International Journal of Human–Computer Interaction

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hihc20

Modeling Trust Dimensions and Dynamics in Human-Agent Conversation: A Trajectory Epistemic Network Analysis Approach

Mengyao Li, Amudha V. Kamaraj & John D. Lee

To cite this article: Mengyao Li, Amudha V. Kamaraj & John D. Lee (2023): Modeling Trust Dimensions and Dynamics in Human-Agent Conversation: A Trajectory Epistemic Network Analysis Approach, International Journal of Human–Computer Interaction, DOI: 10.1080/10447318.2023.2201555

To link to this article: https://doi.org/10.1080/10447318.2023.2201555

Published online: 27 Apr 2023.

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
https://doi.org/10.1080/10447318.2023.2201555

Modeling Trust Dimensions and Dynamics in Human-Agent Conversation: A Trajectory Epistemic Network Analysis Approach

Mengyao Li, Amudha V. Kamaraj, and John D. Lee

Department of Industrial and System Engineering, University of Wisconsin, Madison, WI, USA

ABSTRACT

Human-AI conversation provides a natural, unobtrusive, yet under-explored way to investigate trust dynamics in human-AI teams (HATs). In this paper, we modeled dynamic trust evolution in conversations 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 conversations to capture temporal changes of trust over time. Twenty-four participants performed a habitat maintenance task assisted by a conversational 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, $t(38.88) = 15.18, p < 0.001, Cohen's\ d = 4.72$, 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 for managing trust in HATs.

KEYWORDS

Trust in automation; network analysis; conversational analysis; epistemic network analysis; human-AI teaming

1. Introduction

As artificial intelligence (AI) becomes increasingly capable and can outperform humans in certain tasks, AI may go beyond being tools to cooperate as teammates (Chiou & Lee, 2023; Johnson & Vera, 2019). Considering trust in human-AI teams (HATs) is a step beyond current considerations of trust in automation and poses new challenges for measuring, modeling, and managing trust. An effective interdependent team requires trust models that reflect team processes and how a team's activity unfolds over time. This requires a continuous and observable stream of data to record the cognitive process of trust dynamics. Similar to human-human teams, HATs also need to exchange and update the information to achieve a joint task. Conversation can provide such contextual and process-based means for modeling trust (Cooke et al., 2013). Conversations naturally reflect coordination, which can be used to show changes in human-AI relationships over time. Additionally, trust can be inferred from conversations via people's tone of voice and choice of words (Li et al., 2022). Inferring trust from conversation aligns well with the nature of interdependent HAT and provides an essential means to model and analyze trust dimensions and dynamics through team interactions. Thus, modeling trust dynamics in HATs using conversational data provides a promising yet under-explored approach.

Since trust in HATs conversation is highly contextual, dynamic, and evolves over time, we adopt a novel approach—*trajectory epistemic network analysis (T-ENA)*—to develop a dynamic model of trust evolution in human-agent conversation (Brohinsky et al., 2021). This trust model represents coded conversational data using *epistemic network analysis (ENA)*. ENA helps provide a contextual understanding of conversational data (Shaffer, 2017). Similar to the structure of social network analysis, the nodes in ENA represent trust-related concepts that are defined using a trust framework. These nodes help highlight the multidimensional nature of trust. The edges provide the connections between the concepts based on their co-occurrence in human-AI conversations. The trajectory analysis then characterizes interactions as a trajectory to show temporal changes in trust in AI. In summary, we used T-ENA to model trust dimensions and dynamics in human-AI conversations. We make three contributions to this paper: (1) we modelled trust dynamics using conversational data in human-AI teaming; (2) we adopted a novel method, T-ENA, to show the trust dimensions and temporal dynamics; (3) we identified implications of trust-calibrated conversational agent design.

2. Background

The following sections outline the theoretical foundations for the multidimensional and temporal aspects of trust. To model trust dynamics, we propose using human-agent conversational data. Existing approaches to modeling trust using conversation are discussed and the benefits and drawbacks of each approach are highlighted. Based on the nature of the trust dynamics and limitations of prior approaches,

CONTACT Mengyao Li mengyao.li@wisc.edu Department of Industrial and System Engineering, University of Wisconsin, Madison, WI, USA

we adopt a novel method, trajectory epistemic network analysis for modeling trust.

2.1. Trust dimensions

Trust, defined as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (Lee & See, 2004, p. 51), has been studied for decades to understand and manage the relationships between people and automation. Existing trust frameworks show how trust depends on the interplay between analytic, analogic, and affective cognitive processes. The analytic process refers to cognitive processes that involve deliberative analysis (Hoff & Bashir, 2015; Lee & See, 2004). The analogic process refers to cognitive processes that rely on rules, intermediaries, and environmental context. The affective process refers to cognitive processes guided by emotion. Thus, in understanding and modeling trust in HATs, these cognitive processes must be highlighted. Since trust can manifest in different conversational cues (Elkins & Derrick, 2013; Li et al., 2022; Waber et al., 2015), conversational cues can be used to model cognitive processes of trust.

We know little about the interplay between different cognitive processes underlying trust in human-agent conversation. Miller argued that analogical and affective trust plays a conceptually greater role than the analytic process (Miller, 2005) because people who experience negative affective or analogical trust may stop accumulating information for analytic trust. Yet, in high-risk and safety-critical situations, people are more likely to use analytic processing if they have sufficient cognitive resources (Hoff & Bashir, 2015). The interplay of a more rational analytic process and a more automatic affective process has not been formally studied and modeled. To resolve this research gap, we designed a decision-making task with various levels of conversational agent reliability to reflect their cognitive processes in conversations. The factor, automation reliability, which is governed by the analytic process, has been well-studied and shows the causal relationship on trust (Dzindolet et al., 2003). For the affective process, the circumplex model describes affect using two-dimensional arousal-valence axes to describe human emotions. For example, the affect *excited* is high arousal and positive valence affect, whereas *sad* is low arousal and negative valence. Trust has been shown to be influenced by the valence and arousal stimulus. Dunn and Schweitzer show that positive valence (e.g., happiness, hope) increases trust, while negative valence (e.g., fear, guilt) decreases trust (Dunn & Schweitzer, 2005). Yet, how the affective process of trust is mapped on the affect circumplex model in human-agent conversations is not well understood. Additionally, how the affective process interacts with the analytic process is under-studied. In this study, we aimed to elicit various levels of trust and showed the significant difference and interplay between the *affective* and *analytic* processes.

2.2. Trust dynamics

Another important aspect of trust modeling that merits more attention is its temporal characteristics. A shift from the snapshot view of trust to a dynamic view of trust is important (Yang et al., 2021) because trust is time-dependent and evolves throughout human-agent interaction (Kaplan et al., 2021). The evolution of trust depends on various automation characteristics and experiences as relationships between teammates mature (Korsgaard et al., 2018; Luo et al., 2022). Examining trust dynamics allows us to assess the influence of the recency effect, as interactions that happened more recently may have more influence than those that happened earlier (Desai et al., 2012). Thus, analyzing and modeling the temporal changes offers a nuanced view of the evolution of trust in the HATs.

To model trust evolution, trust should be measured at multiple instances in time and modeled by including a time index in the model. Lee and Moray adopted time series analysis to uncover trust dynamics (Lee & Moray, 1992). Lee and Gao extended decision field theory (DFT) to describe operators' information accumulation and multiple sequential decisions in supervisory control situations. This computational model predicted trust dynamics (Gao et al., 2006; Gao & Lee, 2006). Yang et. al. proposed a computational model which proposes that trust at any time t, follows a Beta distribution with good prediction accuracy (Yang et al., 2021). Although modeling trust evolution is relatively limited, there is a long history of modeling human behaviors and attitudes with a time-dependent dynamical system approach. Gottman et al. (2002) showed how marriage outcomes can be modeled by using dynamical system analysis, which focused on the temporal dynamics of partner conversation. Using such dynamical systems methods to model relationships is becoming more prevalent (Demir et al., 2021).

