Improving Measurement of Trust Dynamics in Human-Agent Teams

Cherrise Ficke, Kendall Carmody, Daniel Nguyen, Isabella Piasecki, Arianna Addis, Mohammed Akib,
Amanda L. Thayer, Jessica L. Wildman, Meredith Carroll
Florida Institute of Technology
Melbourne, Florida
cficke2018@my.fit.edu, kcarmody2016@my.fit.edu, nguyend2018@my.fit.edu, ipiasecki2019@my.fit.edu, aaddis2021@my.fit.edu, makib2021@my.fit.edu, athayer@fit.edu, jwildman@fit.edu, mcarroll@fit.edu

ABSTRACT

Key to studying and assessing trust and other team emergent states in human-agent teams (HATs) is the ability to measure trust, which has predominantly been assessed through self-report survey methodologies. However, on their own, self-report measures are limited by issues such as social desirability (e.g., Arnold & Feldman, 1981; Taylor, 1961), inaccuracies due to retrospective assessments of abstract concepts (Podsakoff & Organ, 1986), the assessment of trust as static rather than a dynamically emerging state (Kozlowski, 2015), and the impracticality of asking team members to pause tasks to complete surveys. There is a clear need for innovative approaches to better capture trust for both research and applied purposes. Recently, researchers have recommended and begun incorporating more unobtrusive measurement methodologies such as physiological measures, event-based behavioral assessments, and analysis of language/communication (Azevedo-Sa et al., 2021; Hill et al., 2014; Marathe et al., 2020; Waldman et al., 2015). Unobtrusive measures offer many benefits beyond self-report measures, including being more objective, more predictive, more dynamic and real-time, and interfering less with taskwork and teamwork. Meanwhile, behavioral measures of trust, such as allocating tasks to autonomous agents and manually controlling agents, are readily available and also correlated with trust (Schaefer et al., 2021; Khalid et al., 2021). On their own, none of these approaches comprehensively measure trust across a variety of HAT domains and interactions. By evaluating and mapping out known measures of trust to use cases, this paper presents a review of the literature in this field and proposes a theoretically-grounded Integrative Measurement Framework of Trust Dynamics in HATs that will more accurately, effectively, and practically capture trust in HATs by combining traditional and contemporary measurement approaches.

ABOUT THE AUTHORS

Cherrise Ficke is a Graduate student in Human Factors in Aeronautics at Florida Tech's College of Aeronautics. She graduated with a BS in Aviation Management in the Spring of 2022. Cherrise currently holds a Private Pilot's License (PPL) and intends to pursue a career in human factors. She has previously worked at the Naval Air Warfare Center Training Systems Division (NAWCTSD) in Corona, California as an intern developing augmented reality (AR) environments for training procedures. Cherrise's research interests include decision making, unmanned aerial systems (UAS) operations, human-agent teaming (HAT), AR, and virtual environments (VE).

Kendall Carmody graduated with a MS in Aviation Human Factors from the College of Aeronautics, and is a lead researcher at ATLAS lab. She has a B.S in Aeronautical Science, and a Minor in Aviation Environmental Science. Kendall has worked as an Airport Operations Specialist intern at RockHill-York County airport and is currently pursuing a Ph.D. in Aviation Sciences, with a focus on human factors, virtual environments, and virtual/augmented reality.

Daniel Nguyen is a Ph.D. candidate in Industrial/Organizational Psychology. He received his B.A. in Psychology at Texas A&M University in 2017, then went on to receive his M.S. in I/O Psychology at Florida Tech in 2020. His research interests and experiences are focused on work teams, with a special focus on human-agent teaming which has led him to a broader interest in related human-factors topics such as human-performance and trust in automation.

Isabella Piasecki is an undergraduate student in Aviation Human Factors in Florida Tech's College of Aeronautics. She currently holds a Private Pilot's License (PPL) and intends to pursue a career in human factors or accident investigation. Her research background includes human-agent teaming (HAT) and training development.

Arianna Addis is a Ph.D. student in Industrial/Organizational Psychology. She received her B.S. in Psychology from the University of Washington in 2020, and is interested in researching the intersection between teams, technology, trust, and well-being.

Mohammed Akib is a graduate student in Industrial/Organizational Psychology at the Florida Institute of technology. He received his B.S. in Psychology from the University of Central Florida in 2016. His research interests and background includes human-agent teaming (HAT), work teams, and LGBTQ+ experience.

Dr. Amanda L. Thayer is an Assistant Professor of Industrial/Organizational Psychology and Director of Growth and Development at the Institute for Culture, Collaboration, and Management (ICCM) at Florida Institute of Technology. She serves on the Editorial Boards for the *Journal of Business and Psychology* and *Group & Organization Management* and has served on the Board of Directors for the Interdisciplinary Network for Group Research and SIOP Scientific Affairs Committee Chair. Dr. Thayer completed her doctorate at University of Central Florida in 2015. To date, she has secured over $7 million in external funding and played an integral role in several interdisciplinary research efforts that have produced more than 80 publications and conference presentations, including outlets such as *American Psychologist, Human Resource Management Review*, and *Organizational Psychology Review*. She has conducted lab- and field-based research for government agencies, the military, and industry. Her research is focused on facilitating teamwork and collaboration across a variety of contexts, including military units, NASA crews, volunteer non-profit teams, and virtual teams, among others. Dr. Thayer's current research focuses on team selection, staffing, and composition; trust development, violation, and repair; team cohesion; team adaptation and resilience; and measurement and methodologies for studying interpersonal relationships and team dynamics.

Dr. Jessica L. Wildman is a tenured Associate Professor of Industrial/Organizational Psychology and the Research Director of the Institute for Culture, Collaboration, and Management (ICCM) at the Florida Institute of Technology. To date, she has co-authored 22 book chapters, 15 journal articles, and presented over 70 national and international conference presentations on topics including team processes and emergent states, team cognition, team performance measurement, global virtual teams, trust development and repair, and cultural competence. Dr. Wildman has been associated with over $2.5 million in funded research for clients including the U.S. Office of Naval Research, Naval Air Warfare Training Systems Division, U.S. Army Research Institute, NASA, the Air Force Office of Scientific Research, and large multinational companies. Exemplar projects include an operational assessment and scientific literature review on team self-maintenance for NASA, developing measures of multiteam system performance for the Navy, an early career grant from ARI to conduct basic research on the development, violation, and repair of trust across cultures, and cultural advising for a Fortune 500 company. Current research interests include diversity, equity, and inclusion (DEI), culture, and teams in the workplace.

