

<b>Fiscal Year:</b>	FY 2023	<b>Task Last Updated:</b>	FY 03/21/2023
<b>PI Name:</b>	Lee, John Ph.D.		
<b>Project Title:</b>	HCAAM VNSCOR: Conversation Analysis to Measure and Manage Trust in Virtual Assistants		
<b>Division Name:</b>	Human Research		
<b>Program/Discipline:</b>			
<b>Program/Discipline-- Element/Subdiscipline:</b>			
<b>Joint Agency Name:</b>	<b>TechPort:</b>	No	
<b>Human Research Program Elements:</b>	(1) <b>HFBP</b> :Human Factors & Behavioral Performance (IRP Rev H)		
<b>Human Research Program Risks:</b>	(1) <b>HSIA</b> :Risk of Adverse Outcomes Due to Inadequate Human Systems Integration Architecture		
<b>Space Biology Element:</b>	None		
<b>Space Biology Cross-Element Discipline:</b>	None		
<b>Space Biology Special Category:</b>	None		
<b>PI Email:</b>	<a href="mailto:jdlee@engr.wisc.edu">jdlee@engr.wisc.edu</a>	<b>Fax:</b>	FY
<b>PI Organization Type:</b>	UNIVERSITY	<b>Phone:</b>	608-890-3168
<b>Organization Name:</b>	University of Wisconsin, Madison		
<b>PI Address 1:</b>	Department of Industrial and Systems Engineering		
<b>PI Address 2:</b>	1513 University Ave		
<b>PI Web Page:</b>			
<b>City:</b>	Madison	<b>State:</b>	WI
<b>Zip Code:</b>	53706-1539	<b>Congressional District:</b>	2
<b>Comments:</b>			
<b>Project Type:</b>	Ground	<b>Solicitation / Funding Source:</b>	2017-2018 HERO 80JSC017N0001-BPBA Topics in Biological, Physiological, and Behavioral Adaptations to Spaceflight. Appendix C
<b>Start Date:</b>	04/15/2019	<b>End Date:</b>	06/01/2024
<b>No. of Post Docs:</b>		<b>No. of PhD Degrees:</b>	
<b>No. of PhD Candidates:</b>	1	<b>No. of Master' Degrees:</b>	1
<b>No. of Master's Candidates:</b>	1	<b>No. of Bachelor's Degrees:</b>	1
<b>No. of Bachelor's Candidates:</b>	1	<b>Monitoring Center:</b>	NASA JSC
<b>Contact Monitor:</b>	Whitmire, Alexandra	<b>Contact Phone:</b>	
<b>Contact Email:</b>	<a href="mailto:alexandra.m.whitmire@nasa.gov">alexandra.m.whitmire@nasa.gov</a>		
<b>Flight Program:</b>			
<b>Flight Assignment:</b>	NOTE: End date changed to 06/01/2024 per A. Beitman/HRP (Ed., 4/18/23) NOTE: End date changed to 04/14/2023 per S. Huppman/HRP and NSSC information (Ed., 3/20/2020) NOTE: End date changed to 3/31/2020 per NSSC information (Ed., 1/22/2020)		
<b>Key Personnel Changes/Previous PI:</b>	Jerri Stephenson is no longer working in the project. James S. Garrett is now a Co-Investigator.		
<b>COI Name (Institution):</b>	Cross, Ernest Ph.D. ( NASA Johnson Space Center ) Garrett, James ( NASA )		
<b>Grant/Contract No.:</b>	80NSSC19K0654		
<b>Performance Goal No.:</b>			
<b>Performance Goal Text:</b>			

**Task Description:**

This task is part of the Human Capabilities Assessments for Autonomous Missions (HCAAM) Virtual NASA Specialized Center of Research (VNSCOR).

The goal of this research is to develop conversation analysis to measure and mitigate inappropriate trust in virtual assistants. These trust measurements will guide system design, particularly the multimodal interactions and mode switching, as well as how to mitigate over trust and trust recovery. We will use conversation analysis to measure trust at multiple time-scales from real-time interactions to longitudinal monitoring of trust over a long duration exploration mission.