The multidimensional and temporal aspects of trust are not independent. Instead, the influencing factors and their effects on various processes of trust dimensions might also vary throughout the human-AI interactions. In modeling interpersonal trust, Korsgaard and colleagues outlined a stage model that captures the trust formation from an early stage of calculus-based to a knowledge-based trust and eventually an identification-based trust that reflects values and goals (Korsgaard et al., 2018; Lewicki & Bunker, 1996). In various stages of trust, predictors affect trust systematically and vary over time (Korsgaard et al., 2018). Kaplan and colleagues proposed a dynamic model of trust with time as the horizontal axis and interaction between various trust antecedents on the y-axis (Kaplan et al., 2021). In sum, prior research highlights the importance of time as a moderator on different trust dimensions. However, to the authors' knowledge, limited research empirically investigates the relationship between the multidimensional and temporal aspects of trust, especially in human-agent conversation. In this paper, we modelled trust dimensions by decomposing the cognitive processes (i.e., analytic and affective) and examine their relationships with temporal dynamics in the human-agent conversation.

2.3. Modeling trust in conversation

A critical challenge in modeling trust is to accommodate the highly contextual, dynamic, and evolving relationship between humans and AI teammates throughout the interaction. To model latent variables such as trust, it must be inferred through indirect indicators, such as subjective, behavioral, and physiological measurements. Although subjective trust ratings are treated as the gold standard in human-automation interaction, they do not fully reflect trust dynamics because it is often obtrusive and one-shot. The interruptions and the deliberate thinking involved while self-reporting attitudes towards automation do not represent the joint cognitive processing that happens in human-AI cooperation. While using behavioral measures of trust such as compliance and reliance provide a more fine-grained sampling of trust throughout the interaction, it strongly depends on the task and is limited to the available decision spaces (Kohn et al., 2021). Physiological measures, such as electrodermal activity, eye movement, and heart rate, can provide real-time trust indicators with greater sensitivity. However, physiological measures also suffer from challenges, such as outcomes that must be analyzed and contextualized with expert knowledge and examination during periods where trust is active and relevant (Kohn et al., 2021). These challenges suggest a need for alternative measures of trust.

One under-explored trust source is conversation data. Conversational data can be considered a mixture of lexical, semantic, phonological, and pragmatic representations of the conversations. In other words, people naturally express their trust attitudes via the words they use, the sentence structure, and the tone of the voices in their conversation, which are all contextualized. People express their trust not only through what they say (e.g., the sentiment of the words), but also via how they say it (e.g., formants) (Li et al., 2022). Based on interactive team cognition theory, communication represents team cognition and can serve as a non-obtrusive measure of team interaction dynamics (Cooke et al., 2013). The conversation is also essential for trust building and calibration, which in turn, can promote effective human-AI teaming (Fuoli & Paradis, 2014).

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 (Oktay, 2012). 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. A machine learning approach can combine lexical and acoustic features to predict trust in the conversational agent (Li et al., 2022); however, machine learning 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, we modelled two aspects: (1) Trust dimensions: the connection to theoretical foundations of trust and cognitive processes in conversations, rather than decontextualized feature; (2) Trust dynamics: the temporal aspect of trust evolution throughout the interactions, rather than aggregation or snapshot of trust.

2.4. Trajectory epistemic network analysis

To characterize trust dimensions and dynamics, we applied Trajectory Epistemic Network Analysis (T-ENA), which can both decompose multidimensional trust using an Epistemic Network Analysis (ENA) and project the trajectory of the network structure over time.

ENA is a quantitative ethnographic technique that estimates the network structure of coded data based on the co-occurrence of coded elements that define connections between the coded data (Shaffer, 2017). Similar to Social Network Analysis (SNA), which analyzes relationships between people, ENA can quantify the changes in both strength and structure of connections between coded elements. In our analysis, coded elements were parts of the conversation but could also include other behaviors. Originally designed to model theories of cognition, discourse, and culture challenges in learning analytics, ENA assumes that the structure of the connections is more important than the mere presence of those elements in isolation (Andrist et al., 2015).

ENA has been applied to many domains because it can quantify complex qualitative data, such as gaze coordination and social interactions in collaborative work (Andrist et al., 2015) and shared agency in online collaborative learning (Tan et al., 2022). Prior research has demonstrated successful applications of ENA to human factors and ergonomics (HFE) domains because the visual representations can help researchers quickly identify and compare the difference between groups (Weiler et al., 2022; Wooldridge et al., 2018). Additionally, the differences can be quantitatively defined with the support of qualitative evidence from the conversation. In this paper, we applied ENA to construct and visualize a multidimensional space of trust based on analytic and affective processes in human-agent conversation.

To model trust dynamics, one major limitation of ENA is that it typically aggregates data across conditions and time while ignoring temporal features. Trajectory ENA considers the temporal structure to reflect process-oriented concepts, such as trust dynamics. T-ENA accounts for the change in the network structure that evolves by incorporating a time index or temporal segmentation. By dividing the complex ENA into time segments, T-ENA shows changes along the temporal dimension, which would otherwise be lost when aggregating data (Tan et al., 2022). Modeling trust dynamics using T-ENA can represent both the multidimensional and temporal aspects of trust.

2.5. Research objectives

This study investigated trust dynamics in human-agent conversations by addressing two research questions: (1) how do humans indicate different trust levels in human-agent conversations? (2) how does human-agent trust conversation change over time? We adopted a novel approach, trajectory epistemic network analysis that can capture the contextual and dynamic nature of trust in HATs. To address the first question, we used ENA to depict the multidimensional structure of trust. To address the second question, we used T-ENA to show the evolution of trust. We adopted long-duration space exploration operations as a context to study trust dynamics. As the National Aeronautics and Space Administration (NASA) moves from moon missions to Mars missions, the longer communication delays lead to an increasing need for cooperation between humans and virtual agents, which requires a close examination of trust dynamics.

3. Method

3.1. Study design

We analyzed data from a 2 × 2 × 3 within-subject study (see Figure 1). Participants completed 12 decision-making tasks where they managed a Carbon Dioxide Removal System (CDRS) that is part of an analog Mars habitat. Because trust changes as a dynamic process that varies across interactions (Yang et al., 2021), we designed repeated exposures with the automated system to ensure we captured multiple measures of trust. Participants were assisted by a conversational agent with two levels of agent reliability (i.e., high, and low). Each level of reliability had two cycles of the CDRS tasks, each including three events (i.e., startup, venting, shutdown). The high-reliability conversational agent provided 100% correct recommendations whereas the low-reliability agent provided 20% correct recommendations. The 12 events were designed to elicit various levels of trust through differing agent reliability. At the end of each event, the agent initiated a conversation by asking six trust-related questions (see Table 1). Once the participant finished the conversation, they then completed a 12-item Checklist for Trust between People and Automation on a 7-point Likert scale (Jian et al., 2000). An example item is "The system is reliable." This survey is the most frequently used and cited survey in trust in the automation domain, which includes a variety of items sampling trust and distrust (Kohn et al., 2021).

