ARL-TR-10403 ● AUG 2026

DEVCOM ARMY RESEARCH LABORATORY

Soldier–AI Integration: AI Trust and Teaming Metrics

by Kimberly Pollard, Shan G. Lakhmani,
Murat Kucukosmanoglu, Cheryl Giammanco,
Samantha K. Berg, and Andrea Krausman

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-10403 ● AUG 2026

Soldier–AI Integration: AI Trust and Teaming Metrics

Kimberly Pollard, Shan G. Lakhmani, Cheryl Giammanco, and Andrea Krausman
DEVCOM Army Research Laboratory

Murat Kucukosmanoglu
D-Prime LLC

Samantha K. Berg
Oak Ridge Associated Universities

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
		START DATE	END DATE
August 2026	Technical Report	April 2025	September 2025

4. TITLE AND SUBTITLE
Soldier–AI Integration: AI Trust and Teaming Metrics

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)
Kimberly Pollard, Shan G. Lakhmani, Murat Kucukosmanoglu, Cheryl Giammanco, Samantha K. Berg, and Andrea Krausman

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
DEVCOM Army Research Laboratory
ATTN: FCDD-RLA-FA
Adelphi, MD 20783

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TR-10403

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

13. SUPPLEMENTARY NOTES
ORCIDs: Kimberly Pollard, 000-0002-5849-1987; Shan G. Lakhmani, 0000-0001-6052-439X; Murat Kucukosmanoglu, 0000-0003-3611-7692; Cheryl Giammanco, 0000-0001-6465-4036; Samantha K. Berg, 0000-0001-5774-8747; Andrea Krausman, 0000-0003-1955-8867

14. ABSTRACT
In support of the Army Command and Control Modernization Priority, ARL is developing automated and AI-enabled technologies, including large language models and adaptive machine learning algorithms to enable faster and more informed decision-making across echelons. Soldiers work in teams with other humans and with automated systems and intelligent agents. It is critical, therefore, to understand how different AI-enabled systems affect team processes and team and user states, such as trust, cohesion, workload, and team adaptation. Advanced AI systems differ in meaningful ways from previous forms of autonomy, raising the question of whether existing methods of measuring team processes and user states are sufficient for use with these new AI-enabled teams. To address this question, we performed a non-exhaustive literature review to uncover what metrics are successfully being used to measure team processes and user states in AI-enabled teams. We found similar methods being used in the literature as have been used with less advanced autonomy, as well as some opportunities to effectively harness human–human measures of team states and processes with AI's new capabilities.

15. SUBJECT TERMS
Humans in Complex Systems, human–machine teaming, human–AI teaming, trust metrics, autonomous systems

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UU	36

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
Kimberly Pollard	(520) 691-6154

STANDARD FORM 298 (REV. 5/2020)
Prescribed by ANSI Std. Z39.18

Contents

List of Tables

1. Introduction

The Army's investment in AI-enabled systems to support future operations stems from the recognition that the speed and complexity of the operational environment is changing rapidly, and with the integration of myriad sensors and advanced technologies on the battlefield, data is produced at a rate that humans alone are unable to process. This creates difficulty in maintaining situation awareness (SA) and making timely, informed decisions. AI, however, possesses capabilities that, when integrated properly, can augment Soldier capabilities and enhance individual and unit performance. For example, AI can process vast amounts of data and identify patterns faster than humans, enabling quicker and better-informed decisions. AI can also reduce Soldier burden by automating repetitive tasks, analyzing complex data, and providing decision support, which frees up Soldiers to focus on more critical tasks that require human judgment, expertise, and creativity. AI can also increase SA of the battlefield due to the ease at which it can fuse data from multiple sources and provide rapid updates to create a more comprehensive and accurate picture of the battlefield. Taken together, these capabilities show promise for enhancing Soldier and unit performance on the future battlefield.

Recent advancements in AI and the proliferation of large language models (LLMs), generative and agentic AI, and machine learning (ML) approaches have created renewed interest in integrating these technologies into military command and control (C2) processes. This is due to the realization that today's C2 processes, although effective, largely stem from "an industrial era" approach, are often linear and time-consuming, and do not scale well to the complexity of future conflicts. The introduction of AI into the Army's C2 process, across echelons, represents a fundamental shift in how the Army operates. This requires a comprehensive reevaluation of existing practices and a deliberate focus on ensuring effective technology integration with Soldiers. In particular, the language capability facilitated by LLMs may allow AI to be better enmeshed into teams and facilitate contributions to team processes and emergent team states. Consequently, the Army is interested in assessing how AI contributes to relevant team processes (e.g., coordination and communication) and emergent states (e.g., shared mental models, team situation awareness, and cohesion).

When integrating Soldiers and AI, the most effective approach is not to view AI as a replacement for humans, but as a complement. Collaborative Soldier–AI teams are the key. In such teams, AI augments human capabilities and allows Soldiers to focus on tasks requiring uniquely human skills. Understanding how to build collaborative Soldier–AI teams that leverage the strengths of both humans and AI

is a critical question to address so that we can develop better assistive AI and integrate it into future AI-enabled teams.

In the early 2020s, the release of capable LLMs and generative AI models, including interactive public formats such as ChatGPT, heralded a veritable explosion of interest and innovation in these and similar models. After just a couple years, LLMs and other forms of generative AI are already starting to feature prominently in consumer technology and are virtually inescapable. This led some users and communities to enthusiastically embrace the profound transformation of technology while others raise ethical concerns or express lack of trust or lack of interest in the new technology. In addition, this integration also poses challenges for understanding and assessing user–AI interactions and the impact on performance and team states such as trust, cohesion, and workload.

The ubiquity of modern AI systems means that Soldiers will increasingly arrive on the battlefield with previous AI experience, preconceived notions about its usefulness and trustworthiness, and existing expectations about its capabilities (Kaplan et al. 2023). Much of this variation in trust and perceived usefulness will depend on the specific AI in question, on the use case and context, and on the user's own personal characteristics (Kaplan et al. 2023; Ng and Zhang 2025; Holder et al. 2025). What remains across all these situations is that the latest forms of AI are arguably *categorically different* than earlier forms of autonomous systems (Chi et al 2021; Kaplan et al. 2023; Ng and Zhang 2025), both in their capabilities and their ubiquity. This raises the need to revisit our understanding of trust development, maintenance, and measurement in human–autonomy teams (Krausman et al. 2022) and revisit our approach to other human–autonomy team processes. We thus seek to uncover whether principles and measures applicable to earlier or simpler forms of autonomy need to be supplemented or replaced with new principles and measures when considering modern advanced AI (including LLMs, multimodal models, and other generative AI).

Modern AI systems act, think, interact, and communicate differently than human teammates do and also differently than earlier or simpler automated systems (Kaplan et al. 2023). For example, an LLM could rapidly generate a natural language summary from a commander's orders and generate a map of important battlefield objects. This could be faster and less burdensome than a human-only approach. However, hallucination is mathematically inevitable in computable LLMs (Xu et al. 2024; Cossio 2025), and the mistakes that LLMs make are different than the mistakes that humans tend to make (e.g., Files and Pollard 2025). LLM mistakes also often come with levels of projected confidence that do not convey accuracy or uncertainty in the same way that indicators of human confidence do, creating difficulties in interactions between humans and AI as humans work

alongside new machines that speak human-like language, use human-like indicators of agreement and confidence, do not operate in a strictly repeatable way, and make unexpected, if not downright odd, mistakes. The differences between modern (and future) AI and previous autonomous systems are substantial (Chi et al. 2021; Kaplan et al. 2023; Ng and Zhang 2025). Therefore, we need to reassess how trust in human–agent teams develops, evolves, and is calibrated with these new, categorically different agents, as compared to previous forms of autonomy/ automated systems.

Further, modern AI-enabled systems, such as LLMs or social robotics, by using more "human" methods of communication and responses to input, might be expected to yield interactions that are more like human–human interactions, such as linguistic entrainment with LLMs leading to greater self-disclosure in what a human might say to the AI (Kian et al. 2025). More "human-like" behavior in AI-enabled systems might be expected to induce trust or teaming patterns that resemble something in between patterns seen for human–human relationships and human-(non-AI)–machine relationships (Chi et al. 2021). However, the behavior and properties of different types of AI may also lead to largely new patterns of trust and team processes (Huang et al. 2025; Ng and Zhang 2025), and novel metrics may be needed.