Conversation analysis provides a promising, but relatively unexplored approach to measuring trust. We propose a conversation analysis at the micro, meso, and macro levels which includes not just the words, but also pauses and facial expressions. Specifically, at the micro-level, conversation elements include voice inflections, pauses between words and keystrokes, gaze shifts, and facial expressions. The meso-level analysis includes words exchanged during interactions with the virtual assistant along with other team interactions as they relate to the automation. At the macro level, conversational analysis considers interaction time, interaction effort, frequency of interaction, turn-taking, bargaining in tendency, and whether it is the person or the virtual assistant who initiates the interaction. Additionally, prior research into conversational analysis indicates there are novel ways of managing or calibrating trust through the presentation of information, e.g., manipulating the tone and cadence of the system when using speech and through facial expressions (Nass & Brave, 2005; DeSteno et al., 2012).

Due to time delays in communication, long duration exploration missions will require greater crew autonomy and greater reliance on automation. For this approach to work trust calibration needs to be engineered into the system. Trust is a critical construct that mediates how well human operators use automated systems, such as virtual assistants, that provide decision support. Trust affects people's willingness to rely on automated systems in situations that have a degree of uncertainty and risk. Trust strongly affects the effectiveness of human-agent collaboration, particularly in the willingness to accept suggestions from a virtual assistant. Knowing whether or not to trust automation can be further complicated by lack of sleep, workload, task risk, and task complexity. Moreover, as we continue to push the limits of intelligent systems and rely on them more as decision aids trust calibration (i.e., operator trust is at a level which matches the automation's capabilities) becomes essential to mission execution.

Appropriate calibration of trust requires matching the operator's trust to the virtual assistant's current capabilities. Calibration of trust is not something that can happen once, but must occur throughout the life cycle of the interaction between operator and automated system (Hoffman et al., 2009). Trust is a dynamic construct that continuously increases and decreases due to a number of factors, primary the performance of the automated system, i.e., higher performance leads to higher trust and vice versa. Although much effort focuses on creating more capable and trustworthy automation, less effort has considered the equally important consideration of creating trustable automation. Trustable automation is automation that is understandable and that naturally promotes calibrated trust. Therefore, we aim to create trustable automation by continuously measuring operators' trust unobtrusively and in real-time, and then use this measure to guide the virtual agent to employ one or more countermeasures to calibrate trust and improve human-system performance.

**References**

DeSteno D, Breazeal C, Frank RH, Pizarro D, Baumann J, Dickens L, Lee JJ. Detecting the trustworthiness of novel partners in economic exchange. Psychol Sci. 2012 Dec;23(12):1549-56. <http://doi.org/>; PubMed [PMID: 23129062](https://pubmed.ncbi.nlm.nih.gov/23129062/)

Hoffman RR, Lee JD, Woods DD, Shadbolt N, Miller J, Bradshaw JM. The dynamics of trust in cyberdomains. IEEE Intelligent Systems. 2009 Nov-Dec;24(6):5-11. [https://](https://doi.org/10.1109/IS.2009.5399999)

Nass C, Brave S. Wired for Speech : How Voice Activates and Advances the Human-Computer Relationship. Cambridge, MA: MIT Press, 2005.

**Rationale for HRP Directed Research:**

The outcomes of this research will make two important contributions to the overall HCAAM VNSCOR effort. First, it will promote more effective interactions and acceptance of virtual assistants. Second, it will provide new analytic techniques for understanding how people work with automated agents as team members.

Virtual assistants and other types of agents enabled by artificial intelligence represent an important opportunity to extend human capabilities, but only if they are accepted and trusted appropriately. If people trust the virtual assistant too much they will rely on it in situations that exceed its capability, and if they trust it too little they will fail to engage it when it could benefit the team. One pathway towards appropriate trust is to make the virtual assistant more trustworthy: increase its technical capabilities to accommodate any situation. Another approach is to make it more trustable: communicate its capability and allow its capability to be challenged in its interactions with people. Such trustable technology requires three important advances to the state of knowledge in the field:

1. An ability to ascertain how much people currently trust the technology
2. An ability to convey uncertainty and its capability, particularly as part of conversational interactions
3. Interaction affordances that provide the opening for people to assess the capability of the assistant, particularly as part of conversational interactions.

**Research Impact/Earth Benefits:**

These three advances for trustable technology require the development of new analytic techniques for understanding human interaction with automated teammates. Real-time, unobtrusive measures of trust represent a particularly valuable, but challenging measure to develop. Trust is most often measured with ratings and indirectly through people's decision to rely on automation, which are obtrusive not diagnostic. Conversation and text-based interactions offer a promising, but unexplored way to assess trust. Text analysis has a 50-year history in domains as diverse as psycholinguistics and cognitive science, and more recently natural language processing, affective state assessment, and sentiment analysis. Building on the foundation of text analysis makes it possible for this research to immediately contribute to data analysis of previous and future studies of automation-human teaming, and to contribute to the foundation of conversational agent design.