Dr. Meredith Carroll is a Professor of Aviation Human Factors and Director of the Advancing Technology-interaction and Learning in Aviation Systems (ATLAS) Lab at Florida Institute of Technology's College of Aeronautics. She has nearly 20 years of experience studying human/team performance and training in complex systems. Her research focuses on decision making in complex systems, cognition and learning, human-machine teaming, performance assessment and adaptive training. She has been funded by the Federal Aviation Administration (FAA), the Air Force Research Laboratory (AFRL), the Air Force Office of Scientific Research (AFOSR), the Office of Naval Research (ONR), and the Army Research Laboratory (ARL) to study different facets of these areas. She also worked at the Kennedy Space Center conducting user-centered design of International Space Station payloads and processing facilities. She teaches a range of human factors courses aimed at giving students practical, hands-on experience in applying theories of cognition and learning to optimize performance in a range of situations. She received her Bachelor's degree in Aerospace Engineering from the University of Virginia, her Master's degree in Aviation Science from Florida Institute of Technology and her Ph.D. in Applied Experimental Psychology and Human Factors from the University of Central Florida.

Capturing Trust Dynamics in Human-Agent Teams through Unobtrusive Measurement

Cherrise Ficke, Kendall Carmody, Daniel Nguyen, Isabella Piasecki, Arianna Addis, Mohammed Akib,
Amanda L. Thayer, Jessica L. Wildman, Meredith Carroll
Florida Institute of Technology
Melbourne, Florida
cficke2018@my.fit.edu, kcarmody2016@my.fit.edu, nguyend2018@my.fit.edu, ipiasecki2019@my.fit.edu,
aaddis2021@my.fit.edu, makib2021@my.fit.edu, athayer@fit.edu, jwildman@fit.edu, mcarroll@fit.edu

INTRODUCTION TO TRUST IN HUMAN-AGENT TEAMS

Due to improvement of agent capabilities, the prevalence of autonomous systems is expanding in contemporary society. For instance, the United States Air Force is preparing their fleet to be 60% unmanned or optionally manned by 2035 (Otto, 2016). Agents can now team with humans to perform complex tasks in settings such as tactical intelligence surveillance and reconnaissance, cyber defense, and suppression of enemy air defenses (Otto, 2016; Chen & Barnes, 2014). As human-agent teams (HATs) operate in these interdependent and complex settings, properly calibrating a human's trust in agents becomes a critical factor that affects team dynamics and effectiveness (Yu et al., 2019; Kohn et al., 2020). To assess human-agent teamwork, trust in agents must be effectively measured.

While many measures of trust are used in the HAT literature, limitations surrounding their effectiveness have been raised (Lu & Sarter, 2020). Self-report measures are the most common approach for measuring trust (Merritt & Ilgen, 2008) but are task-disruptive and impractical for real-world scenarios. For example, in fast-paced and dynamic environments such as search and rescue missions, administering surveys during the mission would interrupt performance and cost valuable time that should be spent rescuing victims. Behavioral measures have been used as an alternative to self-report measures of trust. For example, rejecting agent input has been inferred as low trust (Wang et al., 2017). However, behaviors are task-specific, and although they are highly correlated, behaviors do not equate to attitudes (Ajzen, 1991). Physiological measures, such as electroencephalography (EEG), have also been used (Khawaji et al., 2015; Dong et al., 2015). Physiological measures are objective and able to capture dynamic changes in trust via a constant data stream. Unfortunately, collecting and analyzing this data can be costly, physically intrusive to participants, and confounded with other constructs when used alone. However, when context is taken into account and multiple measures are used in tandem, they can better approximate trust without disrupting a team's momentum. Though there are numerous approaches to measuring HAT trust, there is disparate guidance for capturing dynamic trust over time. To unify the existing literature and guide future research on trust in HATs, this paper 1) reviews current trust measurement approaches within the HAT literature and 2) proposes an integrated framework of HAT trust measures that evaluates and maps known measures to use cases to suggest context-appropriate measures of trust.

REVIEW OF TRUST MEASURES IN HUMAN-AGENT TEAMS

Trust is an elusive construct to measure, as it has been conceptualized in a variety of ways. Due to the diverse options available, choosing the appropriate measure can be difficult and confusing. These measures can be broadly organized into three major categories: self-report measures, behavioral measures, and physiological measures.

Self-Report Measures

Multi-Item Self-Report Scales

Self-report scales are among the most accessible forms of measurement in any given research domain, especially of attitudes (Kohn et al. 2020). Many trust in HAT studies have employed self-report trust scales. However, they often differ, as a myriad of self-report trust scales exist. Among the many scales used, two scales stand out as frequently used measures of trust in the HAT research community: Jian et al's (2000) Trust in Automated Systems survey, and Mayer & Davis's (1999) Measure of Trust & Trustworthiness.

Jian et al's (2000) Trust in Automated Systems survey identifies trust as a construct containing opposite, antithetical dimensions of trust (positive vs. negative). Jian et al's (2000) study conducted a cluster analysis resulting in identification of 12 trust items. This scale treats trust as a singular bipolar construct with trust and distrust on either end. The scale recognizes multiple elements of trust in automation including positive and negative items about the automated system's reliability, security, familiarity, and integrity in HATs (Kohn et al., 2021). The scale can be used either at the sub-facet level to understand particular dimensions of trust, or as a composite scale of overall trust (Jian et al., 2000). The scale has been employed consistently as a valid measure of trust (Kohn et al., 2020).

As agents in HATs transition to more interdependent teammate roles rather than tools, it is also important to examine the interpersonal nature of trust (Mayer et al., 1995). Researchers have adapted Mayer & Davis's (1999) Measure of Trust and Trustworthiness based on the organizational trust framework developed by Mayer et al. (1995). The scale and framework both separate trust from its antecedents, identifying trust as an evolving attitude in which a trustor is willing to be vulnerable to another party based not only on the trustor's propensity to trust, but also on the trustor's perception that the trustee is trustworthy (Alarcon et al., 2018; Mayer et al., 1995). In this scale, trustworthiness encompasses three dimensions: Ability, Benevolence, and Integrity. Propensity to trust and overall trustworthiness are separate constructs that predict trust. Studies of trust in HATs have often adapted one or more dimensions to assess a specific element of trust. Despite the original framework defining propensity to trust as separate from trust, modern adaptations of this scale (1999) have added propensity to trust as a sub-dimension (Kohn et al., 2020; Körber, 2018). The original measure validation studies of employee trust found evidence of reliability and validity and it has since been used regularly in HAT research as a credible source (Kohn et al., 2020)**.** That said, we do not endorse adding propensity to trust as a sub-dimension of dynamic momentary trust, due to its theoretical distinctiveness.

Single-Item Self-report Measures

In addition to validated scales, single-item self-report measures of trust in automation have also been employed (Körber, 2018). Single-item measures are most often used when trust needs to be measured often to capture the dynamic changes in trust. Researchers have embedded the item within missions as a brief pop-up menu that pauses the simulation to capture trust as it fluctuates during task performance (Hergeth et al., 2015). However, some researchers argue that single-item trust measures are inherently unidimensional, while the literature often defines trust as a multi-dimensional construct (Lee & See, 2004; Mayer et al., 1995). The validity of single-item trust measures has thus been criticized for their inability to parse out dimensions of trust (Körber, 2018; Kohn et al., 2021). Researchers have also voiced concerns that single-item measures are inadequate for capturing abstract constructs not easily defined (Körber, 2018; Morgado et al., 2017). Single-item measures have thus been criticized because participant responding reflects their own beliefs and definitions about a construct rather than a standardized definition. However, most people generally have a sense of what it means to trust another entity or at least have their own personal definition of trust (de Fine Licht & Brulde, 2021). Single-item measures also offer multiple benefits. First, single item measures take less time to complete, enabling trust to be captured in the middle of a mission and thus capture changes in trust alongside trust-related events. Second, single-item measures enable more fine-grained analyses of trust networks. Even within single-human HATs with multiple agents, single-item measures still enable investigation of ego-centric networks that connect trust scores amongst the different dyadic relationships that one human has with each agent.