To address the need for appropriate metrics, we performed a non-exhaustive literature review to examine what metrics are available, recommended, or used for measuring human–AI team states and processes. For the purposes of this report, we consider the following to count as "AI": LLMs, other generative AI, ML-based systems/classifiers/recommenders, any system that learns (i.e., reweights its model) from human interaction or data updates, and any ML or similar system that interactively controls or is onboard physical agents (e.g., UAS, ground robots). See Kaplan et al. (2023) for a more restricted definition. Studies are also included that explore simulated AI systems, where the AI capabilities are mimicked via other interactions, including "Wizard of Oz" methods. In Section 2, we call out key categories of metrics and team processes found in the literature.

2. Measuring Team Processes and User States in AI-Enabled Teams

As previously discussed, the advancement and availability of AI capabilities, such as LLMs, generative AI, and ML models, presents opportunities for reducing the task burden on Soldiers, increasing data processing and SA, and improving combat effectiveness. Essentially, these increasing capabilities can allow AI/ML to be better enmeshed into teams as collaborative partners and to improve contributions

to team processes and emergent team states that enhance performance. Consequently, the Army is interested in how to effectively integrate AI into Soldier teams to realize these benefits. A critical component of successful integration is understanding and assessing how relevant team processes and emergent states (such as coordination, situation awareness, and trust) are affected by interacting with AI systems. Current assessment methods rely heavily on observations or subjective data, and while this information is valuable, they only provide point estimates of critical events and fail to capture the continuous flow and evolution of team dynamics. What is needed is a more robust approach that enables monitoring and assessment of the unique aspects of Soldier–AI teams over time. This work builds on a framework (Krausman et al. 2022) and software toolkit (Neubauer et al. 2023) specifically targeting development of suitable methods for continuous assessment within the context of Human Machine Integrated Formations, in support of the Human–Autonomy Teaming Essential Research Program (HAT ERP). To date, the framework and toolkit provide a strategy for continuous multimodal assessment of team processes and states using subjective, communication, behavioral, and physiological measures to capture those changes in team states that occur over time. The purpose of this literature review is to expand the scope to include the unique aspects of AI and determine which of the metrics carry forward to the context of Soldier–AI teams.

2.1 Subjective Methods of Measuring Trust

A variety of questionnaires have been developed for assessing user states and team processes when working with earlier forms of technology, such as computers, artificial agents, robots, and other automated systems. These methods use scales, multiple choice, and free response questions to elicit self-reported levels of trust, cohesion, and other user states. Many such questionnaires exist, including well-cited and well-validated options. Questionnaires can be administered before, during, and/or after interaction with the AI system of interest. Benefits of questionnaires include ease of administration, typically straightforward scoring, and the ready availability of published questionnaires to fit the circumstance. Major drawbacks include the inability to achieve continuous monitoring of human states, which means that changes in trust or other states over time cannot be detected at fine resolution. In addition, self-report measures rely on the accuracy of the respondent, who may not have full conscious awareness of their state or who may not report their state honestly for a variety of reasons (Mehrotra et al. 2024). Below, we list a few popular measures. For extensive catalogs of available subjective measures, refer to Appendix D of Ueno et al.'s paper (2022) and Table 2 in Ng and Zhang's paper (2025).

According to a thorough review by Ueno et al. (2022), the Jian et al. (2000) 12-item Trust in Automation (TIA) scale, called the Checklist for Trust between People and Automation, is the most widely used questionnaire for measuring trust in human–AI interaction. The scale measures individuals' trait-dependent, dispositional trust of automation. The scale was developed by Jian et al. (2000) to assess individuals' "non-directed feelings of trust in automated systems," rather than their trust in a specific system based on experience (i.e., not system specific). The TIA is a 12-item self-report measure in which individuals rate the intensity of their feeling of trust or impression of the system using a Likert-type scale, with 1 = not at all, and 7 = extremely. Despite its widespread use, the TIA scale has not yet been formally assessed for its applicability for human–AI trust (Ueno et al. 2022).

Another popular scale is the Trust in Teams Scale (TTS) (Costa 2003; Costa and Anderson 2011), which measures four distinct dimensions of trust: propensity to trust, perceived trustworthiness, cooperative and monitoring behaviors (Costa 2003; Costa and Anderson 2011). The propensity to trust and perceived trustworthiness are formative indicators, whereas cooperative and monitoring behaviors are reflective indicators—the behavioral consequences of trust. Costa and Anderson (2011) found that perceived trustworthiness and cooperative behaviors tend to correlate strongly with each other. Team members that perceive their teammates as trustworthy exhibit fewer monitoring behaviors and more cooperative behaviors, as compared to members of teams that perceive their teammates as less trustworthy. The TTS is a 21-item scale in which the individuals report their perceptions of trust in the other team members and the team, with responses ranging from 1 = completely disagree, and 7 = completely agree. There are six items for each of three subscales (propensity to trust, perceived trustworthiness, and cooperative behaviors), and three items for the final subscale, monitoring behaviors.

Another commonly used scale is the System Trustworthiness Scale (STS) (Muir and Moray 1996). The STS is a task-dependent state measure of trust calibration based on the predictors of individuals' developing trust, such as the system's predictability, dependability, reliability over time (faith), and competence (Muir and Moray 1996). The STS is system agnostic, referring to "the system" in general. The STS is a 9-item self-report measure in which individuals rate the system's competence, predictability, dependability, responsibility, and reliability over time (faith) using a Likert-type sliding scale scored on a continuum from 0 (very low) to 100 (very high) in 5-point increments (with 1 = none at all, and 100 = extremely high).

U.S. Army Combat Capabilities Development Command (DEVCOM) Army Research Laboratory (ARL) studies have made effective use of scales developed in-house. One such scale is the Trust Perception Scale–Human–Robot Interaction (HRI) (Schaefer 2016), which, for example, was used by ARL in studies of humans teaming with robots that had simulated AI properties (e.g., Lukin et al. 2018). The Trust Perception Scale-HRI consists of 40 questions to assess a user's subjective trust in their robot or AI teammate.

In another effort, ARL investigated the influence of AI transparency on trust calibration, AI adoption, and decision-making (Giammanco, in review). In the study, AI transparency was conveyed through several methods, including application of a graph portraying the AI agents' rationale for its recommended military course of action (COA). The researchers developed measures of the participants' perceived confidence in their selected COA and the AI agents' influence on their decisions (an indirect measure of AI adoption). The Decision Confidence Scale (Giammanco 2021), developed by ARL, is based on Ackerman's (Ackerman 2019; Ackerman and Beller 2017; Ackerman and Thompson 2017) research on problem solving and decision-making processes. Ackerman (2014) found that the participants' final confidence (i.e., subjective probability that your response to a problem is correct) tended to be higher than their intermediate confidence, regardless of the accuracy of their answers. To assess trust calibration, the researchers applied the TTS (Costa 2003). AI adoption was assessed through the participants' use of the AI agents' rationale graph and acceptance of the AI agents' recommended COA. In addition to objective measures of AI adoption, the researchers administered the TTS (Costa 2003) and STS (Muir and Moray 1996) to assess the effect of AI transparency on trust calibration and AI adoption of a novel AI-driven recommender system.

2.2 Behavioral Methods to Measure Trust

A human's active, deliberate behaviors can also be examined as indicators of trust and team states. A major benefit of behavioral methods for measuring team states during human–AI interaction is that these metrics are objective. The human either presses the button or they do not; they either follow the AI's suggestion or they do not. For many experimental paradigms, these sorts of behaviors are easy to observe, score, and log automatically. A significant drawback of behavioral measures is that they are only proxies of underlying subjective states, and behavior is driven by multiple internal states. For example, a person who relies more on an AI's recommendations could reasonably be thought to be more trusting of that AI (Okamura and Yamada 2020a). However, their reliance could also be motivated by fatigue, cognitive overload, or other factors (Okamura and Yamada 2020b;

Mehrotra et al. 2024). In general terms, the rate at which a human agrees with or accepts an AI's recommendation (i.e., agreement percentage) is a common behavioral measure, as is switch percentage (how often a user changes from their plan or answer to the AI's plan or answer).