NOTE: For full citation information on the published papers listed below, please see the Cumulative Bibliography (Ed., 5/22/23).

#### Project status

. Data from the first controlled study has been completed. . Data analysis from the second controlled study has been completed. . A Human Factors and Ergonomics Society (HFES) conference paper on developing a measure of trust based on conversation has been accepted for publication (Li et al., 2022). . An HFES conference paper describing a cognitive simulation model of interdependent agents has been accepted for publication (Li & Lee, 2022). . A Human Factors journal paper on developing a measure of trust based on conversations has been accepted for publication (Li, Erickson, et al., 2023). . A paper has been submitted to the International Journal of Human-Computer Interaction on modeling trust dynamics. This paper has been provisionally accepted for publication pending minor revisions (Li, Amudha, et al., 2023). . Data collection from the NASA Human Exploration Research Analog (HERA) testbed has continued. . Preliminary data analysis of the HERA data has started.

The following summaries describe three specific research accomplishments and the associated papers.

#### Conversational measures of trust

We have analyzed the data from a controlled experiment and created a machine-learning model that estimates trust in an agent from the lexical and acoustical features of conversations with that agent. The objective of this study was to estimate trust from conversations using both lexical and acoustic data. As NASA moves to long-duration space exploration operations, the increasing need for cooperation between humans and virtual agents requires real-time trust estimation by virtual agents. Measuring trust through conversation is a novel, yet unexplored approach.

A 2 (reliability)  $\times$  2 (cycles)  $\times$  3 (events) within-subject study on habitat system maintenance was designed to elicit various levels of trust in a conversational agent. Participants had trust-related conversations with the conversational agent at the end of each decision-making task. To estimate trust, subjective trust ratings were predicted using machine learning models trained on three types of conversational features (i.e., lexical, acoustic, and combined). After training, model inference was performed using variable importance and partial dependence plots. Results showed that a random forest algorithm, trained using the combined lexical and acoustic features, was the highest-performing algorithm for predicting trust in the conversational agent ( $R^2_{adj} = 0.71$ ). The most important predictor variables were a combination of lexical and acoustic cues: average sentiment considering valence shifters and the mean of formants, Mel-frequency cepstral coefficients (MFCC), and standard deviation of the fundamental frequency. Precise trust estimation from conversation requires lexical cues and acoustic cues. We further identified conversational features as mediators between an exposure (i.e., reliability) and a response variable (i.e., trust). Following the mediation analysis criteria, we identified a partial mediation that occurred between reliability on trust via conversational features with a Sobel test for the indirect effect,  $z = -5.86$ ,  $p < .001$ . This suggests that reliability influences how people communicate as an underlying mechanism, which in turn influences people's trust. The proportion of the effect of the reliability on trust that goes through the mediator is 0.17. These results show the possibility of using conversational data to measure trust, and potentially other dynamic mental states, unobtrusively and dynamically. These results have been accepted for publication in the journal, Human Factors, under the title of: It's Not Only What You Say, But Also How You Say It: Machine Learning Approach to Estimate Trust from Conversation (Li, Erickson, et al., 2023).

#### Modeling trust dynamics in conversations

#### Task Progress:

Prior research has used both qualitative and quantitative approaches to identify and model trust in conversational data. Qualitative analysis, such as grounded theory, provides a rigorous and systematic approach to identifying situated meaning and systematic patterns in the data. However, compared to a machine-aided approach, manual coding is often laborious, limited to small volumes of data, and subject to the coders' domain knowledge. For quantitative analysis, such as text analysis, the dominant approach treats the conversations as bag-of-words, which assumes words are independent units. This approach ignores the meaningful context and patterns in the conversation. In the first research aim, we adopted the machine learning approach, which can combine lexical and acoustic features to predict trust in the conversational agent; however, this focuses on the feature level and ignores the rich context and deep meaning of the conversation. In other words, the connections between the features and the meaning associated with features are situated within the context that might benefit from qualitative analysis. Moreover, the sequence of the conversation is often lost when processing using a bag-of-words approach. Thus, to capture trust dynamics, the objective of this study is to model two aspects: (1) Trust dimensions: the connection to theoretical foundations of trust, especially focus on cognitive processes in conversations, rather than feature level or using bag-of-words; (2) Trust dynamics: the temporal aspect of trust evolution throughout the interactions, rather than aggregated or a snapshot of trust.