Behavioral Measures

Usage Behaviors

Overwhelmingly, previous studies utilizing behavioral measures of trust in HATs have focused on usage measures to capture over-reliant and under-reliant behaviors of trust in agents (Wang et al., 2017). Over-relying occurs when the human operator relies on the agent to complete tasks or make decisions when one should not be (Parasuraman & Riley, 1997). Under-relying occurs when the human operator completes tasks or makes decisions without assistance from the agent. Parasuraman and Riley (1997) outline four behaviors to describe how humans use automation: use, disuse, abuse, and misuse. Use, disuse, and misuse are outlined below as primary usage behaviors in identifying trusting behaviors. Automation *abuse* is a behavior that demonstrates operator use of automation without regard for the consequences for human performance. Abuse relies on the design of the automation rather than being trust-dependent, which is why it is not comprehensively reviewed in the following section.

Parasuraman and Riley (1997) define *use* as the intentional activation of automation. A common example would be self-driving capabilities in cars or autopilot in aircraft. Using automation is indicative of trust since the decision to engage in automation is greatly influenced by an operator's trust in it (Lee & Moray, 1992; Muir, 1987). Similarly, following an agent's recommendations is consider use behavior (Freedy et al., 2007). Conversely, *disuse* of automation is defined as the neglect or underutilization of automation (Parasuraman & Riley, 1997). Disuse, such as rejecting agent recommendations, suggests low operator trust (Wang et al. (2017).

The *misuse* of automation refers to overreliance, resulting in failures of supervision from the operator (Parasuraman & Riley 1997) and detrimental effects on performance (Lee & See, 2004). This typically occurs when the operator experiences high workload or has low confidence in their ability to accomplish the task. For example, Freedy et al. (2007) simulated a military escort task where participants worked with agents to eliminate hostile enemies to protect a civilian envoy. The human operator monitored the agent with the option to manually override and eliminate enemies if they felt the enemy was getting too close. Frequently overriding the automation indicated under-trusting, while few instances of overriding indicate over-trust. The study results suggested behavioral actions correlated with self-report scores. Although agents' capabilities have advanced, all automation has a failure rate. Thus, it is important that human teammates appropriately monitor agents and intervene when automation does fail.

Usage behaviors have been widely used in previous literature, demonstrating high correlations with trust. However, the majority of published HAT research paradigms set the agent as a tool rather than a teammate. Some HAT research has pushed toward human-agent cooperation in their experiments. For instance, Demir et al. (2019) developed a space exploration task in which a human and two agents had interdependent roles such that the navigator and pilot had to fly to waypoints for the photographer to take photos of targets. This task paradigm exemplifies collaborative HAT environments in which agents and human work together to achieve a common goal. To understand trust when humans and agents are interdependent, behavioral measures must go beyond use/disuse for other indicators of collaboration.

Context-Specific Action Behaviors

Context-specific action behaviors are behaviors humans exhibit when working with an agent and are highly dependent on the task. For example, in a study examining robots as social partners, waving and speaking to the robot was indicative of trust in the robot (Bainbridge et al., 2011). In another study, participants' comfort levels were determined by their physical proximity to the robot, suddenness of movements around the robot, and physical cues used when responding to the robot (e.g., nodding when agreeing or shaking their head when disagreeing; Fratczak et al., 2021). Furthermore, Buchholz et al. (2017) employed the Give-Some Dilemma mixed-motive game where participants and agents exchanged tokens with the coin value doubling when exchanged. Thus, individual outcomes were maximized when participants kept and received all coins; team outcomes were maximized when both players exchanged all coins. An increased number of positive behaviors (high number of tokens exchanged, positive communication, and quicker response time) correlated with a higher trust. Studies have found using context-specific action behaviors can be a reliable indicator of trust (Albayram et al., 2020); however, this is dependent on properly identifying context-specific trust actions within the right scenario. For example, physical proximity to the agent may be pertinent in an in-person experimental environment, but irrelevant in a computer-based interaction (Bainbridge et al., 2011).

Physiological Measures

Eye Tracking

Gaze, defined as visual fixation over an area of interest in a subject's visual field (Erickson et al., 2020), is one of the most common eye-tracking metrics for measuring trust in automated agents. How many times and how long a participant gazes over a certain area demonstrates a participant's extraction of information from their surroundings to carry out a task (Bales & Kong, 2017). Most studies employing gaze found correlations related to performance rather than trust (Bales & Kong, 2017; Wright et al., 2014; Bagheri et al., 2004). However, newer studies have developed predictive models from eye-tracking data for inferring trust levels in real time. In a study by Lu and Sarter (2020), participants completed an unmanned aerial vehicle (UAV) desktop simulation including visual search and tracking tasks. The visual search task involved the participant monitoring eight UAV camera feeds to detect military targets. Gaze, defined by how much time the participant fixated on the area where the visual search task took place, and transition count, or the number of transitions in gaze between two areas of interest, were used to measure trust in the agent. After collecting baseline information from participants at the beginning of the study, Lu and Sarter (2020) developed predictive models of gaze and transition count, which achieved 80% accuracy. As automation reliability changed from high to low, gaze and transition count increased, inferring potential decreases in trust. Although these measures were not validated through self-report, results showed promise for eye-tracking measures in the HAT space.

Vocal Pitch

Vocal pitch refers to how high or low a participant's voice is (Niculescu et al., 2013). Elkins and Derrick (2013) investigated the relationship between vocal dynamics (pitch and duration) and trust in an Embodied Conversational Agent (ECA). Participants completed a face-to-face interview with the ECA and then reported their trust in it. Vocal pitch was inversely related to trust, and this effect was strongest early in the interview. However, this stronger initial correlation may be due to either increasing trust throughout the study or the participant's recognition of their change in pitch and a conscious effort to fix it. While this study suggests that vocal pitch may be a good indicator of initial trust, further research is needed to examine the reliability of voice pitch as a measure of trust for longer time periods.