Behavioral methods can be measured near-continuously, and equations have been developed for tracking human states across near-continuous behavioral events (e.g., Okamura and Yamada 2020b). Behavioral measures tend to be examples of *demonstrated* trust, which is distinct from the *perceived* trust that is targeted with subjective assessment methods (Mehrotra et al. 2024).

Authors refer to behavioral metrics by different terms, including reliance, compliance, dependence, and others (Chansey et al. 2017; Holland et al. 2024), which focus on the nuanced ways in which a human can agree or disagree, encourage or discourage, when engaging with an AI teammate. The common theme of these measures is that they document how the human actually behaves when interacting with an AI-enabled system and uses objective acts of demonstrated trust (whether compliance, reliance, or other) as a proxy measure of the human's underlying subjective levels of trust. In task domains where human–AI interaction is ongoing, behavioral methods of trust measurement afford the opportunity to assess trust calibration (Okamura and Yamada 2020a) in a more continuous fashion as compared to survey-based methods.

The precise behavioral indicators of trust will vary depending on the nature of the AI system, task, and type of human–AI interaction under study. However, we may expect to frequently see a task domain or interaction system wherein the AI is doing one or more of the following: making recommendations, offering suggestions, presenting rationale, automating task elements, or being encouraged/discouraged in a reinforcement learning paradigm. In any of these circumstances, the human's behavior can be observed and used as an indicator of their (demonstrated) trust in the AI system. Objective behavioral data can thus be collected in terms of how often the human intervenes in an otherwise automated process, when they intervene, what level of automation they choose, when and how often the human accepts versus rejects the AI's recommendations, when or how often the human alters the recommendations, or when or how often they request new or modified recommendations. Similarly, one can measure how often the human accepts or rejects an AI agent's planning assumptions or its offered rationale for its decisions or suggested COA. For systems where the user provides reinforcement training for the AI system, how often the user provides negative or positive reinforcement can be measured.

2.3 Physiological Measures of Trust

Trust has wide-ranging consequences, influencing technology adoption, willingness to delegate control, openness to risk, cooperation, innovation, investment, and ultimately user performance. In contrast, a lack of trust often results in poor user experiences, resistance to adoption, loss of investment, and degraded performance outcomes.

Physiological measures offer a powerful means of capturing continuous and objective indicators of Soldier and team states during interaction with AI-enabled systems. Unlike self-report questionnaires, which are retrospective and subjective, or behavioral logs, which capture only overt actions, physiological signals can uncover underlying cognitive, emotional, and physiological processes as they occur in real time. As such, they provide an essential complement to subjective and behavioral methods for assessing trust, cohesion, workload, and other emergent team states.

Measuring trust through physiological signals requires identifying physiological response patterns associated with different psychological states during the human–machine interaction process. They capture biological responses using electrocardiography (ECG), electroencephalogram (EEG), gaze tracking and skin conductance response (Akash et al. 2018; de Visser et al. 2018; Lu and Sarter 2020; Ayoub et al. 2023). Some of these responses, such as eye movements, can also be considered a form of behavior, but here we discuss them as physiological indicators. Given the complexity of this mapping of the various signals, researchers are encouraged to adopt multipronged strategies, including:

- adapting trust measurement scales to the context,
- integrating multiple physiological modalities,
- distinguishing between different types of trust relationships, and
- employing sophisticated signal analysis and ML techniques.

Common modalities include:

- Metrics such as fixation duration, saccade frequency, blink rate, and pupil diameter are established indicators of attention allocation, stress, and cognitive workload (Marshall 2007; Tsai et al. 2007; Di Stasi et al. 2013). In Soldier–AI teams, eye tracking can be used to evaluate whether operators monitor AI systems, attend to AI recommendations, or divide attention across human and machine teammates. Emerging research also links eye movement patterns directly to trust calibration. For example, Lau et al. (2024) found that larger mean pupil size was significantly associated with

lower trustworthiness ratings, while higher saccade counts corresponded to higher perceived trust toward media presenters. Similarly, Ayoub et al. (2023) demonstrated that physiological signals, including eye-tracking features such as blink rate and pupil diameter, alongside heart rate variability (HRV), can be used for real-time prediction of trust in conditionally automated driving systems, achieving classification accuracies above 80%. Their findings indicate that increased fixation activity is associated with lower trust. Taken together, these results suggest that eye movements may serve as real-time physiological markers of trust dynamics (Ayoub et al. 2023; Lau et al. 2024).

- ECG and HRV provide dynamic and sensitive markers of trust. Prior work showed that team-level heart rate synchrony is associated with higher trust toward AI agents during collaborative tasks (Mitkidis et al. 2015). More recently, Mosaferchi et al. (2024) reported that heart rate responses were negatively correlated with self-reported trust, particularly in the domains of competence and reliability, underscoring the potential of heart rate as a physiological proxy for trust.

Physiological measures bring the benefits of objective signals and continuous monitoring capability. Drawbacks are that specialized equipment is required, as is expertise in equipment setup and signal processing.

2.4 Communication Measures of Trust

Another objectively measurable and near-continuous source of data is communication. Here, we focus on verbal communication, though other modalities can also be measured. Communication is a central driver of coordination, trust, and cohesion in teams. For Soldier–AI teams in particular, communication occurs not only among human teammates but also between humans and AI systems, especially language-enabled systems such as LLMs. Because of this, communication data offers a unique and rich source for assessing team processes and emergent states in AI-enabled teams. Prior work shows that linguistic style matching, or the extent to which speakers align in their word choice and phrasing, can serve as a proxy for trust, cohesion, and rapport in teams (Carmody 2017). More recently, research in human–AI teaming contexts suggests that linguistic analysis of communication, particularly word choice, sentiment, and language patterns, emerges as a promising indicator of trust dynamics, though the field is still underdeveloped (Nguyen et al. 2024).

A first set of methods involves verbal communication metrics, including sentiment and linguistic indicators. Sentiment analysis models can classify the emotional tone

of human or AI utterances, capturing positive, negative, or neutral affect expressed in team dialogue (Kucukosmanoglu et al. 2025). Lexicon-based tools such as Linguistic Inquiry and Word Count (LIWC) or Empath can further reveal word-choice indicators of affect, cognition, or social orientation. For AI-enabled systems that use human language as the input interface, increased complexity in the human's instructions may also be an indicator of trust (Lukin et al. 2018). Beyond sentiment, linguistic style matching and entrainment measures assess how closely team members align in word use, syntax, or phrasing, providing evidence of shared mental models and cohesion. These measures can be applied not only to human–human dialogue but also to human–AI exchanges, offering a window into whether Soldiers treat AI agents as trusted teammates. Studies show that linguistic style matching, the degree of similarity in function-word usage across speakers, predicts group cohesiveness and task performance in team settings (Gonzales et al. 2010), as well as predicting rapport and cooperation (Vail et al. 2022).

Entrainment can include matching between interlocutors in virtually any aspect of communication, including not only syntax or phrasing choices but also semantic content, facial expressions, body postures, rate and timing parameters, and, for spoken language, acoustic properties (Wynn and Borrie 2022; Dideriksen et al. 2023; Kian et al. 2025). A recent study examining linguistic entrainment between humans and LLMs in a digital therapy context found improved team performance (higher levels of intimacy and engagement, in this context) when linguistic entrainment was greater (Kian et al. 2025).

A second set of methods involves structural and network-level communication metrics. These approaches capture patterns of interaction such as frequency of exchanges, turn-taking dynamics, and centrality of individuals or agents in the communication network. For example, examining whether Soldiers defer to an AI agent's suggestions, or how often the AI is positioned as a communication hub, can provide indirect indicators of reliance and trust. Prior work demonstrated that network-level communication measures provide valuable insights into dynamics such as trust, coordination, and performance in human–autonomy teams (Baker et al. 2021).