3.2. Participants

A total of 24 participants (18 female, 6 male) were recruited from the Madison, WI area ($M = 23.7$, $SD = 3.6$). Due to the safety concerns of COVID-19, the study took place online. It was a 2-day study with each day's participation lasting up to 2 h. In total, the study was approximately 4 h. Participants received $30 per hour for a total of up to $120. This research complied with the American Psychological Association Code of Ethics and was approved by the Institutional Review Board at the University of Wisconsin–Madison. In total, each participant had the opportunity for 72 conversational turns with the agent. The cleaned text data contained 1981 lines of utterances, with a mean text length of 38.25 characters ($SD = 26.49$). A t-test showed that the mean trust score for the high-reliability condition ($M = 5.78$, $SD = 0.86$) was significantly higher than the low-reliability condition ($M = 4.37$, $SD = 1.44$), $t(23) = 4.12$, $p = 0.0002$. Thus, we can investigate the conversational indicators associated with high-trust and low-trust situations.

3.3. Apparatus

The experimental task used the Procedure Integrated Development Environment (PRIDE) which is an automated procedure software, to maintain the space station habitat using the Carbon Dioxide Removal System (CDRS) (Schreckenghost et al., 2014). A conversational agent, named Bucky, was preprogrammed with procedure protocols to provide recommendations to help participants follow the PRIDE procedures to maintain the habitat. Google Dialogflow, a Natural Language Understanding (NLU) platform

Table 1. Examples of conversational trust questions.

Number	Question
1	How would you describe your experience selecting the procedure?
2	Why would you feel that? Can you explain your answer in more detail?
3	Can you talk more about my performance in providing the recommendation?
4	That makes sense. Which procedure did you select?
5	Can you tell me more about your strategy for picking that procedure?
6	How can I be more helpful in terms of providing recommendations?

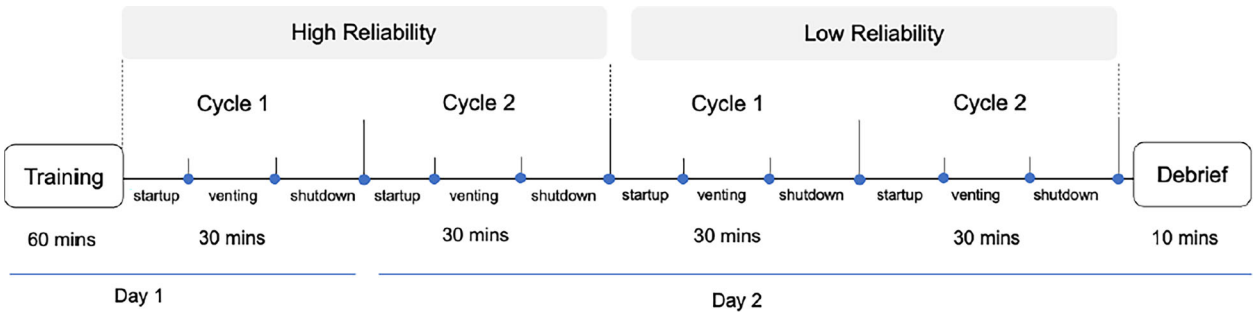

Figure 1. The study design with two levels of reliability conditions.

was used to design and integrate the user interface. Participants were asked to directly speak to the conversational agent using their computer's microphone. Keyboard and button inputs were also provided. Text data was automatically transcribed using speech-to-text technology. Both audio and text data were collected.

3.4. Procedure

After signing the consent form, participants completed training on PRIDE, CDRS, and Bucky systems. During the study, participants had 25 minutes to completely control all three CDRS events (startup, venting, and shutdown) before their crew experienced CO_2 poisoning. For each event, the participant made two essential decisions with Bucky's aid (i.e., procedure recommendation and verification suggestion). Bucky would provide a recommended procedure of CDRS and participants could either accept Bucky's recommendation or reject it and manually select an alternate procedure. Once the procedure was selected, PRIDE automated the procedure execution. After the procedure finished running, Bucky verified the system status and made suggestions on whether participants should rerun the procedure. Again, participants could decide to accept or reject the suggestion. Once the participant finished the event, Bucky initiated a conversation with six questions with some variations to avoid being repetitive (see Table 1). After conversational questions, participants completed the trust questionnaire. The total time of each cycle, including the trust conversation and questionnaire, was approximately 40 min. At the end of the study, participants were debriefed and compensated.

3.5. Trajectory epistemic network analysis

For trajectory epistemic network analysis (T-ENA), we adopted a four-step process as shown in Figure 2: (1) data segmentation, (2) directed content analysis, (3) network analysis, and (4) trajectory analysis.

3.5.1. Data segmentation

Conversation data between participants and the conversational agent were recorded in log files, which were segmented based on the conversational turn. Four types of meta-data were used for data segmentation: (1) Reliability condition that participants experienced. This is used as the grouping variable for comparison between two conditions. (2) Participant ID, which is used as a 'unit' in the ENA. (3) Question ID, which defines as 'conversation' for ENA. Conversations are collections of turns within which ENA models connections between concepts. (4) Codes, which are generated by the directed content analysis (described in section 3.5.2).

3.5.2. Directed content analysis

Trust dynamics models (Yang et al., 2021) were used to identify the six codes shown in Table 1, which include four codes related to the analytical processes of trust and two codes related to the affective processes of trust. For analytical processes of trust, codes were identified in an iterative round of coding. Two researchers combined deductive and inductive coding to refine and validate codes. For affective processes of trust, we adopted the circumplex model of affect (Russell, 1980), which suggests that affect is described in a two-dimensional circular space, containing arousal and valence dimensions. We excluded two quadrants in the valence-arousal affect model: positive valence, high arousal (e.g., excited), and negative valence, low arousal (e.g., sad) since these did not appear in the conversational data.

The directed content analysis identified trust components within the participants' conversations with the virtual agent throughout the task. Two researchers coded each turn using a binary coding structure: "1" if the code exists, or "0" if the code does not exist per each segment. Coders compared codes and categories and re-coded certain segments to resolve disagreements and achieve consensus on final binary codes for each turn in every transcript. Table 2 shows the final codebook used for the ENA analysis.

3.5.3. Network analysis

Based on the codebook from *Directed Content Analysis*, each segment of coded data is represented as a vector of six 1 or 0 s representing the presence or absence of each code. These vectors define the nodes of networks and the cooccurrence of codes in a conversational context defines the links between nodes. The ENA algorithm uses a *moving window* to define the conversational context for each conversational turn in the data, showing how codes in the current turn are connected to codes that occur within the recent temporal context. Here we defined this context as 12 turns (each turn plus the 11 previous turns) within a given conversation. Codes that occurred outside of this window were not considered connected. The size of the moving window is defined by the number of questions after each decision-making point in the experiment (i.e., six questions with answers for each question, $6\times2=12$). Then, the co-occurrence of codes is converted into adjacency matrices

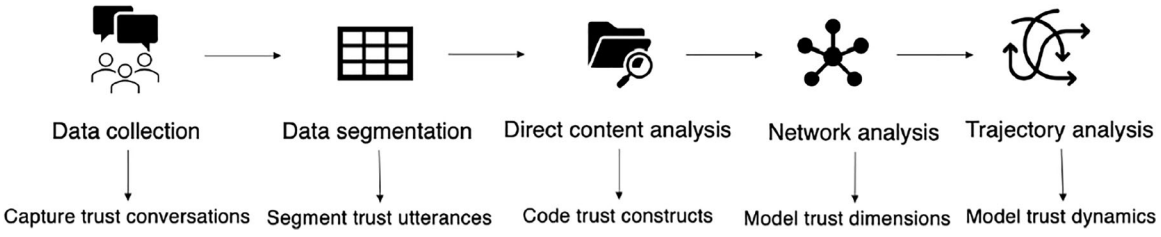

Figure 2. Trajectory epistemic network analysis process and for assessing trust dimensions and dynamics.