Cardiovascular Measures

Cardiovascular measures i include electrocardiograms (ECGs), which quantify heart rate and heart rate variability using electrical leads (Johnson et al., 2021). Heart rate variability has been found to pair well with behavioral trust indicators (Tolston, 2018). In Van Lange and colleague's (2011) study, participants interacted with a team through a consensus wagering task, then placed wagers on the agent's performance for a maze running task. The study found that heart rate variability and self-report measures of trust in the agent from round one were significant predictors of trust behavior in an agent for round two. Mitkidis et al. (2015) conducted a study observing heart rate synchrony and arousal as a predictor of trust during a public goods game. Thirty-seven pairs of participants' constructed LEGO model cars during four consecutive sessions lasting 10 minutes each. Participants' heart rates were measured during the interaction to observe synchrony. Following the car-building sessions (the trust condition), twenty pairs of participants completed a public good game in which theyd contributed any amount of money out of their own pot to a public pool, with total payoff being optimized when players contribute all available funds. They found that participants' heart rate profiles were more synchronized during the trusting condition and heart rate synchrony was a significant predictor of expectations during the public goods game, suggesting heart-rate synchrony may be an indicator of interpersonal trust.

Neural Signatures

Electroencephalogram (EEG) and functional magnetic resonance imaging (fMRI) data are the most common measures for detecting neural signatures for HAT trust. fMRI is a frequently used brain imaging method, which detects changes in blood-oxygenation levels (Menon & Crottaz-Herbette, 2005). EEG measures neuronal processing through electrical brain activity by attaching electrodes to the scalp (Subha et al., 2010). Although neural signatures are not as prevalent as self-report or behavioral measures, studies have found correlations between activation of certain brain regions and HAT interactions in both fMRI and EEG data. In Lee (2018) participants played a mixed motive game whilst collecting fMRI data. When the participant cooperated with the agent, the left region of the parietal lobe was activated, which correlated with self-report capability-based trust. Goodyear and colleagues (2017) also collected fMRI data and revealed greater strength of the lingus gyrus (LG) when participants accepted advice from a human compared to a machine. In Dong et al. (2015), differences in agents' human-like cues were assessed through behavioral performance and EEG results. No significant differences in behavior were found after experiencing a human-like cue from the agent; however, EEG data found significant differences in mean amplitude within event-related potential (ERP) patterns. Although neural signatures are not widely utilized, previous studies have shown promising results. This is largely attributed to their objectivity, which eliminates subjective biases recognized in self-report measures. However, collecting fMRI and EEG data can be strenuous due to expensive and obtrusive equipment along with the difficulty of analyzing the data. More specifically, studies have shown difficulty in accurately representing trust using EEG data due to the high volumes of noise within EEG signals (Subha et al., 2010). Due to the novelty of neural signatures in HATs, there is little uniformity in trust correlations associated with certain brain regions across studies. Identifying which brain regions correlate with trust will be needed for ongoing research in the HAT domain.

Galvanic Skin Response

Galvanic Skin Response (GSR) measures the change in skin conductivity, which fluctuates based on amount of sweat (Sharma et al., 2016). Khawaji et al. (2015) explored the relationship between GSR, cognitive workload, and trust. Pairs of participants were randomly assigned to one of two experimental conditions: enhanced trust and diminished trust. Participants completed four chat sessions in which they played an investment game and GSR data was captured and analyzed to determine average and peak GSR values. A cognitive workload questionnaire was also administered after gameplay for each session. All trust manipulations occurred prior to gameplay to establish baseline. Results revealed that when cognitive load was low, average GSR values and average GSR peaks were significantly lower during the high trust condition compared to during the low trust condition. These results suggest that during a low cognitive load scenario, GSR can be employed to measure trust (Khawaji, et al., 2015). Akash et al. (2018) aimed to identify if physiological metrics could be utilized to develop a classifier-based empirical trust model. Across 100 trials participants engaged in a driving simulator equipped with an obstacle detection sensor that would identify objects and generate a report for the participant to evaluate. The participant then chose to trust or distrust the report. Throughout the study, GSR and EEG data was recorded. The tonic (slow changing) and phasic (fast changing) components of the GSR values were captured and the Net Phasic Components were calculated. Results revealed the phasic component of GSR was a significant predictor of trust, and when combined with EEG data contributed to a model that predicted trust across participants with 70% accuracy. In Hald et al. (2020), participants completed a collaborative drawing task in which the participant held a piece of paper while the robot drew a square. Sudden changes in the robot's speed and motor noise were designed to elicit a decrease in trust, which was measured by self-report, motion tracking, and GSR. GSR results were not significant between conditions despite the questionnaire showing significance. This may be due to a small warm-up period and poor baseline data. GSR appears to show promise as a trust measure.

Summary

Based on this review, measures including vocal pitch and eye-tracking require additional research before they can be soundly recommended as measures of HAT trust. Furthermore, measurement of neural signatures requires incredibly specialized expertise and are often impractical to implement by most HAT researchers. However, there appears to be sufficient support, and feasibility of implementation, for the use of both multi-item, validated scales and single-item self-report measures, usage and context specific action behavioral measures, and cardiovascular and GSR physiological measures. The following section evaluates these measures and proposes a framework for utilizing these measures in isolation and together to capture the various aspects of human trust in agents.

MEASUREMENT FRAMEWORK FOR TRUST IN HUMAN AGENT TEAMS

Measure Evaluation Criteria

To evaluate effectiveness of each of these measures, we propose eight criteria, including six criteria stemming from Wickens & Holland (2000): sensitivity, diagnosticity, selectivity, obtrusiveness, temporal resolution, and reliability. To account for practical resources, we also add affordability and resource intensiveness as additional criteria.

Sensitivity

Sensitivity refers to the threshold of input required to register change in a construct's value (Vig & Walls, 2000; Wickens & Holland, 2000). Sensitivity, in this case, reflects how easily a measure responds to input from a person with respect to trust. A high sensitivity measure will register changes from very small inputs, while a low sensitivity measure will require significant inputs (Vig & Walls, 2000). In general, higher sensitivity is desirable because it better reflects the overall construct, especially for highly dynamic constructs that may subtly change over time.

Diagnosticity

Diagnosticity refers to how well a measure explains the drivers of a construct's score. A highly diagnostic measure better explains why scores on the construct are high or low (Wickens & Holland, 2000). For example, multi-dimensional trust scales explain different driverof trust allowing for increased diagnosticity (Jian, et al., 2000).

Selectivity

Selectivity is similar to discriminant validity and refers to a measure's ability to delineate between a construct of interest and other irrelevant constructs (Rönkkö & Cho, 2022; Wickens & Holland, 2000). Measures that demonstrate low selectivity are correlated with the intended construct as well as with other constructs, presenting an obstacle for measuring trust as individuals cannot be certain if their measure assesses trust or another construct such as stress or workload (Khawaji et al., 2015, Sharma et al., 2016). High selectivity measures would parse out correlations between the measure and the construct of interest, with less correlational overlap from other constructs.

Obtrusiveness

Obtrusiveness refers to how disruptive a measure is when capturing data (Wickens & Holland, 2000). A highly obtrusive measure is very noticeable and often interferes with a task because it requires attention. Obtrusiveness is undesirable because it reduces the measure's ability to capture a construct naturally without interference (Webb et al., 1999). Generally, measures are more obtrusive when they require more attention to collect data (e.g., questionnaires) or are more physically noticeable (e.g., on-body equipment).