The integration of AI-enabled communication also enables novel metrics unique to human–AI teaming. Unlike traditional automated systems, LLMs produce natural language outputs that can themselves be analyzed for coherence, fluency, adaptiveness, or responsiveness to human input. Assessing the quality of AI-generated communication, alongside human responses to it, offers opportunities to measure mutual understanding, trust calibration, and the extent to which AI contributions are integrated into team discourse (de Visser et al. 2023; Behzad et al. 2025). Recent research in healthcare conversational agents identifies quality,

fluency, relevance, empathy, and personalization as key user-centered evaluation metrics for LLM-based systems, which directly speak to qualities such as coherence, adaptability, and human integration (Abbasian et al. 2025).

Strengths of communication-based measures include their capacity to capture continuous, objective data that reflects both the content (what is said) and the process (how it is said). However, these measures also carry limitations, including sensitivity to task context, dependence on robust natural language processing tools, and the risk that surface-level linguistic features may not fully capture latent states such as trust or cohesion. Even with these challenges, communication remains a critical lens for examining Soldier–AI teaming, particularly given the inherently language-driven nature of many modern AI systems (Baker et al. 2021; McNeese et al. 2021).

2.5 Trust Calibration

Effective human–AI teamwork relies on appropriate levels of trust—too little, and the human will not fully utilize the AI's capabilities—too much, and the human may fail to adequately monitor the AI's actions (Okamura and Yamada 2020b). This is called trust calibration—the process of ensuring the level of trust within a team corresponds to the actual capabilities present for a given situation (Muir 1987). It is crucial because both excessive and insufficient trust can negatively impact team performance. Trust calibration is particularly complex and vital within human–AI teams, as it involves establishing appropriate trust both between humans and the AI system and among the human team members regarding the AI's contributions (Krausman et al. 2022). Several challenges are unique to these teams, stemming from differing cognitive styles; humans rely on intuition while AI operates algorithmically—with an often-opaque reasoning (Johnson and Vera 2019). This opacity hinders trust development, even with explainable AI (XAI), and is compounded by varying levels of technical expertise and preexisting biases among team members (Mercado et al. 2016). Crucially, trust is not static; it fluctuates with performance, task complexity, and perceived risk, necessitating continuous recalibration. The metrics presented in this paper are an important aspect of trust calibration, and effectively measuring trust over time is critical for predicting inappropriate reliance and enabling interventions to promote appropriate calibration (deVisser et al. 2019). Understanding when to measure trust and interpreting the results considering mission parameters and team dynamics is complex, but essential for effective team function.

2.6 Measuring Team Adaptation

One of the major precursors to team effectiveness is team adaptation. Several team processes and emergent states are involved in the team adaptive cycle, and thus effective assessment of these processes and states is paramount when determining the effectiveness of team adaptation, especially when a team is coupled with AI. The relevant processes and emergent states for understanding team adaptation include shared mental models, team SA, and team communication and coordination, which are discussed below.

2.6.1 Assessing Shared Mental Models

Shared mental models, organized knowledge shared by team members (Cannon-Bowers et al. 1993), has been identified as a major factor positively affecting team performance (Johnson et al. 2007). Shared mental models also undergird the work done during the cycle of team adaptation, wherein a team changes its performance in response to a salient cue to reach a functional outcome (Burke et al. 2006). The exploration and assessment of shared mental models in human–AI teams is limited, but ongoing, and is currently informed by existing methods (Andrews et al. 2023). Knowledge elicitation is one of the more common methods of measuring shared mental models (e.g., surveys and questionnaires, interviews, observation, similarity ratings, concept mapping, and card sorting), and it entails soliciting the content or components of team members' mental models and determining the extent to which that information is shared (Mohammed et al. 2000; Johnson et al. 2007; DeChurch and Mesmer-Magnus 2010). Structure representation techniques (e.g., card sorting, concept mapping, paired sentence comparison, network analysis algorithms [such as Pathfinder and UCINET], and multidimensional scaling) focus on capturing the structure or organization of the knowledge in mental models—though some of these techniques can also include the content, which can be used to establish similarity in concept structure among teammates (Mohammed et al. 2000; DeChurch and Mesmer-Magnus 2010). Schelble et al. (2022a) used paired sentence comparison and the Pathfinder network scaling algorithm to determine content and structure of mental models in human–human–human and human–human–agent teams, suggesting that the effect of the inclusion of an AI on a team's shared mental model can be measured in a multi-human human–AI team. An AI's mental model, depending on its individual construction, may be measurable, but the extent to which that model is shared with human teammates during joint activity, and the extent to which that overlap can be measured, is still being pursued.

2.6.2 Measuring Team Situation Awareness

Team SA is another emergent state that feeds into and emerges from the cycle of team adaptation (Burke et al. 2006). One definition of team situation awareness is the degree to which every team member possesses the SA—the perception of elements in the environment within time and space, the comprehension of their meaning, and the projection of future states (Endsley 1988)—needed to complete their responsibilities (McNeese et al. 2021; Endsley 2023). Individuals within a team each have their own awareness of the situation, and one approach to measuring a team's SA is to aggregate the assessments of each team member's individual SA—using their scores on measures such as the Situation Awareness Global Assessment Technique (SAGAT), the Situation Present Assessment Method (SPAM), or other individual SA measures (Cooke et al. 2004). An alternative approach to team SA is oriented not around shared knowledge, but around team interaction; essentially, team SA emerges from the interactions that occur as teams assess a situation, relay relevant information to relevant actors, and take action (Cooke et al. 2013; McNeese et al. 2021). One means of measuring team SA, using the interaction approach, is by identifying team coordination—a team process critical to the plan execution phase of team adaptation (Burke et al. 2006), whereby behavioral mechanisms are enacted to perform a task and transform team resources into outcomes (Salas et al. 2015)—in the face of "roadblocks," a change in the environment that pushes the team's task into an unfamiliar region of its state space that has performance consequences for the team (Gorman et al. 2005). McNeese et al. (2021) used this team interaction-based measure of SA, the Coordinated Awareness of Situation by Teams (CAST), to compare the team SA of an all-human team, a team with a synthetic pilot, and a team with a "synthetic teammate" that was in fact a human confederate. They found that the human team's SA remained stable, while the team SA of the team with a synthetic pilot improved over time, and the team SA of the team "Wizard of Oz" pilot improved and then stabilized over time (McNeese et al. 2021). As such, behavioral measures of team SA, such as CAST, may be useful for AI that uses sensors to identify changes in the environment and can convey that information to human teammates.

2.6.3 Measuring Team Communication and Coordination

Communication plays a pivotal role in team adaptation, specifically in the development and updating of shared knowledge structures (Burke et al. 2006), so it is particularly relevant for the examination of behaviors in teams using AI. Communication refers to a reciprocal process of team members sending and receiving information that establishes and revises teams' attitudes, behaviors, and cognitions outcomes (Salas et al. 2015). AI can support human communication in

teams or share information with team members (Sukthankar and Sycara 2006); while this information may or may not reflect actual attitudes or cognitions on behalf of the AI (Ferdaus et al. 2024), communication behaviors can reflect the knowledge of the team and its interaction (Cooke et al. 2013). As such, these behaviors can be assessed similarly. Common measures using communication (e.g., communication frequency, language style matching, and prosody) can be used when comparing teams comprised solely of humans with teams using AI. When measuring communication in human–AI teams, measures of prosody (e.g., voice pitch, loudness, and vocal quality) may not be appropriate; current AI with vocal outputs cannot truly *emote*. In human–AI teams, where the AI can contribute to the team using LLM outputs, content-based approaches to measuring communication can be used. One approach is to present an LLM system with a scenario (e.g., asking for conflict management approaches to an interpersonal conflict scenario), request a solution, and then code the textual output to determine the type of solution or effectiveness of output (Cooper et al. 2024). If an AI does not have a dynamic text output capability, but instead uses pre-specified responses, then researchers can still use frequency-based analyses, such as patterns of interaction or category frequency (Lakhmani et al. 2022). One approach, an examination of the effect of spatial awareness on team cognition in human–agent teams using NeoCITIES, measured communication using the frequency of chats between teammates (Schelble et al. 2022b).

