Exploring Trust Indicators in Human-Agent Conversation with Epistemic Network Analysis

MENGYAO LI, VARSHINI KAMARAJ, and JOHN D. LEE, University of Wisconsin-Madison, USA

Trust greatly affects human-AI cooperation, especially under uncertain and risky situations, such as long-duration manned space missions. These situations usually require an unobtrusive and dynamic trust measure. Measuring trust using conversational data is a promising yet under-explored approach. Epistemic network analysis, a quantitative ethnographic approach, was used to explore the meaningful connections between trust indicators based on human-agent conversation in a simulated habitat maintenance task. We used the well-validated variable, reliability as a proxy to induce levels of trust in the study design. Results suggested that trust indicators significantly differed between reliability levels in conversation codes related to the system processes, t(36.16) = 4.69, p=0, Cohen's d=1.48. Compared to reported high level of trust, participants with low trust focused more on system process and misalignment information. These results help explain black box machine learning models that predict trust based on conversation data, and can help guide conversational agent design to manage trust.

CCS Concepts: • **Human-centered computing** → **HCI theory, concepts and models**; Laboratory experiments; **User models**.

Additional Key Words and Phrases: Trust in automation, network analysis, conversational analysis, epistemic network analysis, human-AI teaming

ACM Reference Format:

Mengyao Li, Varshini Kamaraj, and John D. Lee. 2022. Exploring Trust Indicators in Human-Agent Conversation with Epistemic Network Analysis. In *CHI-trAIt '22: ACM CHI Workshop on Trust and Reliance in Human-AI Teams, April 30, 2022, New Orleans, LA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

As artificial intelligence (AI) becomes increasingly capable and can possibly outperform humans in certain tasks , in the near future, human and AI may gradually work and cooperate as coworkers than tools, which shifts the human-AI relationships towards an interdependent teaming rather than a supervisory control [3, 7]. One example of this trend is the need for intelligent agents to support long duration space missions, where time delays can prevent communication with ground control. Without ground control, crew members will need to cooperate with a virtual agent to solve technical issues in the system. In such scenarios, distrust in the virtual agent can undermine coordination and cooperation throughout the task [2, 3]. Measuring trust in these interactions can better monitor and manage these interactions. However, current trust measures are generally self-reported, which is obtrusive. Additionally, trust measurements are often static, which cannot reflect the dynamic human-agent teaming. Using some types of continuous streams of data to infer and measure human trust during the human-agent interaction in the mission is ideal. Prior works have explored on physiological (e.g. heart rate and eye glancing) and behavioral measures of trust (e.g. decision time, delegation) in automation (TiA)[8]. Measuring trust using conversational data is a promising yet under-explored approach [10].

Manuscript submitted to ACM

To understand trust indicators in the conversation, prior research has shown that using a machine learning approach can combine lexical and acoustic features to predict trust in the conversational agent [11]. However, the meaning and connections between the features is not easily interpreted in a machine learning approach. Additionally, machine learning methods can not offer a complete picture of the dynamic nature of trust as they lack the qualitative information that is embedded in conversations between the human and the agent. Epistemic network analysis, a quantitative ethnographic technique, provides an alternative way to examine the meaningful connections between conversation content and trust. Thus, this paper explores trust indicators by leveraging epistemic network analysis to analyze human-agent conversations.

2 BACKGROUND

Trust, defined as "as the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability"[9], has emerged as a prevalent construct to describe relationships between people and and technology in myriad domains. Trust is a multidimensional (capable, ethical, sincere, and reliable), multilevel construct (dispositional, history-based, and situational) that spans over multiple domains (automation, e-commerce, and human)[1, 18]. For the purpose of this study, understanding the trust in conversational agent for long-duration space mission, we would focus on "automation" domain and evaluated the "history-based", and 'capable' and 'reliable' aspects of trust.

A critical challenge for advancing trust research is being able to measure trust in automation (TiA) precisely, and in a way that can generalize across multiple contexts. In contrast with behavioral and physiological measures, questionnaires (i.e., self-report) are more direct measures of trust because trust is inherently subjective; asking a person her beliefs, attitudes, and expectations is an important dimension of understanding that person's trust, which are straightforward to administer, can be rigorously developed based on trust theory, validated empirically, and can more easily generalize across task contexts. Although subjective rating has been treated as the gold standard of trust measurement, this approach suffers from some limitations when assessing human-agent relationships: 1) Since the survey is heavily text-based, the administration process often forces an interruption while people are interacting with the agent. 2) Direct descriptive statement in survey does not leave respondents with adequate freedom to identify, form, and explain their feelings and opinions based on the context. The contexual nature of trust is missing. Therefore, there is a need for an alternative or complementary trust measurement[10].