Temporal Resolution

Temporal Resolution refers to how granular the data provided by a measure is with respect to time (Wickens & Holland, 2000). Temporal resolution can also be thought of as how often data points are recorded, or the bandwidth of time that a measure is capable of capturing (e.g. milliseconds, seconds, minutes; Huang et al., 2021). Higher temporal resolution is generally more desirable because it provides more data, however this may come at the expense of being inundated with too much data. With highly granular data, it becomes important to make intentional aggregations of periods of data to make sense of it. Temporal resolution is particularly important when capturing dynamic constructs that are expected to change quickly in small periods of time.

Reliability

Reliability refers to how consistently a measure reproduces similar data (Traub, 1994). Reliability can be thought of as how much the measure's data varies each time it is used. Highly reliable measures will reproduce similar data when used repeatedly under the same conditions. Higher reliability is always desirable.

Affordability
Realistically, the feasibility of using a measure depends on the resources available. We include affordability to acknowledge that the accessibility of a measure is important when considering whether or not to employ it. Actual costs of a measure fluctuate depending on the vendor and type of measure, therefore we provide general affordability ratings. Although these thresholds are subjective, the hope is to provide a relative price gauge.

Resource Intensiveness
In addition to affordability, the amount of time and energy required to implement a measure and analyze its resulting data is an important consideration. Resource Intensiveness refers to the time and effort demands to effectively implement a measure and analyze associated data. For example, physiological measures can be incredibly resource intensive to implement due to the time and training necessary to learn the technical set-up and configuration as well as to conduct the analyses necessary to make meaning of the data. Conversely, self-report measures tend to be very non-resource intensive as they have already been validated and produce direct scores that are easy to analyze.

Measure Evaluation

Table 1 reflects an informal evaluation of the measures using the proposed criteria. Plus signs denote positive attributes related to the measurement evaluation criteria; double plus signs indicate very high levels of the positive attribute; minus signs denote negative attributes. As is evident in Table 1, there is currently no single best measure that can be recommended to effectively capture HAT trust. Various measures have different advantages and disadvantages. Validated self-report scales have high levels of diagnosticity given their multidimensional nature, which explains key aspects that contribute to trust. These are the only measures to date that capture these key aspects separately. They are also high in selectivity as they have been demonstrated to have divergent validity from other related constructs (Flake et al., 2017), high reliability (Körber., 2018), and are also low-cost and non-resource intensive to implement and analyze. However, self-reports tend to be highly obtrusive to the task as respondents must be interrupted to respond. They are typically administered once after the task and are thus low in temporal resolution, making it hard to assess dynamic changes in trust. In contrast, single-item self-report measures maintain selectivity, cost effectiveness, and non-resource-intensiveness of self-report scales, while increasing temporal resolution as they can be injected during tasks through a quick pop-up. This is still obtrusive to the task, although less so, and single item measures do not provide the diagnosticity and reliability of multi-item scales.

Table 1. Measure Evaluation for Trust in HATs

	Self Report Measures		***Behavioral Measures***		***Physiological Measures***	
	Validated Scales	**Single Item**	**Usage Behaviors**	**Context-Specific Actions**	**Cardio-vascular**	**GSR**
Sensitivity	+	-	-	+	+*	+*
Diagnosticity	++	-	-	+	-	-
Selectivity	++	+	-	+	-	-
Unobtrusiveness	-	+	++	++	+	+
Temporal Resolution	-	-	+	+	++	++
Reliability	++	-	-	+	+	+
Affordability	++	++	+	+	-	-
Non-Resource Intensive	++	++	+	-	-	-

* Although research on cardiovascular and GSR measures as indicators of trust have not definitively shown these measures are specifically sensitive to changes in trust, this may be a research limitation rather than a limitation of the measures themselves. Cardiovascular and GSR measures are generally sensitive to input and have the potential to be sensitive to changes in trust. Further research is needed to verify this.

Behavioral measures provide the least obtrusive means of measuring trust and moderate to high levels of temporal resolution, depending on the behavior being captured (e.g., response to an event may be low frequency, but body language can fluctuate rapidly). However, because attitudes and behaviors are distinct (Ajzen, 1991), usage behaviors may not reflect trust attitudes exclusively and thus have low levels of sensitivity and selectivity. Usage behaviors are also low on diagnosticity as the behaviors themselves do not explain trust scores (although sometimes performance context can increase this), and reliability is questionable until validated as these measures must be carefully developed for each specific context. Context-specific measures, if designed well, can provide increased sensitivity, diagnosticity and selectivity as scenarios can be designed to elicit specific behaviors indicative of trust. Usage measures in particular can typically be captured without excessive cost or resources through use of observer checklists and system collected response measures. However, for well-designed context-specific action measures, there can be intense resources required for both scenario design and post-hoc coding of behaviors and communications from video transcripts.

Finally, physiological measures including cardiovascular and GSR measures, provide extremely high levels of temporal resolution as they can be captured second-by-second, and in a manner fairly unobtrusive to the task (although some argue that some of the sensors can be physically obtrusive). They also have the potential to be fairly sensitive as has been demonstrated in their ability to capture small changes in workload (Akash, et al., 2018) and sensor technology is becoming increasingly reliable. The biggest disadvantage of physiological measures is that sensor technology can be extremely expensive, especially if there is a need for software to support analysis by a non-expert, and resource intensive to learn to administer and analyze. Physiological measures also have low levels of selectivity as many different physical and emotional states can cause fluctuations, and low levels of diagnosticity because you do not know why they are fluctuating (although performance context can assist with this as well).

Measure Selection and Implementation

Given these advantages and disadvantages, decisions regarding which measure of trust to utilize depend on the research goals and the HAT performance context. Further, it is often necessary to utilize multiple measures to triangulate in on trust at different times and increase construct validity. To assist in measure selection, we present three categorical use-cases that vary based on (a) level of interactivity with the agent, (b) how dynamically we anticipate human trust in the agent varying (i.e., static, moderately dynamic, highly dynamic), and (c) the research goals. In the first use case, the goal is to understand what influences trust in an agent when there are few, if any, interactions with the agent (e.g., vignette-based research) and therefore human trust in the agent is unlikely to fluctuate over time. In the second use case, the goal is to study factors that influence trust when there are low to moderate levels of interaction with the agent and thus, human trust in the agent may fluctuate to a degree, but the focus is not on dynamic changes in trust. Finally, in the third use case the goal is to study dynamic changes in trust over time in a HAT performance context that is highly interactive, with less focus on why trust changes as this is controlled with experimental design. Table 2 illustrates which measures to implement in each of these use cases. It should be noted that there are additional considerations in measure selection that were not the focus of this effort. For example, administering measures in a multi-level context with multiple agents and humans can increase the obtrusiveness of the self-report measures as each entity must be rated, and the cost and resources necessary to collect physiological data increase since multiple sensors must be purchased and multiple data streams analyzed. Further, the practicality of using measures in an applied context (e.g., real-world military operations) may be of additional interest if collecting real-time data during performance is important so that interventions can be implemented.