We modeled dynamic trust evolution in the conversation using a novel method, trajectory epistemic network analysis (T-ENA). T-ENA captures the multidimensional aspect of trust (i.e., analytic and affective), and trajectory analysis segments the conversations to capture temporal changes in trust over time. Twenty-four participants performed a habitat maintenance task assisted by a virtual agent and verbalized their experiences and feelings after each task. T-ENA showed that agent reliability significantly affected people's conversations in the analytic process of trust,  $t(38.88) = 15.18$ ,  $p = 0.00$ , Cohen's  $d = 144.72$ , such as discussing agents' errors. The trajectory analysis showed that trust dynamics manifested through conversation topic diversity and flow. These results showed trust dimensions and dynamics in conversation should be considered interdependently and suggested that an adaptive conversational strategy should be considered to manage trust in human-agent teaming (HATs). These results have been provisionally accepted for publication in the International Journal of Human Computer Interaction: Modeling Trust Dimensions and Dynamics in Human-Agent Conversation: A Trajectory Epistemic Network Analysis Approach (Li, Amudha, et al., 2023).

#### A computational model of interdependent agents

We also developed a computational cognitive model of interdependent agents, where one agent is a person and the other is a conversational agent. Conversational agents are likely to represent automation that has more authority and autonomy than simple automation. Greater authority may lead the agents' goals to diverge from those of the person. Such misaligned goals can be amplified by the situation and strategic interactions, which can further impact the teaming process and performance. These interrelated factors lack a systematic and computational model. To address this gap, we developed a dynamic game theoretical framework simulating the human-Artificial Intelligence (human-AI) interdependency by integrating the Drift Diffusion Model simulating the goal alignment process.

A 3 (Situation Structure)  $\times$  3 (Strategic Behaviors)  $\times$  2 (Initial Goal Alignment) simulation study of human-AI teaming

	was designed. Results showed that teaming with an altruistic agent in a competitive situation leads to the highest team performance. Moreover, the goal alignment process can dissolve the initial goal conflict. Our study provides a first step in modeling goal alignment and implies a tradeoff between a balanced and cooperative team to guide human-AI teaming design. These results showed how the AI teammate's strategic behavior interacts with the situational factors to influence outcomes. These results have been accepted for publication in the HFES conference proceedings: Modeling Goal Alignment in Human-AI Teaming: A Dynamic Game Theory (Li & Lee, 2022).
<b>Bibliography Type:</b>	Description: (Last Updated: 07/06/2025)
<b>Articles in Peer-reviewed Journals</b>	Li M, Kamaraj AV, Lee JD. "Modeling trust dimensions and dynamics in human-agent conversation: A trajectory epistemic network analysis approach." Int J Hum-Comput Interact. 2023 Apr 27;1-12. <a href="https://doi.org/10.1080/10447318.2023.2201555">https://doi.org/10.1080/10447318.2023.2201555</a> , Apr-2023
<b>Articles in Peer-reviewed Journals</b>	Li M, Erickson IM, Cross EV, Lee JD. "It's not only what you say, but also how you say it: Machine learning approach to estimate trust from conversation." Hum Factors. 2023 Apr 28;187208231166624. Online ahead of print. <a href="https://doi.org/10.1177/00187208231166624">https://doi.org/10.1177/00187208231166624</a> ; PMID: 37116009 , Apr-2023
<b>Articles in Peer-reviewed Journals</b>	Li M, Lee JD. "Modeling goal alignment in human-AI teaming: A dynamic game theory approach." Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 2022 Oct 27;66(1):1538-42. <a href="https://doi.org/10.1177/1071181322661047">https://doi.org/10.1177/1071181322661047</a> , Oct-2022
<b>Papers from Meeting Proceedings</b>	Li M, Erickson I, Cross E, Lee J. "Estimating trust in conversational agent with lexical and acoustic features." 66th International Annual Meeting of the Human Factors and Ergonomics Society, Atlanta, GA, October 10-14, 2022. Abstracts. 66th International Annual Meeting of the Human Factors and Ergonomics Society, Atlanta, GA, October 10-14, 2022. , Oct-2022