Table 2. Codebook of trust-related constructs included in the Epistemic Network Analysis.

Code	Definition	Example from data
System capability	Participants commented on Bucky's past and/or current performance and ability to provide the appropriate recommendation for the task.	*I think your **performance** was good since it worked out well.*
System error	Participants commented on errors in Bucky's recommendations.	*The procedures didn't line up to **what I thought the right procedure would be**.*
User capability	Participants commented on their self-efficacy and their belief in his or her capacity to execute the task.	*Bucky is incorrect this round, but I'm **confident** in myself for choosing the correct one.*
System process scrutiny	Participants recalled the specific system knowledge to understand or clarify how the system operates.	*It was the only option in which the **EPS was powered up before the ATCS was activated. If the EPS is not powered up, then the ATCS can't be activated**, therefore I assumed this was the only procedure that would be effective.*
Positive valence, low arousal	Participants expressed their affect that is positive and low aroused, such as calm, contented, and relaxed.	*I feel like I've reached a routine with my method of choosing the procedure. So, I enter the same state of **calm**.*
Negative valence, high arousal	Participants expressed their affect that is negative and highly aroused, such as confused, frustrated, stressed, nervous, and annoyed.	*I get even more **confused** with Bucky's recommendation*

and summed across the moving window into a cumulative adjacency matrix. Each matrix is then converted into an adjacency vector by copying the cells from the upper diagonal of the matrix row by row into a single vector. Next, ENA is normalized using spherical normalization by dividing each vector by its length, representing frequencies of co-occurrence code pairs. Once data is normalized, ENA performs a singular value decomposition (SVD) using two SVD dimensions. The resulting network is then visualized by locating the network nodes using an algorithm that minimizes the projection of the point under SVD and the centroid of the network graph under the node positioning being tested. The optimization allows the structure of the ENA space to account for the most variance between different networks. Links are then constructed between the positioned network nodes according to the adjacency matrix.

For the resulting network, nodes correspond to codes; edges correspond to the co-occurrence between each pair of codes and the thickness of the edges shows the strength of the connection between nodes defined by the relative frequency of co-occurrence. The centroids are the mean values based on node position weighted by each edge weight in the networks. To compare the trust indicators in the conversation, the network was grouped based on the two reliability conditions of the conversational agent. To determine if the high-reliability conditions are statistically different from the low-reliability conditions, we conducted *t*-tests on the network centroids.

3.5.4. Trajectory analysis

We used the R package *trajectoryENA* (Brohinsky et al., 2021) to create trajectories. We coded the conversations with 12 time segments, which is each conversation after each event (startup, venting, shutdown). Thus, each reliability group had six-time segments. The mean for each time segment was projected in the aggregated ENA space described above. Group means were plotted and sequentially connected, which produce curves between successive time points. Adding the time unit to the ENA allows us to investigate how people's trust evolves from the beginning to the end, which the aggregated ENA analysis ignores.

4. Results

4.1. Epistemic network analysis

The graphs summarizing the Epistemic Network Analysis (ENA) contain: (1) plotted points, which represent the location of that unit's network in the low-dimensional projected space, and (2) weighted edges connecting these points. The positions of the network graph nodes are determined by minimizing the difference between the plotted points and their corresponding network centroids. This co-registration of network graphs and projected space, the positions of the network graph nodes, can describe the dimensions of the projected space. Our model had co-registration correlations of 0.98 (Pearson) and 0.98 (Spearman) for the first dimension and co-registration correlations of 0.92 (Pearson) and 0.90 (Spearman) for the second. These measures indicate that there is a strong fit between the visualization and the model.

Figure 3 shows the network for high and low reliability. In these network graphs, nodes correspond to the codes identified that are relevant to trust indicators in the conversations, and edges reflect the relative frequency of co-occurrence of these codes within each conversation between participants and the conversational agent. Thus, the thicker the edges, the more frequent the co-occurrence and the stronger the node connection in the human-agent conversation. Figure 4 shows subtracted network graphs, which subtract the high and low-reliability networks nodes and connections from each other to create a different network graph. The resulting network provides the visual representation to show the difference between node connections and edge width in high and low-reliability conditions. The centroids in Figure 4 summarize the dimensions of each network. Centroids, indicated by square points and confidence intervals (dotted lines), enable comparisons of networks statistically as well as visually.

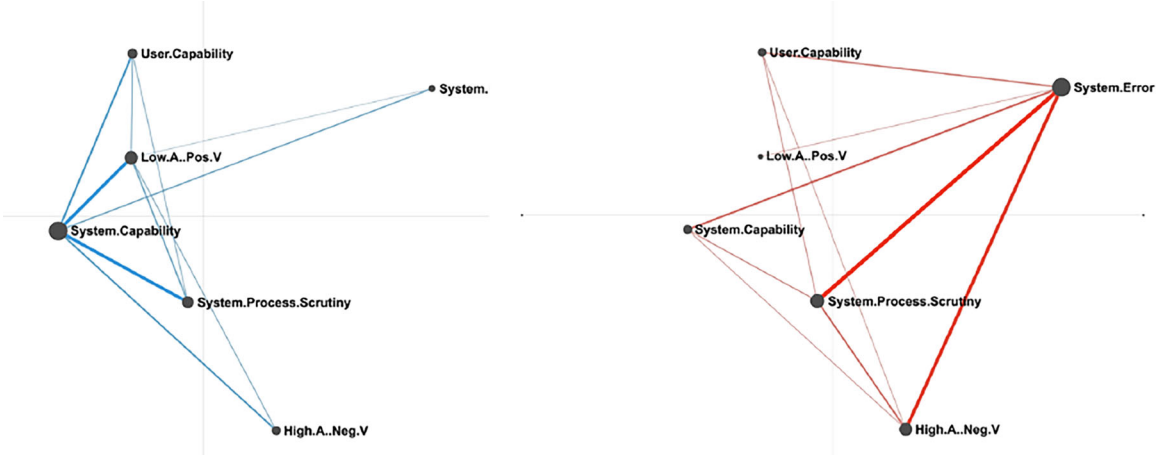

Figure 3. ENA network for the high reliability (left) versus low reliability (right) conditions.

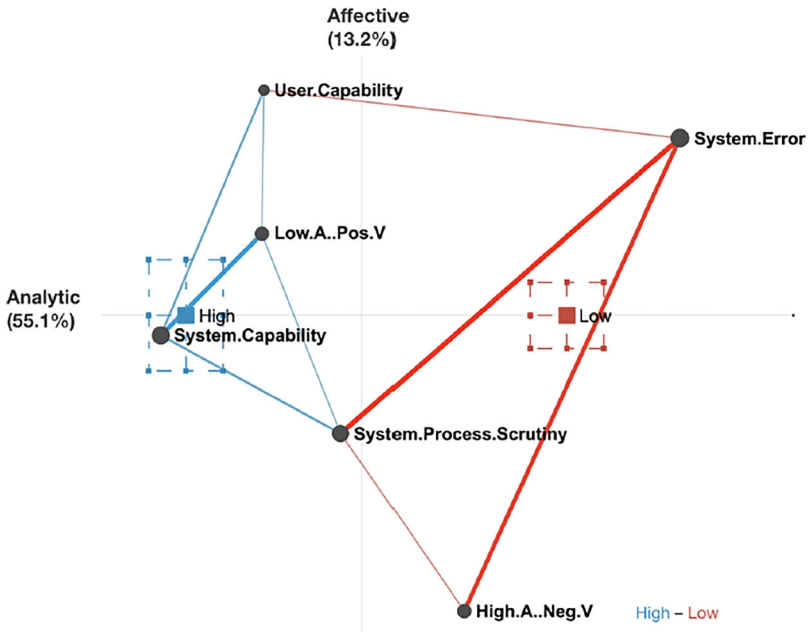

Figure 4. ENA network of subtracted connections for high reliability (blue) versus low reliability (red). The points represent coded topics, and the edges represent the co-occurrence of the topics. The thicker the edges, the more frequently the topics co-occur in the human-agent conversation. The square points and associated error bars represent the centroids and the confidence interval of the network.