Table 2. Measure Implementation Guidance with Use Case Scenarios

	Self-Report		***Behavioral***		***Physiological***	
	Validated Scales	**Single Item**	**Usage Behaviors**	**Context-Specific Actions**	**Cardio**	**GSR**
1. Low interactivity, static trust	X		X			
2. Moderate interactivity, moderately dynamic trust	X	X	X	X		
3. High interactivity, highly dynamic trust		X	X	X	X	X

Low Interactivity, Static Trust

When there is low interactivity, few expected fluctuations in trust over time, and the goal is to understand what influences trust based on manipulation of various variables (e.g., in a vignette-based scenario), validated self–report scales provide a great tool to capture a snapshot of the multidimensional nature of trust and shed light on factors that influence trust. To bolster this understanding and identify how trust attitudes may impact human behaviors, usage measures can be utilized alongside scales, providing a more robust understanding of *why* a participant trusts an agent. To implement these measures, the researcher should administer validated scales before and after each scenario. Usage can be captured by having participants report their response to the vignette. For example, in Reig et al. (2021)'s vignette study, participants used a self-report scale to indicate their trust in an agent after viewing one of four videos that reflect different trust repair strategies enacted by the agent. In this low-interactivity scenario, context specific action measures and physiological measures are not recommended as the researcher is not examining trust fluctuations.

Moderate Interactivity, Moderately Dynamic Trust

When there is moderate interactivity, small fluctuations in trust, and the goal is to understand factors that influence trust, such as in a brief simulation-based scenario, validated self–report scales continue to provide a great tool to capture a snapshot of multidimensional trust. However, higher temporal resolution is needed as trust is expected to fluctuate, and it is preferable to minimally disrupt HAT interactions. Single-item self-report measures can be layered on through brief pauses that are minimally obtrusive. Behavioral measures can also be captured unobtrusively to gain insight on how trust changes after an intervention or scenario event. Coupling behavioral measures with self-report measures increases temporal resolution while being less disruptive. As usage behaviors are typically used for low-level HAT tasks where the operator may choose to engage or disengage the automation, applying use/disuse criterion may be challenging to tease apart trust from required task-based interactions. This reliance-trust behavioral confound is less prevalent if tasks promote interactivity and cooperation amongst teammates with additional options to use/disuse an agent's supplementary capabilities, as well as provide opportunities to capture context-specific behaviors indicative of trust. For example, Azevedo-Sa et al. (2021) examined how automated driving systems' (ADS) reliability and road visibility influence trust. Trust was assessed through self-report and behavioral measures such that high frequency of emergency braking by participants inferred low trust. Validated scales could also be administered before and after the task or between conditions to provide diagnostic information that explains a human operator's trust.

High Interactivity, Highly Dynamic Trust

Finally, when there is high interactivity, highly dynamic fluctuations in trust over time, and the goal is to understand these dynamic fluctuations, such as in a more extensive simulation-based scenario, validated self–report scales will provide limited assistance given their low temporal resolution. In this use case, capturing responses to single-item trust scales, usage, and context-specific behavioral measures will allow for study of trust dynamics over time with minimal task interruption. Further, physiological measures will increase temporal resolution and sensitivity, allowing for more granular measure of changes in trust. This is desirable as trust is thought to fluctuate rapidly in response to trust-related events like violation and repair (Glikson & Woolley, 2020). For example, Mitkidis et al. (2015) used heart rate synchrony as a predictor of trust in an experimental setting where participants engaged with an agent for four consecutive sessions that each lasted ten minutes. Using heart rate data collected both within and across the four 10- minute sessions, it was found that heart rate was more synchronized in the trusting condition.

FUTURE RESEARCH

As HATs are instantiated into more complex environments such as teams with multiple humans interacting with multiple types of agents, accurately assessing trust in HATs is imperative for future research. Previous research has illustrated conflicting results amongst various trust measures, demonstrating a need for numerous measures to gather precise results. Future research should examine measurement of trust in multiple agents and elaborate on the implications of having multiple trust targets. Developing a solidified assessment of trust is important for differentiating participant trust levels between each agent and overall team trust. Another important element to consider is how to capture trust over time. Since relationship dynamics differ before and after trust violations and repair events, understanding how trust is built, broken, and repaired is critical. Future research should incorporate multiple measures in concert to understand the trajectory of trust across these different phases of interaction, especially in heterogeneously composed HATs (e.g., multi-human, multi-agent, or multi-team HATs).

There are also multiple future research opportunities to improve upon these measures of trust. For example, as HAT roles become more interdependent, activating automation will no longer be an option for human operators since automation will already be deeply embedded into the system. This will inherently affect the usage criterion since humans may no longer have the option to activate or deactivate automation. Therefore, as HATs continue to progress into more elaborate environments, upcoming behavioral measures should also evolve past automation activation or supervisory-related tasks to make usage more relevant to advancing HAT missions, such as multi-agent teams. For context-specific behaviors, no research to date has created a generalizable framework for identifying and solidifying these. Therefore, it is recommended that emerging context-specific behaviors are validated and supported by prior literature and a framework for identifying these behaviors be developed to assist in supporting the choice of behaviors. Furthermore, future research in eye-tracking measures should develop individualized eye-tracking baselines through machine learning models to increase validity and uniformity, similar to Lu and Sarter's (2020) technique. Future research is also needed to establish the relationship with trust and various vocal properties (e.g., pitch, speed).

To assess the validity of using numerous measures in tandem, developing a series of different testbed/tasking solutions with the same measurements is crucial in assessing the validity of this framework. Although trust is an elusive construct, triangulation from numerous measurements can combat this issue and help us better understand the development and degradation of trust in a realistic HAT environment.

ACKNOWLEDGEMENTS

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-21-1-0294. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the United States Air Force.

REFERENCES

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179-211. https://doi.org/10.1080/08870446.2011.613995

Akash, K., Hu, W.-L., Jain, N., & Reid, T. (2018). A classification model for sensing human trust in machines using EEG and GSR. *ACM Transactions on Interactive Intelligent Systems*, *8*(4), 1–20. https://doi.org/10.1145/3132743

Alarcon, G. M., Lyons, J. B., Christensen, J. C., Klosterman, S. L., Bowers, M. A., Ryan, T. J., Jessup, S. A., Wynne, K. T. (2018). The effect of propensity to trust and perceptions of trustworthiness on trust behaviors in dyads. *Behavior Research Methods*, *50*(5), 1906–1920. https://doi.org/10.3758/s13428-017-0959-6

Albayram, Y., Jensen, T., Khan, M. M. H., Fahim, M. A. A., Buck, R., & Coman, E. (2020, November). Investigating the effects of (empty) promises on human-automation interaction and trust repair. *Proceedings of the 8th International Conference on Human-Agent Interaction*, 6-14. https://doi.org/10.1145/3406499.3415064

Arnold, H. J., & Feldman, D. C. (1981). Social desirability response bias in self-report choice situations. *Academy of Management Journal*, *24*(2), 377-385. https://doi.org/10.5465/255848