2.6.4 Measuring Cohesion

While team cohesion is not necessarily a direct indicator of effective team adaptation, it is an emergent state that affects numerous aspects of teamwork. Cohesion, a state that emerges from team interaction yielding an attraction to the group or a resistance to leaving it, has been described as one of the most important determinants of team success (Dion 2000; Salas et al. 2015). Consequently, determining how the integration of AI tools can affect the team's cohesion is a particularly relevant factor, especially given how some AI, such as LLMs, can now interact in significantly more human-like ways. At an individual level, an AI cannot feel anything about the group with which it works, and thus trying to measure that feeling may not be feasible. Cooper et al. (2024) map certain interpersonal processes in teamwork (e.g., conflict management, motivation and confidence building, and affect management) with measurable activities that can be done by AI with LLM capabilities (e.g., a negotiation task, a motivation problem task, and a probabilistic learning task), but while these tasks may correspond with the interpersonal aspect of cohesion, they do not encompass the construct as a whole. At a group level, however, AI tools can contribute to a group's feelings of cohesion, either through direct management of group affect, contributions to a team task, or

facilitating the work needed to complete a team's shared task (Lakhmani et al. 2022). A survey of cohesion that encompasses AI as part of the larger team (such as Neubauer et al.'s [2025] team cohesion scale for use in human–autonomy teams) is one option to measure this group-level construct.

2.7 Team Measures: Aggregation or Interaction?

Many of the measures discussed here assess team-level constructs (e.g., team SA, team cohesion, and shared mental models). Furthermore, these measures are specifically aimed at teams that are using advanced AI systems, such as LLMs that can communicate in a human-like manner, or AI that can fulfill a role in a team similar to a human. As such, there are different aspects to consider when using these measures in human–AI teams. The first aspect is the role of the AI in the team. AI can act as a tool that can help individual team members accomplish their specific tasks, as a facilitator of team interaction, or as an independent actor with a distinct role in the team (Sukthankar and Sycara 2006). The complexity of the AI and its role in the team determines how its behavior is measured. An AI that has its own role and acts as an independent agent within a team will affect and be affected by the group dynamics of a team, so it would be useful to measure its inputs, outputs, and processes as part of the larger team's work (Ososky et al. 2013; You and Robert 2018).

An AI that is used as a tool may be better examined as part of the user's workflow, and thus its output would be measured in terms of how it affects an individual's performance or a team's processes. The second aspect to consider when assessing team-level constructs in human–AI teams is whether the team-level construct is assessed by aggregating individual outcomes or through the activity of the team as a whole (Cooke et al. 2013). The aggregation approach, also described as the shared knowledge approach, measures individuals' cognitive states (e.g., SA, and mental models) with the understanding that these states can be combined and shared (McNeese et al. 2021). Alternatively, though not exclusively, the interaction approach suggests that interaction among team members reflects the team's cognitive state, so team states can be assessed by measuring relevant team interactions (Cooke et al. 2013; McNeese et al. 2021). Depending on the AI and how it is implemented in a team, different kinds of measures would be relevant, so that information should be remembered when determining appropriate assessments for human–AI teams.

2.8 Measuring the AI Teammate

Many metrics discussed in this review are only feasible to measure on the human members of human–AI teams. Many of the metrics are not appropriate to measure on AI itself as a teammate. An AI does not have subjective experiences in the way that humans do, and thus an AI cannot experience trust, cohesion, emotional sentiment, or other subjective states in the way that a human would. Similarly, the AI also does not have human physiology to assess. It also does not "reason" in the same way, nor does it approximate SA in the same way. Nonetheless, if measuring the AI teammate is desired, one could consider, for example, looking at its numerical estimates of confidence when predicting teammate behavior as a proxy for its level of trust. One could examine the gaze direction of onboard cameras or sensors as indicators of (potentially shared) attention. One could probe an AI's memory of relevant information to assess its approximation of SA. Measures of power usage or run time could function as indicators of "cognitive" workload. LLMs can and do verbally entrain with their human interlocutors (and humans can entrain to them), and a model sophisticated enough to include prosody could have the potential for prosodic entrainment as well. However, it must be remembered that what entrainment (or confidence levels, or SA, etc.) *means* to an AI is not what it means to a human.

In addition, some measures of individual or team-level states or processes may be appropriate for some types of AI but not others. Language-based probes or measures are of course not appropriate for AI models that do not use the natural language modality. Other AI systems, such as an LLM that simply summarizes a document, may not predict teammate behavior as part of their function and therefore would have nothing approximating trust or reliance that could be measured. The AI's capabilities and functions need to be considered when choosing what metrics to employ.

3. Table of Papers and Metrics

In this section, we present a table of several different types of methods used for measuring various constructs in human–AI teams. Table 1 highlights a subset of distinct example metrics and studies and is not intended to be exhaustive. The table provides a guide or starting point for researchers when choosing what types of metrics to use for assessing trust and other user states and team processes for human–AI teams. Each highlighted measure or algorithm is organized by how the relevant data is acquired (for example, by administering a questionnaire or examining communication). Additional methods are discussed within the text of different sections of this report.

Table 1. Non-exhaustive list of selected papers discussing metrics for measuring team and user states in human–AI interactions.

Measure or algorithm	Measure type	What does it measure?	When to measure	Citations
Trust in Automation (TIA)	Questionnaire	Trait-dependent dispositional trust of automation	Pre-study	Jian et al. 2000
Trust in Teams Scale (TTS)	Questionnaire	Propensity to trust, perceived trustworthiness, and cooperative and monitoring behaviors	Pre-/ post-study	Costa 2003; Costa and Anderson 2011
System Trustworthiness Scale (STS)	Questionnaire	Trust calibration of system's predictability, dependability, reliability over time, and competence	Pre-/ post-study	Muir and Moray 1996
Transactive Memory Scale (TMS)	Questionnaire	Specialization, credibility, and team coordination	Pre-/ post-study	Lewis 2003
Scale of Social Service Robot Interaction Trust (SSRIT)	Questionnaire	Trust	Before, or pre-/ post-study	Chi et al. 2021
Decision Confidence Scale (DCS)	Questionnaire (1 item)	Decision confidence	Post-decision	Giammanco 2021
AI Agents' Influence Scale	Questionnaire (1 item)	AI adoption	Post-decision	Giammanco 2021
Decision Styles Scale	Questionnaire	Rational and intuitive decision styles	Pre-study	Hamilton et al. 2016, 2017
Human–Automony Team (HAT) Cohesion Scale	Questionnaire	Cohesion	During or post-study	Neubauer et al. 2025
RoboGPT – ChatGPT-Enabled Human–Robot Collaboration System	Questionnaire	Trust in human–robot collaboration	After each condition	Ye et al. 2023
Trust Assessment in Human–Machine Interaction Study	Questionnaire + physiology	Dynamic trust in automation (autonomous driving system)	Pre-/post-sessions	Lu and Sarter 2019
Dynamic Trust Measurement in Human–Robot Interaction	Questionnaire + physiology	Dynamic, moment-to-moment trust in a robot teammate during HRI	After each phase of the task	Zhang et al. 2024
Trust in Autonomous Vehicles Questionnaire + Heart Rate	Questionnaire + physiology	Trust in autonomous vehicles	Post-study	Mosaferchi et al. 2024

Table 1. Non-exhaustive list of selected papers discussing metrics for measuring team and user states in human–AI interaction (continued).