To test the differences between the reliability conditions, we applied a two-sample *t*-test on the distribution of the centroids of each group. Along the *x*-axis, a two-sample *t*-test showed that the high-reliability condition ($M = -1.15, SD = 0.55, N = 22$) was statistically significantly different at the $\alpha = 0.05$ level from low-reliability condition ($M = 1.33$, $SD = 0.50$, $N = 19$); $t(38.88) = 15.18$, $p < 0.001$, $Cohen'sd = 4.72$. Along the *y*-axis, the two-sample *t*-test assuming showed high-reliability condition ($M = 0.00, SD = 0.82, N = 22$) was not statistically significantly different at the $\alpha = 0.05$ level from the low-reliability condition ($M = 0.00$, $SD = 0.45$, $N = 19$); $t(34.45) = 0.00$, $p = 1.00$, $Cohen'sd = 0$.

The *x*-axis and *y*-axis of the ENA network should be interpreted based on the code placement and researchers' domain knowledge. Nodes at the extreme edges of the space provide more information for labeling the axis. The codes that capture the degree of *analytic processes* of trust define the *x*-axis. These codes include system capability, system errors, system process scrutiny, and user capability. Moving from left to right along the *x*-axis indicates conversation topics shift from positive aspects of system capability to negative aspects, such as system errors. Codes that reflect the *affective processes* of trust in the system, which include the high/low arousal and positive/negative valence of affect define the *y*-axis. The significant result on the *x*-axis in Figure 4 indicates that the conversation between high-reliability versus low-reliability conditions differed along the analytical level conversation codes. In the high-reliability condition, the conversation is centered around the system performance (e.g., The performance is good). In the low-reliability condition, the conversation was more centered on the errors that occurred in the system and its connections with the system scrutiny (e.g., The CO_2 level should be lower).

Based on Figure 4, in the high-reliability condition, the strongest connection is *System Capability* and *Low Arousal,*

Positive Valence, indicating that when the conversational agent is performing well, people usually commented on the system performance along with positive valence and low arousal affect words, such as "calm" and "relax". Additionally, the connection between *System Capability* and *User Capability* indicates that people often reflected on their self-efficacy and talked about their capability when the system performs well. In the low-reliability condition, between the affective and analytical processes of trust, we noticed a strong connection between *System Error*, *High Arousal, and Negative Valence*. This means that low performance and lower levels of trust were associated with high arousal and negative valence words (e.g., annoyed). Additionally, there is a strong connection between *System error* and *System Process Scrutiny*. This suggests that in the low-reliability condition, people expressed their low level of trust by thinking aloud about the specific system processes, such as reflecting on what states CDRS should have been in certain situations (i.e., *System Process Scrutiny*).

4.2. Trajectory ENA

Figure 5 shows the trajectory model for the high- and low-reliability conditions across 12 interactions. Every point on the graph shows the mean for each time segment, which is each conversation after each event (a total of 12 conversations). A total of three time-series ENA trajectories were plotted: one with two-dimensional ENA showing both affective and analytic processes (Figure 5(a, c)) and two one-dimensional ENA with each process along with the time (Figure 5(b)). To interpret the trajectory, three crucial variables need to be disentangled: changes in the *x*-dimension, changes in the *y*-dimension, and progression in time. The subplots were co-registered with the main plot, thus the comparison between plots also allowed changes in the subplots to be interpreted with the dimensions of the main ENA space.

Figure 5(a) tracks changes in the y-dimension (affective process) of the original ENA space over time (time as the *x*-axis). This shows the evolution of trust in changes in affective trust over interactions. The variation across the y-dimension indicates a mixed and continuous emotion related to trust shown in the conversation. For example, the subject commented: "*Since this is the first time Bucky's been incorrect, it confused me for a little bit and made me second-guess myself just because Bucky's been so accurate.*"

Figure 5(c) tracks changes in the *x*-dimension (analytical process) of the original ENA space over time (time as *y*-axis). This shows the analytic process of trust as a function

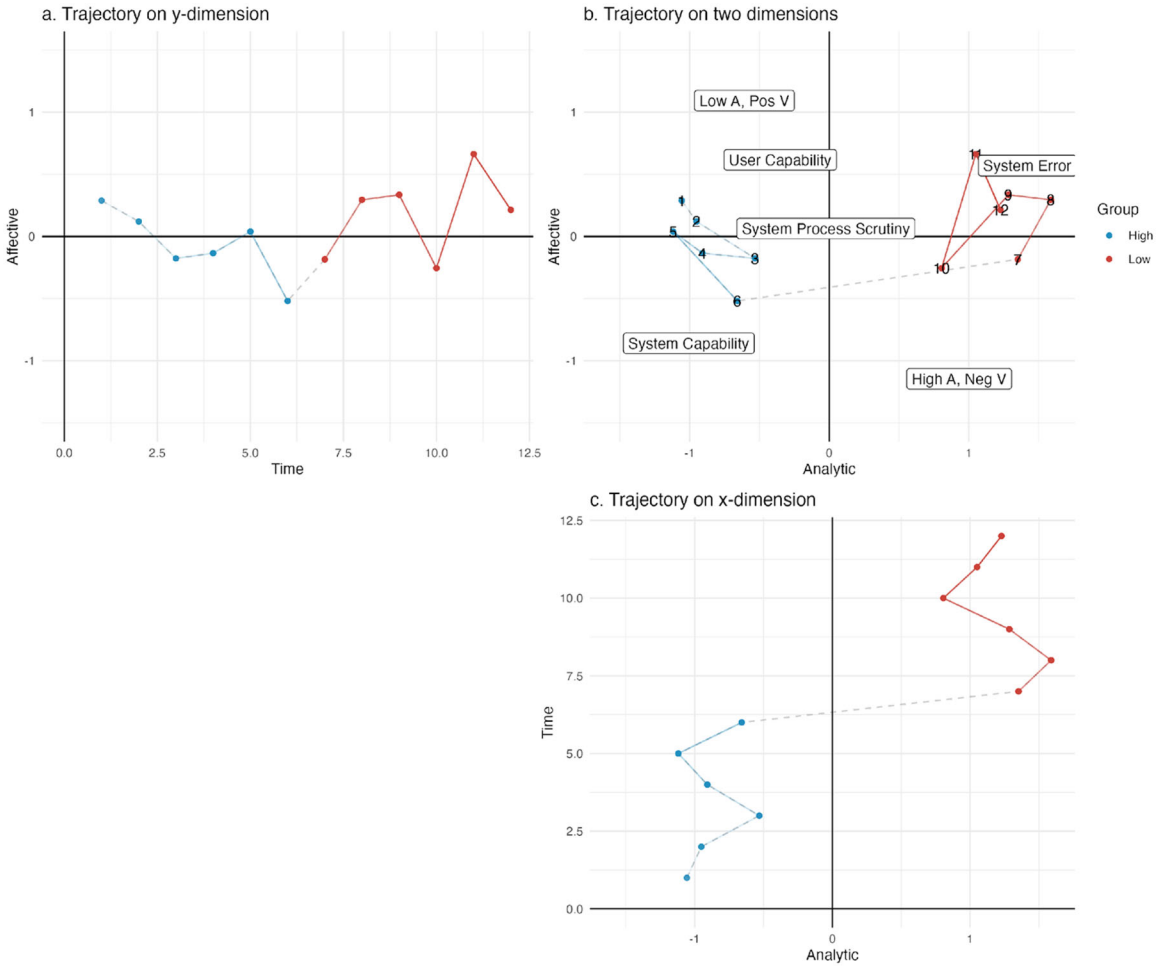

Figure 5. Trajectory ENA. Figure (a) shows the affective process of trust changes in the *y*-dimension as a function of time. Figure (b) shows the two-dimensional trajectory mapping onto the network result. Figure (c) shows the analytic process of trust changes in *x*-dimension as a function of *t* time. The increasing transparency indicates the increasing time throughout the interaction.

of time. Compared to the affective process of trust, which is more continuous and non-significant between high and low-reliability groups, the analytic processes showed a distinct pattern difference. This suggested that the analytical process of trust follows a swift transition.