Azevedo-Sa, H., Zhao, H., Esterwood, C., Yang, X. J., Tilbury, D. M., & Robert Jr, L. P. (2021). How internal and external risks affect the relationships between trust and driver behavior in automated driving systems. *Transportation Research Part C: Emerging Technologies*, *123*, 102973. Advance online publications. https://doi.org/10.1016/j.trc.2021.102973

Bagheri, N., & Jamieson, G. A. (2004, October). The impact of context-related reliability on automation failure detection and scanning behaviour. In *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)* (Vol. 1, pp. 212-217). IEEE. https://doi.org/10.1109/ICSMC.2004.1398299

Bainbridge, W. A., Hart, J. W., Kim, E. S., & Scassellati, B. (2011). The benefits of interactions with physically present robots over video-displayed agents. *International Journal of Social Robotics*, *3*(1), 41-52. https://doi.org/10.1007/s12369-010-0082-7

Bales, G., & Kong, Z. (2017, October). Neurophysiological and behavioral studies of human-swarm interaction tasks. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 671-676). IEEE. https://doi.org/10.1109/SMC.2017.8122684

Buchholz, V., Kulms, P., & Kopp, S. (2017). It's (not) your fault! Blame and trust repair in human-agent cooperation. In *Kognitive Systeme 2017.* https://doi.org/10.17185/duepublico/44538

Chen, J. Y., & Barnes, M. J. (2014). Human–agent teaming for multirobot control: A review of human factors issues. *IEEE Transactions on Human-Machine Systems*, *44*(1), 13-29. https://doi.org/10.1109/THMS.2013.2293535

de Fine Licht, K., & Brülde, B. (2021). On defining “reliance” and “trust”: Purposes, conditions of adequacy, and new definitions. *Philosophia*, *49*(5), 1981–2001. https://doi.org/10.1007/s11406-021-00339-1

Demir, M., McNeese, N. J., & Cooke, N. J. (2019). The Evolution of human-autonomy teams in remotely piloted aircraft systems operations. *Frontiers in Communication*, *4*, [50], 1-12. https://doi.org/10.3389/fcomm.2019.00050

Dong, S. Y., Kim, B. K., Lee, K., & Lee, S. Y. (2015, October). A preliminary study on human trust measurements by EEG for human-machine interactions. *Proceedings of the 3rd International Conference on Human-Agent Interaction*, 265-268. https://doi.org/10.1145/2814940.2814993

Elkins, A. C., & Derrick, D. C. (2013). The sound of trust: Voice as a measurement of trust during interactions with embodied conversational agents. *Group Decision and Negotiation*, *22*(5), 897-913. https://doi.org/10.1007/s10726-012-9339-x

Erickson, A., Norouzi, N., Kim, K., LaViola, J. J., Bruder, G., & Welch, G. F. (2020). Effects of depth information on visual target identification task performance in shared gaze environments. *IEEE Transactions on Visualization and Computer Graphics*, *26*(5), 1934-1944. https://doi.org/10.1109/TVCG.2020.2973054

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370-378. https://doi.org/10.1177/1948550617693063

Fratczak, P., Goh, Y. M., Kinnell, P., Justham, L., & Soltoggio, A. (2021). Robot apology as a post-accident trust-recovery control strategy in industrial human-robot interaction. *International Journal of Industrial Ergonomics*, *82*, 103078. https://doi.org/10.1016/j.ergon.2020.103078

Freedy, A., DeVisser, E., Weltman, G., & Coeyman, N. (2007, May). Measurement of trust in human-robot collaboration. In *2007 International symposium on collaborative technologies and systems* (pp. 106-114). IEEE. https://doi.org/10.1109/CTS.2007.4621745

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627-660. https://doi.org/10.5465/annals.2018.0057

Goodyear, K., Parasuraman, R., Chernyak, S., de Visser, E., Madhavan, P., Deshpande, G., & Krueger, F. (2017). An fMRI and effective connectivity study investigating miss errors during advice utilization from human and machine agents. *Social Neuroscience*, *12*(5), 570-581. https://doi.org/10.1080/17470919.2016.1205131

Hald, K., Rehmn, M., & Moeslund, T. B. (2020, October). Human-robot trust assessment using motion tracking & galvanic skin response. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 6282-6287). IEEE. https://doi.org/10.1109/IROS45743.2020.9341267

Hergeth, S., Lorenz, L., Krems, J. F., & Toenert, L. (2015, June). Effects of take-over requests and cultural background on automation trust in highly automated driving. *Proceedings of the 8th International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design* (Vol. 2015, pp. 330-336). https://doi.org/10.17077/drivingassessment.1591

Hill, A. D., White, M. A., & Wallace, J. C. (2014). Unobtrusive measurement of psychological constructs in organizational research. *Organizational Psychology Review*, *4*(2), 148-174. https://doi.org/10.1177/2041386613505613

Huang, L., Cooke, N. J., Gutzwiller, R. S., Berman, S., Chiou, E. K., Demir, M., & Zhang, W. (2021). Distributed dynamic team trust in human, artificial intelligence, and robot teaming. In C. S. Nam and J. B. Lyons (Eds.), *Trust in human-robot interaction* (pp. 301-319). Academic Press. https://doi.org/10.1016/B978-0-12-819472-0.00013-7

Jian, J. Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53-71. https://doi.org/10.1207/S15327566IJCE0401_04

Johnson, C. J., Demir, M., McNeese, N. J., Gorman, J. C., Wolff, A. T., & Cooke, N. J. (2021). The impact of training on human–autonomy team communications and trust calibration. *Human Factors*, 1-17. Advance online publication. https://doi.org/10.1177/00187208211047323

Khalid, H. M., Helander, M. G., & Lin, M. H. (2021). Determinants of trust in human-robot interaction: Modeling, measuring, and predicting. In C. S. Nam and J. B. Lyons (Eds.), *Trust in Human-Robot Interaction* (pp. 85-121). Academic Press. https://doi.org/10.1016/B978-0-12-819472-0.00004-6

Khawaji, A., Zhou, J., Chen, F., & Marcus, N. (2015, April). Using galvanic skin response (GSR) to measure trust and cognitive load in the text-chat environment. *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 1989-1994).

