsalamah & Callinan, 2021) provides connections with an extensive body of research on such scaffolding for values-based behavior because it views learning on four levels—reaction (engagement), learning (knowledge), behavior, and outcomes—that evolve and intertwine over different time scales. The premise of the work reported here is that values are reflected, to varying degrees, in all four of these levels. The extrapolation to the development of human-AI trust is that trust is based on more than just predictability. It requires empirical evidence that the behavior of others is consistent with the values of organizational culture; that is, others must have shown that they are worthy of trust.

We conclude by outlining implications of this theory. While much remains to be achieved with respect to AI development and integration, these implications provide novel guidance regarding what matters, how, and why:

- Values matter for trust of another intelligence (human or AI), and they will be specific to the culture of the organization in which the intelligences are embedded.
    - Intelligence implies choice, choice demands responsibility, responsible choice reflects values.
    - The actions of any intelligence must be consistent with organizational values (e.g., Army Values, Warrior Ethos) to be considered trustworthy.
    - Here we addressed the Army, but other service components, joint and coalition environments will have their own value landscape that will determine whether actions are trustworthy or not.
- Focus on AI-trustworthiness as it develops empirically through experience with other intelligences.
    - What matters is the context and trajectory over time of values-based interactions that are observable and ubiquitous.
    - Before getting "inside the head" of an intelligence, start by asking what information is available and can be obtained about value-based actions and the settings in which they occur.
    - Ask what is needed to capture, measure, track, and leverage this information iteratively over time.
- Build trustworthiness through formative assessment and values-based learning given mission requirements.
    - Given expected mission requirements (e.g., multidomain operations), ask if/when human-AI collaboration is necessary and what attributes of decision-making and action are needed for it.
    - Deliberately shape microexperiences within the context of those mission requirements to enable assessment and growth of trustworthiness by developing instructional events in which actions have value-based meaning and feedback is provided about the implication of those actions.

ACKNOWLEDGEMENTS

In developing the concepts presented here, we gratefully acknowledge the insights of our many colleagues associated with OBTE, ASLTE, and more generally, methods to deliberately assess and grow attributes. While the present work focuses on other intelligences, we owe a debt to prior research related to Soldier development.

REFERENCES

Agarwal, N., Pande, N., & Ahuja, V. (2014). Expanding the Kirkpatrick evaluation model-towards more efficient training in the IT sector. *International Journal of Human Capital and Information Technology Professionals (IJHCITP)*, *5*, 19-34.

Albert, B., & Tullis, T. (2013). *Measuring the user experience: collecting, analyzing, and presenting usability metrics*. Amsterdam: Elsevier.

Alsalamah, A., & Callinan, C. (2021). The Kirkpatrick model for training evaluation: bibliometric analysis after 60 years (1959–2020). *Industrial and Commercial Training*, *54*, 36-63.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, *58*, 82-115.

Balliet, D., & Van Lange, P. A. M. (2013). Trust, conflict, and cooperation: A meta-analysis. *Psychological Bulletin, 139*, 1090–1112.

Belle, V., & Papantonis, I. (2021). Principles and practice of explainable machine learning. *Frontiers in big Data*, 39.

Braitenberg, V. (1984). *Vehicles: Experiments in synthetic psychology*. Cambridge, MA: MIT Press.

Branson, R. K., Rayner, G. T., Cox, J. L., Furman, J. P., King, F. J., & Hannum, W. H. (1975). *Interservice procedures for instructional systems development: Executive summary, Phases I, II, III, IV, V*. Ft. Benning, GA: US Army Combat Arms Training Board. (DTIC No. A019486).

Brunyé, T., Riccio, G., Sidman, J., Darowski, A., & Diedrich, F. J. (2006). Enhancing Warrior Ethos in initial entry training. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, San Francisco, CA.

Colquitt, J. A., Lepine, J. A., Piccolo, R. F., Zapata, C. P., & Rich, B. L. (2012). Explaining the justice-performance relationship: Trust as exchange deepener or trust as uncertainty reducer? *Journal of Applied Psychology, 97*, 1-15.

Dein, J. P., Ingurgio, V., Ratwani, K. L., Diedrich, F., & Flanagan, S. (2019). Tools and measures for NCO talent assessment. *NCO Journal*, July 2019.

De Jong, B. A., Dirks, K. T., & Gillespie, N. (2016). Trust and team performance: A meta-analysis of main effects, moderators, and covariates. *Journal of Applied Psychology, 101*, 1134–1150.

Dewey, J. (1938). *Experience and education*. New York: Simon and Shuster.

Flanagan, S., Horn, Z., Knott, C., Diedrich, F., Halverson, K., Lucia, L., & Weil, S. (2015). Teaching social interaction skills with stealthy training techniques. 6th International Conference on Applied Human Factors and Ergonomics. *Procedia Manufacturing, 3*, 4036 - 4043.