Measure or algorithm	Measure type	What does it measure?	When to measure	Citations
Equations for Continuous Assessment of Trust (Reliance)	Behavior	Reliance, treated as a proxy for trust	During interaction with AI system	Okamura and Yamada 2020a; Okamura and Yamada 2020b
Acceptance/Rejection of the AI Agents' Recommendations	Behavior	AI adoption	Post-decision	Mercado et al. 2016; Stowers et al. 2020; Giammanco, in review
Recognition of the AI Agents' Plan Rationale and Planning Assumptions	Behavior	AI adoption	Post-decision	Mott et al. 2010; Giammanco, in review
Coordinated Awareness of Situation by Teams (CAST)	Behavior	Team situation awareness/ coordination	During Cycle (tracking responses to "roadblocks")	Gorman et al. 2005; McNeese et al. 2021
Card Sort	Elicitation	Shared mental models	Before, after, or pre-/post-study	Righi et al. 2013
Similarity Rating	Elicitation (questionnaires and behaviors)	Shared mental models	Pre-/post-study	Schelble et al. 2022a
Chat Frequency	Communication	Communication	During cycle	Lakhmani et al. 2022; Schelble et al. 2022b
LIWC (various)	Communication	Plan execution, affect, etc.	During or post-study	Jordan et al. 2019
Custom Language Model of Collaborative States	Communication	Collaborative states	Post-study	Doherty et al. 2025
Linguistic Entrainment	Communication	Indicator of cohesion, rapport, shared mental models	During or post-study	Dideriksen et al. 2023; Kian et al. 2025

4. Conclusion, Limitations, and Path Forward

In this report we presented a guide for researchers to use as they select different metrics for assessing trust and other user states and team processes for human–AI teams. Because this is a rapidly evolving field, our review provides a snapshot in an ongoing process as new measurement methods are being devised and existing measurements are being tested in new contexts. It must also be considered that this

literature review was not exhaustive. Hundreds of papers have measured team processes or user states when interacting with some form of "AI," and the pace of new publications is rapid. New metrics will likely be proposed as research progresses. We did not attempt to examine them all. In addition, the AI field is changing fast, and a particular questionnaire or metric might become obsolete a year after its creation. This also means there is little time to wait for new metrics to be thoroughly validated.

Another consideration when selecting measures is that "AI" is defined differently by different authors, and it was not always easy to tell how advanced the AI in any particular study was. This makes it challenging to know whether a particular metric would be suitable for a different AI use case, so we caution the reader to carefully consider the nature of the AI they wish to study and to assess the appropriateness of metrics for that AI on a case-by-case basis.

AI itself is changing quickly and may soon gain new capabilities that again represent a qualitative change from the previous state of the art. This may again necessitate a search for new metrics and reassessment of whether older metrics are still suitable. Going forward, we believe that a multimodal approach to assessing team states is an important step toward metrics that are suitable in the context of AI-enabled teams. We welcome feedback and new insights for other researchers working in this area.

5. References

Abbasian M, Azimi I, Rahmani AM, Jain R. Conversational health agents: a personalized large language model-powered agent framework. JAMIA Open. 2025;8(4). https://doi.org/10.1093/jamiaopen/ooaf067

Ackerman R. The diminishing criterion model for metacognitive regulation of time investment. J Exp Psychol Gen. 2014;143(3):1349–1368. https://doi.org/10.1037/a0035098

Ackerman R. Heuristic cues for meta-reasoning judgments: review and methodology. Psychol Topics. 2019;28(1):1–20. https://doi.org/10.31820/pt.28.1.1

Ackerman R, Beller Y. Shared and distinct cue utilization for metacognitive judgments during reasoning and memorization. Think Reason. 2017;23(4):376–408. https://doi.org/10.1080/13546783.2017.1328373

Ackerman R, Thompson V. Meta-reasoning: monitoring and control of thinking and reasoning. Trends Cogn Sci. 2017;21. https://doi.org/10.1016/j.tics.2017.05.004

Akash K, Hu W-L, Jain N, Reid T. A classification model for sensing human trust in machines using EEG and GSR. ACM Trans Interact Intell Syst. 2018;8(4):1–20. https://doi.org/10.1145/3132743

Andrews RW, Lilly JM, Srivastava D, Feigh KM. The role of shared mental models in human–AI teams: a theoretical review. Theor Iss Ergon Sci. 2023;24(2):129–175. https://doi.org/10.1080/1463922X.2022.2061080

Ayoub J, Avetisian L, Yang XJ, Zhou F. Real-time trust prediction in conditionally automated driving using physiological measures. IEEE Trans Intell Transp Syst. 2023;24(12):14642–14650. https://doi.org/10.1109/TITS.2023.3295783

Baker AL et al. Approaches for assessing communication in human–autonomy teams. Hum–Intell Syst Integr. 2021;3(2):99–128. https://doi.org/10.1007/s42454-021-00026-2

Behzad T, Gurney N, Wang N, Pynadath DV. Beyond predictions: a study of AI strength and weakness transparency communication on human–AI collaboration. arXiv; 2025 Aug 12. arXiv:2508.09033. https://doi.org/10.48550/arXiv.2508.09033

Burke CS, Stagl KC, Salas E, Pierce L, Kendall D. Understanding team adaptation: a conceptual analysis and model. J Appl Psychol. 2006;91(6):1189–1207. https://doi.org/10.1037/0021-9010.91.6.1189

Cannon-Bowers JA, Salas E, Converse S. Shared mental models in expert team decision making. In: Castellan JN Jr, editor. Individual and group decision making: current issues. Lawrence Erlbaum Associates, Inc.; c1993. p. 221–246.

Carmody PC, Mateo JC, Bowers D, McCloskey MJ. Linguistic coordination as an unobtrusive, dynamic indicator of rapport, prosocial team processes, and performance in team communication. Proceedings of the Human Factors and Ergonomics Society Annual Meeting; 2017 Oct 9–13; Austin, TX. 2017;61(1):140–144. Human Factors and Ergonomics Society; c2017. https://doi.org/10.1177/1541931213601518

Chancey ET, Bliss JP, Yamani Y, Handley HAH. Trust and the compliance–reliance paradigm: the effects of risk, error bias, and reliability on trust and dependence. Hum Factors. 2017;59(3):333–345. https://doi.org/10.1177/0018720816682648

Chi OH, Jia S, Li Y, Gursoy D. Developing a formative scale to measure consumers' trust toward interaction with artificially intelligent (AI) social robots in service delivery. Comput Hum Behav. 2021;118:106700. https://doi.org/10.1016/j.chb.2021.106700

Cooke NJ, Gorman JC, Myers CW, Duran JL. Interactive team cognition. Cogn Sci. 2013;37(2):255–285. https://doi.org/10.1111/cogs.12009

Cooke NJ, Salas E, Kiekel PA, Bell B. Advances in measuring team cognition. In: Salas E, Fiore SM, editors. Team cognition: understanding the factors that drive process and performance. American Psychological Association; 2004. p. 83–106.

Cooper PS, Irons JL, McGrath MJ, Duenser A. How well do large language models perform as team members? Testing teamwork capabilities of LLMs. PsyArXiv. 2024 Nov 15. https://doi.org/10.31234/osf.io/qyfrw

Cossio M. A comprehensive taxonomy of hallucinations in large language models. arXiv; 2025 Aug 3. arXiv:2508.01781. https://doi.org/10.48550/arXiv.2508.01781

Costa AC. Work team trust and effectiveness. Pers Rev. 2003;32(5):605–622. https://doi.org/10.1108/00483480310488360

Costa AC, Anderson N. Measuring trust in teams: development and validation of a multifaceted measure of formative and reflective indicators of team trust. Euro J Work Organ Psychol. 2011;20(1):119–154. https://doi.org/10.1080/13594320903272083

de Visser EJ et al. Learning from the slips of others: neural correlates of trust in automated agents. Front Hum Neurosci. 2018;12:309. https://doi.org/10.3389/fnhum.2018.00309

de Visser EJ et al. Towards a theory of longitudinal trust calibration in human–robot teams. Int J Soc Robot. 2019;12(2):459–478. https://doi.org/10.1007/s12369-019-00596-x

de Visser EJ et al. Mutually adaptive trust calibration in human–AI teams. HHAI-WS 2023: Workshops at the Second International Conference on Hybrid Human–Artificial Intelligence (HHAI). CEUR Workshop Proceedings; 2023 June 26–27; Munich, Germany. Vol. 3456, p. 188–193. CEUR-WS.org.