Kohn, S. C., Kluck, M., & Shaw, T. (2020). A Brief review of frequently used self-report measures of trust in automation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *64*(1), 1436–1440. https://doi.org/10.1177/1071181320641342

Kohn, S. C., De Visser, E. J., Wiese, E., Lee, Y. C., & Shaw, T. H. (2021). Measurement of trust in automation: A narrative review and reference guide. *Frontiers in Psychology*, *12*. 1-23. https://doi.org/10.3389/fpsyg.2021.604977

Körber, M. (2018, August). Theoretical considerations and development of a questionnaire to measure trust in automation. In *Congress of the International Ergonomics Association* (pp. 13-30). Springer, Cham. https://doi.org/10.1007/978-3-319-96074-6_2

Kozlowski, S. W. J. (2015). Advancing research on team process dynamics: Theoretical, methodological, and measurement considerations. *Organizational Psychology Review, 5*, 270-299. https://doi.org/10.1177/2041386613505613

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243-1270. https://doi.org/10.1080/00140139208967392

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50-80. https://journals.sagepub.com/doi/abs/10.1518/hfes.46.1.50_30392

Lee, S. Y. (2018). *Impact of human like cues on human trust in machines: Brain imaging and modeling studies for human-machine interactions*. Korea Advanced Institute of Science and Technology.

https://apps.dtic.mil/sti/citations/AD1046152

Lu, Y., & Sarter, N. (2020). Modeling and inferring human trust in automation based on real-time eye tracking data. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 64*(1), 344–348. https://doi.org/10.1177/1071181320641078

Marathe, A., Brewer, R., Kellihan, B., & Schaefer, K. E. (2020). Leveraging wearable technologies to improve test & evaluation of human-agent teams. *Theoretical Issues in Ergonomics Science*, 1-21. https://doi.org/10.1080/1463922X.2019.1697389

Mayer, R. C., & Davis, J. H. (1999). The effect of the performance appraisal system on trust for management: A field quasi-experiment. *Journal of Applied Psychology*, *84*(1), 123. https://doi.org/10.1037/0021-9010.84.1.123

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709-734. https://doi.org/10.2307/258792

Menon, V., & Crottaz-Herbette, S. (2005). Combined EEG and fMRI studies of human brain function. *International Review of Neurobiology*, *66*(05), 291-321.

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors, 50* (2), 194-210.

Mitkidis, P., McGraw, J. J., Roepstorff, A., & Wallot, S. (2015). Building trust: Heart rate synchrony and arousal during joint action increased by public goods game. *Physiology & Behavior*, *149*, 101-106. https://doi.org/10.1016/j.physbeh.2015.05.033

Morgado, F. F., Meireles, J. F., Neves, C. M., Amaral, A., & Ferreira, M. E. (2017). Scale development: Ten main limitations and recommendations to improve future research practices. *Psicologia: Reflexão e Crítica*, *30,* 1-20. https://doi.org/10.1186/s41155-016-0057-1

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-machine Studies*, *27*(5-6), 527-539. https://doi.org/10.1016/S0020-7373(87)80013-5

Niculescu, A., van Dijk, B., Nijholt, A., Li, H., & See, S. L. (2013). Making social robots more attractive: The effects of voice pitch, humor and empathy. *International Journal of Social Robotics*, *5*(2), 171–191. https://doi.org/10.1007/s12369-012-0171-x

Otto, R. P. (2016). *Small unmanned aircraft systems (SUAS) flight plan: 2016-2036. Bridging the gap between tactical and strategic*. United States Air Force. https://apps.dtic.mil/sti/citations/AD1013675

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230-253. https://doi.org/10.1518/001872097778543886

Podsakoff, P. M., & Organ, D. W. (1986). Self-reports in organizational research: Problems and prospects. *Journal of Management*, 12(4), 531-544. https://doi.org/10.1177/014920638601200408

Rönkkö, M., & Cho, E. (2022). An updated guideline for assessing discriminant validity. *Organizational Research Methods*, *25*(1), 6-14. https://doi.org/10.1177/1094428120968614

Schaefer, K. E., Perelman, B. S., Gremillion, G. M., Marathe, A. R., & Metcalfe, J. S. (2021). A roadmap for developing team trust metrics for human-autonomy teams. In *Trust in Human-Robot Interaction* (pp. 261-300). Academic Press. https://doi.org/10.1016/B978-0-12-819472-0.00012-5

Schaefer, K. E., Straub, E. R., Chen, J. Y., Putney, J., & Evans III, A. W. (2017). Communicating intent to develop shared situation awareness and engender trust in human-agent teams. *Cognitive Systems Research*, *46*, 26-39. https://doi.org/10.1016/j.cogsys.2017.02.002

Sharma, M., Kacker, S., & Sharma, M. (2016). A brief introduction and review on galvanic skin response. *Int J Med Res Prof*, *2*(6), 13-17. https://doi.org/10.21276/ijmrp.2016.2.6.003

Subha, D. P., Joseph, P. K., Acharya U, R., & Lim, C. M. (2010). EEG signal analysis: A survey. *Journal of Medical Systems*, *34*(2), 195-212. https://doi.org/10.1007/s10916-008-9231-z

Taylor, J. B. (1961). What do attitude scales measure: The problem of social desirability. *Journal of Abnormal and Social Psychology*, *62*(2), 386-390. https://doi.org/10.1037/h0042497

Tolston, M. T., Funke, G. J., Alarcon, G. M., Miller, B., Bowers, M. A., Gruenwald, C., & Capiola, A. (2018). Have a heart: Predictability of trust in an autonomous agent teammate through team-level measures of heart rate synchrony and arousal. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *62*(1), 714–715. https://doi.org/10.1177/1541931218621162

Traub, R. E. (1994). *Reliability for the social sciences: Theory and applications* (Vol. 3). Sage.

Van Lange, P. A. M., Finkenauer, C., Popma, A., & van Vugt, M. (2011). Electrodes as social glue: Measuring heart rate promotes giving in the trust game. *International Journal of Psychophysiology*, *80*(3), 246–250. https://doi.org/10.1016/j.ijpsycho.2011.03.007

Vig, J. R., & Walls, F. L. (2000, June). A review of sensor sensitivity and stability. In *Proceedings of the 2000 IEEE/EIA International Frequency Control Symposium and Exhibition (Cat. No. 00CH37052)* (pp. 30-33). IEEE. https://doi.org/10.1109/FREQ.2000.887325

Waldman, D. A., Wang, D., Stikic, M., Berka, C., & Korszen, S. (2015). Neuroscience and team processes. In

Organizational Neuroscience (Vol. 7, pp. 277-294). Emerald Group Publishing Limited. https://doi.org/10.1108/S1479-357120150000007012

Wang, N., Pynadath, D. V., Hill, S. G., & Merchant, C. (2017, August). The dynamics of human-agent trust with POMDP-generated explanations. In *International Conference on Intelligent Virtual Agents* (pp. 459-462). Springer, Cham. https://doi.org/10.1007/978-3-319-67401-8_58

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1999). *Unobtrusive Measures* (Vol. 2). Sage Publications.

Wickens, C. D., & Hollands, J. G. (2000). *Engineering psychology and human performance*. Prentice Hall.

Wright, J. L., Quinn, S. A., Chen, J. Y. C., & Barnes, M. J. (2014). Individual Differences in Human-Agent Teaming: An Analysis of Workload and Situation Awareness through Eye Movements. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *58*(1), 1410–1414. https://doi.org/10.1177/1541931214581294

Yu, K., Berkovsky, S., Taib, R., Zhou, J., & Chen, F. (2019, March). Do I trust my machine teammate? An investigation from perception to decision. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (pp. 460-468). https://doi.org/10.1145/3301275.3302277