Fowler, G.A. (2023, April 3). We tested a new ChatGPT-detector for teachers. It flagged an innocent student. *The Washington Post*. https://www.washingtonpost.com/technology/2023/04/01/chatgpt-cheating-detection-turnitin/

Gagne, R. M., Wager, W. W., Golas, K. C., Keller, J. M., & Russell, J. D. (2005). Principles of instructional design. *Performance Improvement*, *44*(2), 44-46.

Guion, R. M. (2011). *Assessment, measurement, and prediction for personnel decisions*. New York, NY: Routledge.

Hancock, P. A., Kessler, T. T., Kaplan, A. D., Stowers, K., Brill, J. C., Billings, D. R., Schaefer, K. E., & Szalma, J. L. (2023). How and why humans trust: A meta-analysis and elaborated model. *Frontiers in Psychology,* 14:1081086. doi: 10.3389/fpsyg.2023.1081086

Handley, H. A., & Knapp, B. G. (2014). *Where are the people? The human viewpoint approach for architecting and acquisition*. Defense Acquisition University: Ft. Belvoir, VA.

Kim, P. H., Dirks, K. T., Cooper, C. D., & Ferrin, D. L. (2006). When more blame is better than less: The implications of internal vs. external attributions for the repair of trust after a competence- vs. integrity-based trust violation. *Organizational Behavior and Human Decision Processes, 99*, 49-65.

Kraft, A. (2016, March 25). Microsoft shuts down AI chatbot after it turned into a Nazi. *CBS News*. https://www.cbsnews.com/news/microsoft-shuts-down-ai-chatbot-after-it-turned-into-racist-nazi/

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Newbury Park, CA: Sage Publications.

Mace, W. M. (1977) James Gibson's strategy for perceiving: Ask not what's inside your head, but what your head's inside of. In R. Shaw and J. Bransford (eds.), *Perceiving, Acting and Knowing*. Erlbaum, Hillsdale, NJ.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*, 709–734.

National Institute of Standards & Technology (2023). *Artificial intelligence risk management framework* (AI RMF 1.0). U.S. Department of Commerce, https://doi.org/10.6028/NIST.AI.100-1

Park, E., Jenkins, Q., & Jiang, X. (2008 September). *Measuring trust of human operators in new generation rescue robots.* Paper presented at the 7th JFPS International Symposium on Fluid Power, Toyama, Japan.

Riccio, G., Diedrich, F., & Cortes, M. (2010). *An initiative in Outcomes-Based Training and Education*. Technical Report, U.S. Army Asymmetric Warfare Group, Ft. Meade, MD.

Sebo, S. S., Krishnamurthi P., & Scassellati, B. (2019*). "I don't believe you": Investigating the effects of robot trust violation and repair.* Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Daegu, South Korea.

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology, 47*, 149-155.

Sousa, S., Cravino, J., Martins, P., & Lamas, D. (2023). Human-centered trust framework: An HCI perspective. *arXiv preprint arXiv:2305.03306*.

Stahl, N. A., & King, J. R. (2020). Expanding approaches for research: Understanding and using trustworthiness in qualitative research. *Journal of Developmental Education, 44*(1), 26-28.

Toumbeva, T. H., Diedrich, F. J., Flanagan, S. M., Naber, A., Reynolds, K., Shenberger-Trujillo, J., Cummings, C., Ratwani, K. L., Ubillus, G., Nocker, C., Gerard, C. M., Uhl, E. R., & Tucker, J. S. (2019). *Assessing character in U.S. Army Initial Entry Training* (ARI Technical Report 1373). Fort Belvoir, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (DTIC No. AD1077839)

Toumbeva, T. H., Uhl, E. R., Wittig, A. H., Sanders, C. N., Diedrich, F. J., Flanagan, S. M, & Koschny, R. L. (2021). *Development and evaluation of a revised peer assessment for the U.S. Army Officer Candidate School* (ARI Technical Report 1390). Fort Belvoir, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (DTIC No. AD1126519)

Uhl, E. R., Toumbeva, T. H., Wittig, A. H., Diedrich, F., Flanagan, S., Koschny, R. L., & Kirshenbaum, J. M. (2022). Predictors of peer assessment in junior leader training. *Military Psychology, 34*, 1-11.

U.S. Army Asymmetric Warfare Group (2013). *Implementing the Army Learning Model*, unpublished workbook.

U.S. Department of the Army (2018). *The U.S. Army in multi-domain operations 2028* (TRADOC PAM 525-3-1). Washington, DC.

U.S. Department of the Army (2019a). *Army leadership and the profession* (ADP 6-22). Washington, DC.

U.S. Department of the Army (2019b). *Mission command* (ADP 6-0). Washington, DC.

U.S. Department of the Army (2022). *Operations* (FM 3-0). Washington, DC.

van Dijk, L. (2021). Psychology in an indeterminate world. *Perspectives on Psychological Science, 16*, 577-589.

Verma, P., & Oremus, W. (2023, April 5). ChatGPT invented a sexual harassment scandal and named a real law prof as the accused. *The Washington Post*. https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/

Zumbado, J. R. (2015). *Human Systems Integration (HSI) Practitioner's Guide* (No. JSC-CN-34987).