Figure 5(b) maps the two-dimensional trajectory on top of the ENA node positions, which shows trust dynamics on both the multidimensional and temporal aspects of trust. The annotated numbers correspond to the centroid of the topics over twelve decision-making interactions. The direction of the trajectory indicates the changes and convergence of topics over time. When people have high trust in an agent, they attribute their capability with positive sentiment and later confirm the system's capability. When people interact with low-reliability agents with low trust, conversations note the system error and then converge towards checking the system process with a large variance in the affective processes. Additionally, we noted a distinct difference in the topic variance of the trajectory between high and low-reliability groups. The variance of the trajectory suggests the diversity of the conversational topics. The low-reliability group shows a wider range on both the x and y dimensions, indicating higher volatility when people express a low level of trust. This shows that people have mixed emotions (affective) and analytical judgment when interacting with a poorly performing agent. In sum, our T-ENA results show that trust changes as a function of time, and the prominence of analytical and affective dimensions changes over time.

To better understand trust dynamics and evolution in HATs, we applied a novel approach, trajectory epistemic network analysis (T-ENA), to twenty-four human-agent conversations. Specifically, we explored the multidimensional aspect of trust using ENA and temporal change of trust using the trajectory analysis of ENA. For the trust dimensions, the ENA plots provided meaningful connections between analytic and affect processes of trust concerning agent reliability. For trust dynamics, the temporal analysis segmented the change of trust throughout the courses of the human-agent interactions and mapped it with analytic and affective dimensions of trust.

4.3. ENA showed an interplay between analytic and affective processes of trust

A significant difference between high and low-reliability conversations was shown in the x-axis, which is interpreted and labeled as an analytic process of trust. Results suggest that people express different trust states by using distinct analytic information, such as commenting on system performance and noticing errors. This is expected since we manipulated the reliability of the conversational agent, which maps to the analytic process. No significant difference was found in the affective process. This suggested that the manipulation of reliability showed less influence on the affective process, which aligns with prior literature that the affective process has a greater influence on the analytic process than the analytic has on the affective (Lee & See, 2004). Especially in low risks and self-relevant decisions, the effect of the affective process on trust is much weaker (Midden & Huijts, 2009). In our case, the simulated CO_2 removal procedure did not affect the participants' physical environment, thus the physical and psychological distances to the potential hazards were far. Participants experienced a low level of risk and low self-relevance, which would induce less changes in the affective process of trust.

The network analysis also revealed interactions between analytic and affect processes of trust under the influence of automation reliability. When people show high trust in conversational agents, on the affective dimension of trust, there is a stronger connection between low arousal and positive valence affect with the system and user capability. Complementing our prior paper using machine learning models which showed that positive sentiment predicts trust (Li et al., 2022), ENA results provided more context-relevant information: the positive sentiment is associated with the system capability and users' capability. In the low capability, people indicate high arousal and negative valence and discuss system errors with a detailed inspection of the system process. For the future design of the conversational agent, when people have lower levels of trust, the agent should provide more details on system processes to support the cognitive processes.

Additionally, the different links of conversational topics between high and low reliability conditions have been observed: a stronger connection with user capability in the high condition whereas a stronger connection with system process in the low reliability condition. These results show people's self-serving attribution bias since people often credit positive events internally with their capability and attribute negative events externally by scrutinizing the system processes and errors (Miller & Ross, 1975). A prior study showed when a robot gave people credit, people would trust the robot more (Kaniarasu & Steinfeld, 2014; You et al., 2011). Our study provided the potential for using people's self-serving bias and blame attribution when designing agents' communication strategies. People might be more likely to accept and trust the virtual agent if the agent credits users' capability when the joint task went well and accepts blame if the joint performance was poor. Future empirical studies can further validate the hypotheses and show the effects on trust processes.

4.4. ENA trajectory showed temporal dynamics of trust

For the temporal aspect of trust dynamics, T-ENA showed the temporal change of trust throughout human-agent interactions by mapping the temporal changes in trust to the analytic and affective dimensions of trust. Kaplan and colleagues hypothesized a dynamic mode of trust where trust, at each measurement point, is based on a triangle form of three antecedents (human, robot, contextual)(Kaplan et al., 2021). Our T-ENA results are the first empirical validation of Kaplan's dynamic model and show the relationships between trust antecedents and time. We showed that trust at each measurement can show different processes of trust in an evolving relationship. We observed clear differences in conversation trajectory on affective and analytic dimensions.

For the affective dimension of trust, people show mixed and continuous emotions related to trust shown in the conversation. For the analytic dimension of trust, a distinct difference was observed in the conversation. This implied that using analytic information to estimate people's trust transitions can be more effective in human-agent conversation.

The variance and direction of the conversational trajectory on the two-dimensional trust dynamics also suggested the differences in conversational topic diversity and flow. When people have high trust in the agent, people's conversation topics are more consistent and converged to the system's capability. When in a low trust state, conversational topics are more scattered. One potential explanation is that the scattered conversational topics reflect heavier cognitive processing because topic shifts would cause unexpectedness for people to process the connections between topics (Dessalles, 2017). Our results on human-agent conversations in various trust states shed light on their cognitive processes. Compared to high trust, which leads to a familiar congruent flow of cognitive processing (thus consistent conversational topics), low trust or distrust triggers a spontaneous activation of alternatives and incongruent associations, which can be shown as a diverse topic or verbose examination of the system (Mayo, 2015). These results implied the cognitive mechanism of epistemic vigilance in human-AI conversation. Epistemic vigilance refers to a processing cost to be minimized when the information communicated is of no relevance to oneself (Sperber et al., 2010). Trust is buttressed by epistemic vigilance: during the trust calibration process, in the high trust state, people are less vigilant and minimize the processing cost, which manifested as a consistent topic trajectory; in the low trust state, a higher epistemic vigilance, thus, it leads to a heavier processing cost, which manifested as a more complex conversational topic and structure. Prior research has studied the relationship between trust and vigilance in the detection performance of automation monitoring (Molloy & Parasuraman, 1996). Future research is needed to further validate the influence of trust and epistemic vigilance on human-AI conversation.