DeChurch LA, Mesmer-Magnus JR. Measuring shared team mental models: a meta-analysis. Group Dyn Theor Res Pract. 2010;14(1):1–14. https://doi.org/10.1037/a0017455

Di Stasi LL et al. Microsaccade and drift dynamics reflect mental fatigue. Eur J Neurosci. 2013;38(3):2389–2398. https://doi.org/10.1111/ejn.12248

Dideriksen C, Christiansen MH, Tylén K, Dingemanse M, Fusaroli R. Quantifying the interplay of conversational devices in building mutual understanding. J Exp Psychol Gen. 2023;152(3):864–889. https://psycnet.apa.org/doi/10.1037/xge0001301

Dion KL. Group cohesion: from "field of forces" to multidimensional construct. Group Dyn Theory Res Pract. 2000;4(1):7–26. https://psycnet.apa.org/doi/10.1037/1089-2699.4.1.7

Doherty E et al. Piecing together teamwork: a responsible approach to an LLM-based educational jigsaw agent. CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems; 2025 Apr 26–May 1; Yokohama, Japan. Association for Computing Machinery; c2025. p. 1–17. https://doi.org/10.1145/3706598.3713349

Endsley MR. Design and evaluation for situation awareness enhancement. Proceedings of the Human Factors Society 32nd Annual Meeting; 1988 Oct 24–28; Anaheim, CA. 1988;32(2):97–101. Human Factors and Ergonomics Society; c1988. https://doi.org/10.1177/154193128803200221

Endsley MR. Supporting human–AI teams: transparency, explainability, and situation awareness. Comput Hum Behav. 2023;140:107574. https://doi.org/10.1016/j.chb.2022.107574

Ferdaus MM et al. Towards trustworthy AI: a review of ethical and robust large language models. arXiv; 2024 June 1. arXiv:2407.13934. https://doi.org/10.48550/arXiv.2407.13934

Files BT, Pollard KA. Using large language models (LLMs) to summarize mission data for mission planning and after-action review (AAR). DEVCOM Army Research Laboratory (US); 2025. Report No.: ARL-TR-10239.

Giammanco C. Decision confidence scale. 2021. Unpublished.

Giammanco C. Human–artificial intelligence decision support for command and control (HAIDS C2) program closeout report. DEVCOM Army Research Laboratory (US); in review.

Gonzales AL, Hancock JT, Pennebaker JW. Language style matching as a predictor of social dynamics in small groups. Commun Res. 2010;37(1):3–19. https://doi.org/10.1177/0093650209351468

Gorman JC, Cooke NJ, Pederson HK, Olena OC, DeJoode JA. Coordinated awareness of situation by teams (CAST): measuring team situation awareness of a communication glitch. Proceedings of the Human Factors and Ergonomics Society Annual Meeting; 2005 Sep 26–30; Orlando, FL. 2005;49(3):274–277. Human Factors and Ergonomics Society; c2005. https://doi.org/10.1177/154193120504900313

Hamilton K, Shih S-I, Mohammed S. The development and validation of the rational and intuitive decision styles scale. J Pers Assess. 2016;98(5):523–535. https://doi.org/10.1080/00223891.2015.1132426

Hamilton K, Shih S-I, Mohammed S. The predictive validity of the decision styles scale: an evaluation across task types. Pers Individ Diff. 2017;119:333–340. https://doi.org/10.1016/j.paid.2017.08.009

Holder E, Pollard K, Krausman A. Integrating AI and emergent technology at the lower echelon. DEVCOM Army Research Laboratory (US); 2025. Report No.: ARL-TR-10277.

Holland C, Perry G, Neyedli HF. Calibrating trust, reliance and dependence in variable-reliability automation. Proceedings of the Human Factors and Ergonomics Society Annual Meeting; 2024 Sep 9–23; Phoenix, AZ.

2024;68(1):604–610. Human Factors and Ergonomics Society; c2024. https://doi.org/10.1177/10711813241277531

Huang L et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Trans Informat Syst. 2025;43(2):1–55. https://doi.org/10.1145/3703155

Jian J-Y, Bisantz AM, Drury CG. Foundations for an empirically determined scale of trust in automated systems. Int J Cogn Ergon. 2000;4(1):53–71. https://doi.org/10.1207/S15327566IJCE0401_04

Johnson M, Vera AH. No AI is an island: the case for teaming intelligence. AI Magazine. 2019;40(1):16–28. https://doi.org/10.1609/aimag.v40i1.2842

Johnson TE et al. Measuring sharedness of team-related knowledge: design and validation of a shared mental model instrument. Hum Res Dev Int. 2007;10(4):437–454. https://doi.org/10.1080/13678860701723802

Jordan KN, Sterling J, Pennebaker JW, Boyd RL. Examining long-term trends in politics and culture through language of political leaders and cultural institutions. Proc Natl Acad Sci. 2019;116(9):3476–3481. https://doi.org/10.1073/pnas.1811987116

Kaplan AD, Kessler TT, Brill JC, Hancock PA. Trust in artificial intelligence: meta-analytic findings. Hum Factors. 2023;65(2):337–359. https://doi.org/10.1177/00187208211013988

Kian M et al. Using linguistic entrainment to evaluate large language models for use in cognitive behavioral therapy. In: Chiruzzo L, Ritter A, Wang L, editors. Findings of the Association for Computational Linguistics: NAACL 2025; 2025 Apr. p. 7739–7758. https://doi.org/10.18653/v1/2025.findings-naacl.430

Krausman A et al. Trust measurement in human–autonomy teams: development of a conceptual toolkit. ACM Trans Hum–Robot Interact. 2022;11(3):1–58. https://doi.org/10.1145/3530874

Kucukosmanoglu M et al. Exploring trust in AI-supported military teams using sentiment analysis. Interact Stud. 2025;26(2):229–266.

Lakhmani SG et al. Guidelines for collecting laboratory speech data. DEVCOM Army Research Laboratory (US); 2022. Report No.: ARL-TR-9406.

Lau GR et al. Predicting trust toward media content by eye movement. APCV 2024: Asia Pacific Conference on Vision; 2024 July 10–12; Singapore. https://doi.org/10.17605/osf.io/5fnyr

Lewis K. Measuring transactive memory systems in the field: scale development and validation. J Appl Psychol. 2003;88(4):587–604. https://doi.org/10.1037/0021-9010.88.4.587

Lu Y, Sarter N. Eye tracking: a process-oriented method for inferring trust in automation as a function of priming and system reliability. IEEE Trans Hum–Mach Syst. 2019;49(6):560–568. https://doi.org/10.1109/THMS.2019.2930980

Lu Y, Sarter N. Modeling and inferring human trust in automation based on real-time eye tracking data. Proceedings of the Human Factors and Ergonomics Society Annual Meeting; 2020 Oct 5–9; virtual meeting. 2020;64(1):34–348. Human Factors and Ergonomics Society; c2020. https://doi.org/10.1177/1071181320641078

Lukin SM et al. Consequences and factors of stylistic differences in human–robot dialogue. In: Komatani K et al., editors. Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue; 2018 July; Melbourne, Australia. Association for Computational Linguistics; c2018. p. 110–118. https://doi.org/10.18653/v1/W18-5012

Marshall SP. Identifying cognitive state from eye metrics. Aviat Space Environ Med. 2007;78(5):B165–B175.