4.5. Trust-calibrated conversational agent design

With the proliferation of intelligent conversational agents (e.g., Siri, Amazon Alexa, ChatGPT), designing trustworthy conversational agents becomes more important because it is the foundation for both social (interactional) and transactional (task-based) conversations (Clark et al., 2019; Rheu et al., 2021). While currently most conversational agents are designed to be exclusively task-oriented and transactional, more research has focused on long-term and multiple-turn agents to not only fulfill service requests but also address social needs (Clark et al., 2019). Trust dynamics become increasingly important since it is core to achieving common ground for long-term relationship building. In this study, we investigated trust dynamics in human-agent conversation and provided design applications for the conversational agent: an adaptive conversational strategy should be adopted to better manage people's trust in various states. The conversation can be designed using the principle of conversational entrainment. Entrainment refers to the phenomenon where communication partners synchronize their behaviors, which can promote productive conversation and enhance cooperation by supporting predictive processes and reducing cognitive processing Manson et al., 2013). Specifically, our results imply when users are in high-trust, the conversational agent can provide minimal input with validation for the system performance; when in a low trust state, the conversational agent should proactively provide more analytic information regarding the system process to reduce cognitive processes. Additionally, implementing acoustic-prosodic entertainment (e.g., speech rate, pitch properties) in the conversational agent can be beneficial to manage trust in human-AI conversation (Li et al., 2022). The question that remains is how to mitigate users' over- and under-trust using conversational cues. Future research should focus on designing conversational strategies for adaptive trust calibration.

4.6. ENA trajectory for human-computer interaction (HCI) research

In this paper, we demonstrated that the T-ENA method can be used to model trust dimensions and dynamics in human-agent conversations: ENA provides qualitative meaning to structured networks with statistical tests and trajectory analysis can further analyze the temporal changes. Our results provide empirical evidence for future HCI research topics. Prior research has shown promise for understanding individual contributions versus team interactions in human-human conversations (Siebert-Evenstone et al., 2017). Future studies can apply T-ENA to understand the structures of human-AI teams.

Application of ENA is not limited to conversational data. Because of ENA's characteristics of quantifying meaningful connections among elements, researchers have analyzed social gaze coordination using eye-tracking data in a joint task (Andrist et al., 2015), identified areas of co-activated brain areas using fMRI data (Collier, 2015), and visualized people's spatial movements in clinical team simulations (Fernandez-Nieto et al., 2021). Essentially, ENA makes it possible to treat qualitative data quantitatively to extract insights that might otherwise be lost with a purely qualitative approach.

4.7. Limitations and future studies

It is important to note several limitations in our study to better generalize the findings. First, the human-agent conversation has a pre-defined decision-tree structure due to the limits of the state-of-art conversational agent capabilities. On the one hand, we were able to compare the difference in answers systematically across the interactions. However, compared to the human-human conversation, the conversations can appear to be limited in terms of the potential topics discussed. Thus, the coverage of the topics can be less diverse than human-human conversation, which cannot provide rich information for coupled conversational analysis. Future studies using a more robust conversational agent can

generate more dynamic conversations and trust-related findings. Second, since conversations are heavily contextual, the conversation in our study is domain focused. For example, for the node of *System Process Scrutiny*, people used jargon related to our study design, such as the carbon dioxide removal system. Thus, when generalizing findings from our study to another domain, the coding for the nodes in the network should be adjusted to the context. Additionally, researchers should consider whether the task situations and relationship between humans and agents can be generalized. Our study manipulated the reliability conditions of the agent, and the task was safety-critical with heavy cognitive loads. Future studies should also consider social and non-critical conversations between humans and AI.

5. Conclusion

To support human-AI teaming, the AI needs to monitor and manage trust dynamics in real time. Conversational data provides a novel approach to measuring, modeling, and managing trust. Prior approaches using quantitative analysis (e.g., machine learning, text analysis) or qualitative analysis (e.g., grounded theory) cannot provide *deep connections* between the trust indicators. We employed trajectory epistemic network analysis, a quantitative ethnographic approach that identifies time-series patterns in the data while providing interpretable construct connections, using human-agent conversational data. ENA mapped the multidimensional aspect of trust and showed that reliability affected the analytic process of trust. People scrutinized system processes and misaligned information when they were in a low-trust state. T-ENA showed the temporal dynamics of trust throughout human-agent interaction. Results showed a distinct difference in conversational topic diversity and flow over time, which suggested that the agent's conversational strategy should be adaptive based on people's trust states. Our study enhanced the understanding of trust dimensions and dynamics in human-AI conversation and teaming.

Acknowledgments

We thank members of the University of Wisconsin-Madison Cognitive Systems Laboratory for their insightful discussions and comments.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by NASA Human Research Program No. 80NSSC19K0654.

ORCID

Mengyao Li http://orcid.org/0000-0002-0819-4693
Amudha V. Kamaraj http://orcid.org/0000-0002-8602-1448
John D. Lee http://orcid.org/0000-0001-9808-2160

References

Andrist, S., Collier, W., Gleicher, M., Mutlu, B., & Shaffer, D. (2015). Look together: Analyzing gaze coordination with epistemic network analysis. *Frontiers in Psychology*, *6*, 1016. https://doi.org/10.3389/fpsyg.2015.01016

Brohinsky, J., Marquart, C., Wang, J., Ruis, A. R., & Shaffer, D. W. (2021). Trajectories in epistemic network analysis. In A. R. Ruis & S. B. Lee (Eds.), *Advances in quantitative ethnography* (Vol. 1312, pp. 106–121). Springer International Publishing. https://doi.org/10.1007/978-3-030-67788-6_8

Chiou, E. K., & Lee, J. D. (2023). Trusting automation: Designing for responsivity and resilience. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *65*(1), 137–165. 001872082110099. https://doi.org/10.1177/00187208211009995

Clark, L., Pantidi, N., Cooney, O., Doyle, P., Garaialde, D., Edwards, J., Spillane, B., Gilmartin, E., Murad, C., Munteanu, C., Wade, V., & Cowan, B. R. (2019). What makes a good conversation?: Challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–12). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300705

Collier, W. (2015). *Application of ENA-based network analyses to fMRI data on school-children's acquisition of number symbols* [Paper presentation]. Discovery Challenge Research Symposium, Wisconsin Institutes for Discovery, Madison, WI.

Cooke, N. J., Gorman, J. C., Myers, C. W., & Duran, J. L. (2013). Interactive team cognition. *Cognitive Science*, *37*(2), 255–285. https://doi.org/10.1111/cogs.12009

Demir, M., Mcneese, N. J., Gorman, J. C., & Cooke, N. J. (2021). Exploration of team trust and interaction dynamics in human-autonomy teaming. *Cognitive Systems Research*, *46*, 3–12. https://doi.org/10.13140/RG.2.2.32213.55528

Desai, M., Medvedev, M., Vázquez, M., McSheehy, S., Gadea-Omelchenko, S., Bruggeman, C., Steinfeld, A., & Yanco, H. (2012). *Effects of changing reliability on trust of robot systems* [Paper presentation]. Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction - HRI '12, 73, Boston, MA, USA. https://doi.org/10.1145/2157689.2157702

Dessalles, J.-L. (2017). Conversational topic connectedness predicted by simplicity theory. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society, 1914–1919*. Cognitive Science Society.

Dunn, J. R., & Schweitzer, M. E. (2005). Feeling and believing: The influence of emotion on trust. *Journal of Personality and Social Psychology*, *88*(5), 736–748. https://doi.org/10.1037/0022-3514.88.5.736

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

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

Fernandez-Nieto, G. M., Martinez-Maldonado, R., Kitto, K., & Buckingham Shum, S. (2021). Modelling spatial behaviours in clinical team simulations using epistemic network analysis: Methodology and teacher evaluation. In *LAK21: 11th International Learning Analytics and Knowledge Conference* (pp. 386–396). Association for Computing Machinery. https://doi.org/10.1145/3448139.3448176

Fuoli, M., & Paradis, C. (2014). A model of trust-repair discourse. *Journal of Pragmatics*, *74*, 52–69. https://doi.org/10.1016/j.pragma.2014.09.001

Gao, J., & Lee, J. D. (2006). Extending the decision field theory to model operators' reliance on automation in supervisory control situations. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *36*(5), 943–959. https://doi.org/10.1109/TSMCA.2005.855783

Gao, J., Lee, J. D., & Zhang, Y. (2006). A dynamic model of interaction between reliance on automation and cooperation in multi-operator

multi-automation situations. *International Journal of Industrial Ergonomics*, *36*(5), 511–526. https://doi.org/10.1016/j.ergon.2006.01.013
Gottman, J., Swanson, C., & Swanson, K. (2002). A general systems theory of marriage: Nonlinear difference equation modeling of marital interaction. *Personality and social psychology review*, *6*(4), 326–340. https://doi.org/10.1207/S15327957PSPR0604_07
Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570
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, M., & Vera, A. (2019). No AI is an island: The case for teaming intelligence. *AI Magazine*, *40*(1), 16–28. https://doi.org/10.1609/aimag.v40i1.2842
Kaniarasu, P., & Steinfeld, A. M. (2014). Effects of blame on trust in human robot interaction. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication* (pp. 850–855). IEEE.
Kaplan, A. D., Kessler, T. T., Sanders, T. L., Cruit, J., Brill, J. C., Hancock, P. A., & LyOns, J. B. (2021). Chapter 6 - a time to trust: Trust as a function of time in human-robot interaction. In C. S. Nam (Eds.), *Trust in human-robot interaction* (pp. 143–157). Academic Press. https://doi.org/10.1016/B978-0-12-819472-0.00006-X
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*, 604977. https://doi.org/10.3389/fpsyg.2021.604977
Korsgaard, M. A., Kautz, J., Bliese, P., Samson, K., & Kostyszyn, P. (2018). Conceptualising time as a level of analysis: New directions in the analysis of trust dynamics. *Journal of Trust Research*, *8*(2), 142–165. https://doi.org/10.1080/21515581.2018.1516557
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://doi.org/10.1518/hfes.46.1.50_30392
Lewicki, R. J., & Bunker, B. B. (1996). *Developing and maintaining trust in work relationships*. Sage.
Li, M., Erickson, I., Cross, E., & Lee, J. (2022). Estimating trust in conversational agent with lexical and acoustic features. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *66*(1), 544–548. https://doi.org/10.1177/1071181322661147
Luo, R., Du, N., & Yang, X. J. (2022). Evaluating effects of enhanced autonomy transparency on trust, dependence, and human-autonomy team performance over time. *International Journal of Human–Computer Interaction*, *38*(18–20), 1962–1971. https://doi.org/10.1080/10447318.2022.2097602
Manson, J. H., Bryant, G. A., Gervais, M. M., & Kline, M. A. (2013). Convergence of speech rate in conversation predicts cooperation. *Evolution and Human Behavior*, *34*(6), 419–426. https://doi.org/10.1016/j.evolhumbehav.2013.08.001
Mayo, R. (2015). Cognition is a matter of trust: Distrust tunes cognitive processes. *European Review of Social Psychology*, *26*(1), 283–327. https://doi.org/10.1080/10463283.2015.1117249
Midden, C. J. H., & Huijts, N. M. A. (2009). The role of trust in the affective evaluation of novel risks: The case of CO_2 storage. *Risk Analysis : An Official Publication of the Society for Risk Analysis*, *29*(5), 743–751. https://doi.org/10.1111/j.1539-6924.2009.01201.x
Miller, C. A. (2005). Trust in adaptive automation: The role of etiquette in tuning trust via analogic and affective methods. In *Proceedings of the 1st international conference on augmented cognition* (pp. 22–27). American Psychological Association.
Miller, D. T., & Ross, M. (1975). Self-serving biases in the attribution of causality: Fact or fiction? *Psychological Bulletin*, *82*(2), 213–225. https://doi.org/10.1037/h0076486
Molloy, R., & Parasuraman, R. (1996). Monitoring an automated system for a single failure: Vigilance and task complexity effects. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *38*(2), 311–322. https://doi.org/10.1177/001872089606380211
Oktay, J. S. (2012). *Grounded theory*. Oxford University Press.
Rheu, M., Shin, J. Y., Peng, W., & Huh-Yoo, J. (2021). Systematic review: Trust-building factors and implications for conversational agent design. *International Journal of Human–Computer Interaction*, *37*(1), 81–96. https://doi.org/10.1080/10447318.2020.1807710
Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714
Schreckenghost, D., Milam, T., & Billman, D. (2014). Human performance with procedure automation to manage spacecraft systems. In *Proceedings of the 35th International Conference for Aerospace Experts, Academics, Military Personnel, and Industry Leaders* (pp. 1–16). IEEE.
Shaffer, D. W. (2017). *Quantitative ethnography*. Lulu.com.
Siebert-Evenstone, A. L., Irgens, G. A., Collier, W., Swiecki, Z., Ruis, A. R., & Shaffer, D. W. (2017). In search of conversational grain size: Modeling semantic structure using moving stanza windows. *Journal of Learning Analytics*, *4*(3), 123–139. https://doi.org/10.18608/jla.2017.43.7
Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, *25*(4), 359–393. https://doi.org/10.1111/j.1468-0017.2010.01394.x
Tan, S. C., Wang, X., & Li, L. (2022). The development trajectory of shared epistemic agency in online collaborative learning: A study combing network analysis and sequential analysis. *Journal of Educational Computing Research*, *59*(8), 1655–1681. https://doi.org/10.1177/07356331211001562
Waber, B., Williams, M., Carroll, J., & Pentland, A. (2015). A voice is worth a thousand words: The implications of the micro-coding of social signals in speech for trust research. In *Handbook of research methods on trust: Second edition* (pp. 302–312). Edward Elgar Publishing.
Weiler, D. T., Lingg, A. J., Eagan, B. R., Shaffer, D. W., & Werner, N. E. (2022). Quantifying the qualitative: Exploring epistemic network analysis as a method to study work system interactions. *Ergonomics*, *65*(10), 1434–1449. https://doi.org/10.1080/00140139.2022.2051609
Wooldridge, A. R., Carayon, P., Shaffer, D. W., & Eagan, B. (2018). Quantifying the qualitative with epistemic network analysis: A human factors case study of task-allocation communication in a primary care team. *IISE Transactions on Healthcare Systems Engineering*, *8*(1), 72–82. https://doi.org/10.1080/24725579.2017.1418769
Yang, X. J., Schemanske, C., & Searle, C. (2021). Toward quantifying trust dynamics: How people adjust their trust after moment-to-moment interaction with automation. *Human Factors*. Advance online publication. https://doi.org/10.1177/00187208211034716
You, S., Nie, J., Suh, K., Sundar, S. S. (2011). When the robot criticizes you … Self-serving bias in human-robot interaction. In *Proceedings of the 6th International Conference on Human-Robot Interaction* (pp. 295–296). IEEE.

About the authors

Mengyao Li is a Ph.D. candidate in the Industrial and System Engineering Department at the University of Wisconsin–Madison. Mengyao does research on trust in human-agent communication and cooperation. Her current studies focus on measuring and managing trust in virtual assistants during long-duration space exploration operations.

Amudha V. Kamaraj is pursuing a Ph.D. in Industrial and systems engineering from the University of Wisconsin-Madison. Her current research interests lie in human-technology interaction in the surface transportation and healthcare domain.

John D. Lee is Emerson Electric Quality & Productivity Professor at the Department of Industrial and Systems Engineering, University of Wisconsin–Madison. John does research in Cognitive Engineering with a focus on human-automation interaction.