McNeese NJ, Demir M, Cooke NJ, She M. Team situation awareness and conflict: a study of human–machine teaming. J Cogn Eng Decis Mak. 2021;15(2–3):83–96. https://doi.org/10.1177/15553434211017354

Mehrotra S, Degachi C, Vereschak O, Jonker CM, Tielman ML. A systematic review on fostering appropriate trust in human–AI interaction: trends, opportunities and challenges. ACM J Respons Comput. 2024;1(4):1–45. https://doi.org/10.1145/3696449

Mercado JE et al. Intelligent agent transparency in human–agent teaming for multi-UxV management. Hum Factors. 2016;58(3):401–415. https://doi.org/10.1177/0018720815621206

Mitkidis P, McGraw JJ, Roepstorff A, Wallot S. Building trust: heart rate synchrony and arousal during joint action increased by public goods game. Physiol Behav. 2015;149:101–106. https://doi.org/10.1016/j.physbeh.2015.05.033

Mohammed S, Klimoski R, Rentsch JR. The measurement of team mental models: we have no shared schema. Organ Res Meth. 2000;3(2):123–165. https://doi.org/10.1177/109442810032001

Mosaferchi S et al. Beat trustfully: the correlation between heart rate and a multi-dimensional trust questionnaire. In: Kocian A, Milazzo P, Henriques Martins AL, Nanni M, Pappalardo L, editors. Intelligent transport systems. INTSYS 2024. Lecture notes of the institute for computer sciences, social informatics and telecommunications engineering. Vol. 608; p. 3–15. Springer, Cham; c2025. https://doi.org/10.1007/978-3-031-86370-7_1

Mott D, Giammanco C, Dorneich MC, Braines D. Hybrid rationale for shared understanding. Proceedings of the Fourth Annual Conference of the International Technology Alliance (IITA 2010); 2010 Nov 4–5; Bangkok, Thailand.

Muir BM. Trust between humans and machines, and the design of decision aids. Int J Man–Mach Stud. 1987;27(5–6):527–539. https://doi.org/10.1016/S0020-7373(87)80013-5

Muir BM, Moray N. Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. Ergonomics. 1996;39(3):429–460. https://doi.org/10.1080/00140139608964474

Neubauer C et al. Human–autonomy teaming trust toolkit (HAT3) software development documentation and user guide. DEVCOM Army Research Laboratory (US); 2023. Report No.: ARL-TR-9623.

Neubauer C et al. 4 – Development of a team cohesion scale for use in human–autonomy team research. Interdependent human–machine teams. In: Lawless W, Mittu R, Sofge D, Fouad H, editors. Interdependent Human–machine teams. The path to autonomy. Academic Press; c2025. p. 67–99. https://doi.org/10.1016/B978-0-443-29246-0.00010-9

Ng SWT, Zhang R. Trust in AI chatbots: a systematic review. Telemat Informat. 2025;97:102240. https://doi.org/10.1016/j.tele.2025.102240

Nguyen D et al. Exploratory models of human–AI teams: leveraging human digital twins to investigate trust development. arXiv; 2024 Nov 1. arXiv:2411.01049. https://doi.org/10.48550/arXiv.2411.01049

Okamura K, Yamada S. Adaptive trust calibration for human–AI collaboration. PLOS ONE. 2020a;15(2):e0229132. https://doi.org/10.1371/journal.pone.0229132

Okamura K, Yamada S. Calibrating trust in human–drone cooperative navigation. 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN 2020); 2020b Aug 31–Sep 4; Naples, Italy. IEEE

Robotics and Automation Society; c2020. p. 1274–1279. https://doi.org/10.1109/RO-MAN47096.2020.9223509

Ososky S, Schuster D, Phillips E, Jentsch FG. Building appropriate trust in human–robot teams. 2013 AAAI Spring Symposium Series; 2013. Vol. SS-13-07; p. 60–65. Association for the Advancement of Artificial Intelligence; c2013.

Righi C et al. Card sort analysis best practices. J Usability Stud. 2013;8(3):69–89.

Salas E, Shuffler ML, Thayer AL, Bedwell WL, Lazzara EH. Understanding and improving teamwork in organizations: a scientifically based practical guide. Hum Res Manage. 2015;54(4):599–622. https://psycnet.apa.org/doi/10.1002/hrm.21628

Schaefer KE. Measuring trust in human robot interactions: development of the "*trust perception scale–HRI*." In: Mittu R, Sofge D, Wagner A, Lawless WF, editors. Robust intelligence and trust in autonomous systems. Springer; 2016. p. 191–218. https://doi.org/10.1007/978-1-4899-7668-0_10

Schelble BG, Flathmann C, McNeese NJ, Freeman G, Mallick R. Let's think together! Assessing shared mental models, performance, and trust in human–agent teams. In: Nichols J, editor. Proceedings of the ACM on Human–Computer Interaction. 2022a;6(GROUP):1–29. Association for Computing Machinery; c2022. https://doi.org/10.1145/3492832

Schelble BG, Flathmann C, Musick G, McNeese NJ, Freeman G. I see you: examining the role of spatial information in human–agent teams. Proceedings of the ACM on Human–Computer Interaction. 2022b;6(CSCW2):1–27. https://doi.org/10.1145/3555099

Stowers K et al. The IMPACT of agent transparency on human performance. IEEE Trans Hum–Mach Syst. 2020:1–9. https://doi.org/10.1109/THMS.2020.2978041

Sukthankar G, Sycara K. Simultaneous team assignment and behavior recognition from spatio-temporal agent traces. In: Cohn A, editor. AAAI '06. Proceedings of the 21st National Conference on Artificial Intelligence – Vol. 1; 2006 July 16–20; Boston, MA. 2006;6:716–721. AAAI Press; c2006.

Tsai YF, Viirre E, Strychacz C, Chase B, Jung T-P. Task performance and eye activity: predicting behavior relating to cognitive workload. Aviat Space Environ Med. 2007;78(5Suppl):B176–B185.

Ueno T et al. Trust in human–AI interaction: scoping out models, measures, and methods. CHI EA '22: CHI Conference on Human Factors in Computing

Systems Extended Abstracts. 2022;(254):1–7. https://doi.org/10.1145/3491101.3519772

Vail A et al. Toward causal understanding of therapist–client relationships: a study of language modality and social entrainment. In: Tumuluir R, Sebe N, Pingali G, editors. ICMI ’22: Proceedings of the 2022 International Conference on Multimodal Interaction; 2022 Nov 7–11; Bengaluru, India. 2022;487–494. Association for Computing Machinery; c2022. https://doi.org/10.1145/3536221.3556616

Wynn CJ, Borrie SA. Classifying conversational entrainment of speech behavior: an updated framework and review. J Phon. 2022;94(6):101173. https://doi.org/10.1016/j.wocn.2022.101173

Xu Z, Jain S, Kankanhalli M. Hallucination is inevitable: an innate limitation of large language models. arXiv; 2024 Jan 22. arXiv:2401.11817. https://doi.org/10.48550/arXiv.2401.11817

Ye Y, You H, Du J. Improved trust in human–robot collaboration with ChatGPT. IEEE Access. 2023;11:55748–55754. https://doi.org/10.1109/ACCESS.2023.3282111

You S, Robert LP Jr. Human–robot similarity and willingness to work with a robotic co-worker. In: HRI ’18: Proceedings of the 2018 ACM/IEEE International Conference on Human–Robot Interaction; 2018 Mar 5–8, Chicago IL. 2018:251–260. Association for Computing Machinery; c2018. https://doi.org/10.1145/3171221.3171281

Zhang Y, Yadav A, Hopko SK, Mehta RK. In gaze we trust: comparing eye tracking, self-report, and physiological indicators of dynamic trust during HRI. In: HRI ’24: Companion of the 2024 ACM/IEEE International Conference on Human–Robot Interaction; 2024 Mar 11–15; Boulder, CO. 2024;1188–1193. Association for Computing Machinery; c2024. https://doi.org/10.1145/3610978.3640649

List of Symbols, Abbreviations, and Acronyms

AI	Artificial Intelligence
ARL	Army Research Laboratory
C2	Command and Control
CAST	Coordinated Awareness of Situation by Teams
COA	Course of Action
DEVCOM	U.S. Army Combat Capabilities Development Command
ECG	Electrocardiography
EEG	Electroencephalogram
HAT ERP	Human–Autonomy Teaming Essential Research Program
HRI	Human–Robot Interaction
HRV	Heart Rate Variability
LIWC	Linguistic Inquiry and Word Count
LLM	Large Language Model
ML	Machine Learning
SA	Situation Awareness
SAGAT	Situation Awareness Global Assessment Technique
SPAM	Situation Present Assessment Method
STS	System Trustworthiness Scale
TIA	Trust in Automation
TTS	Trust in Teams Scale
UAS	Unmanned Aerial System
XAI	Explainable AI

1 (PDF)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA
1 (PDF)	DEVCOM ARL FCDD RLB CB TECH LIB