Communication is essential for trust building and calibration, which in turn, can promote effective human-AI teaming [4]. Although conversations are a rich source of information regarding trust, trust is not often explicitly referred to or verbalized in conversations. Thus, there is a need to identify the trust-related indicators. Prior research has used both qualitative and quantitative approaches to identify conversational indicators of trust. 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. Qualitative analysis, such as grounded theory, provides a rigorous and systematic approach to identify the situated meanings and systematic patterns in the data [13]. 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. To overcome this, we adopt an Epistemic Network Analysis (ENA), which is a quantitative ethnographic technique that leverages coded data and uses network representations along with metrics such as co-occurrences to analyze and interpret the connections between the coded data [14, 15]. ENA can systematically identify a set of meaningful features in the data based on the triangulation between human coders and computers. ENA networks can help visualize the structure of the conversation,

and the network metrics can also be analyzed statistically. Together, in the context of measuring trust in human-agent conversations, ENA helps to identify and understand trust indicators in human-agent conversations.

3 METHOD

3.1 Data

The data we analyzed came from a $2 \times 2 \times 3$ within-subject study. Participants conducted 12 decision-making tasks moderated by a human where they managed a Carbon Dioxide Removal System (CDRS) that is part of a analog Mars habitat. Participants were assisted by a conversational agent with 2 levels of agent reliability (i.e., high, and low). Each level of reliability had 2 cycles of the CDRS tasks, each including 3 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 total events (Table 1) 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. Once the participant finished the conversation, they then completed a trust survey [6].

A total of 24 participants (18 female, 6 male) were recruited from the Madison, WI area (M=23.7, SD =3.6). In total, each participant had the opportunity for 72 conversational turns with the agent. The cleaned text data contained 1981 lines of utterances, with the mean text length of 38.25 characters (SD=26.49). Additionally, we evaluated the relationship between reliability and trust. 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 condition (M= 4.37, SD = 1.44), $t(23)$=4.12, p=0.0002. Thus, for high-reliability condition, we can investigated the conversational indicators associated with the high trust, and vice versa.

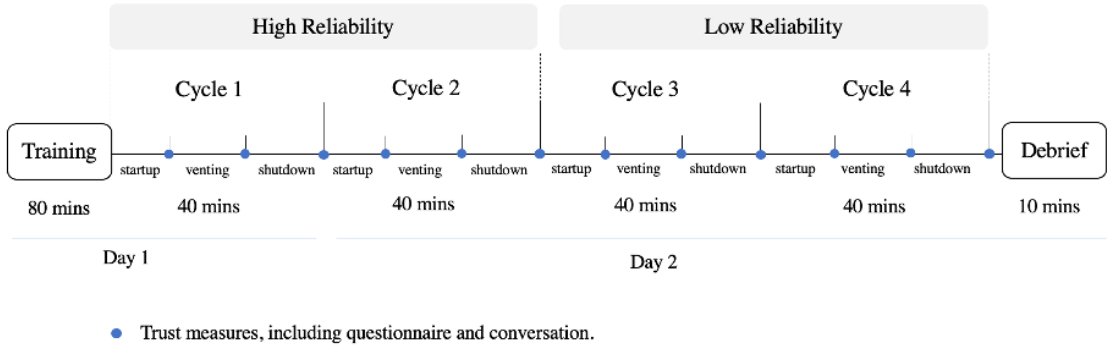

Fig. 1. Study Overview

3.2 Epistemic Network Analysis

To code the conversational data, we segmented the data into sentences and coded using an automated coding software (nCoder)[5], which uses regular expression matching. Eight codes were identified as shown in Table 1. These codes were selected and defined in an iterative round of coding, two researcher combined deductive and inductive coding to refine and validate codes. They discussed regularly to compare codes and categories, and re-coded certain discourses if needed. Any disagreements were resolved until all codes had reached Cohen's > 0.65 and Shaffer's ρ (0.9) < 0.05

Table 1. Code book of constructs included in Epistemic Network Analysis.

Code	Definition	Example From Data	Kappa	Rho
Familiarity	People getting more exposure with the system.	*"I felt stress-free and comfortable because I'm pretty **familiar** with the system from yesterday."*	0.92	0.01
User Capability	Self-efficacy. People's belief in his or her capacity to execute behaviors.	*"Bucky is incorrect this round but I'm **confident** in myself for choosing the correct one."*	0.94	0.04
Trust and Reliance	People's trust or distrust in the system and willingness to use the recommendations.	*"Bucky was wrong again so I'm starting to lose **trust** in him."*	0.97	0.01
Task Complexity	Perception of complexity of the task.	*"It was **pretty easy** honestly."*	0.98	0.02
Mismatched Procedure	A misalignment between the procedure and the provided diagram.	*"The procedure **lined up** well to the diagram."*	0.88	0.04
System Interpretability	System's transparency and clarity in providing information and reasoning.	*"Possibly getting the **reasons** of the diagram instead of having to click and wait for more. "*	0.97	0.03
Feedbacks	System updates and feedbacks.	*"Maybe providing like **verification** along the way that parts of the procedure were correct "*	0.79	0.05
System Process Scrutiny	People recall the specific system knowledge.	*"I am wondering if it is normal for the **CO2 levels** to decrease like that."*	0.84	0.01

between two human raters and the automated classifier. After validating each code, we applied the automated classifiers to the data set to code the data.

To compare the trust indicators in the conversation, ENA web tool was used by defining all lines of data as units and conversations as the collection of each participant's utterances [12, 16]. We defined units as each participant grouped by reliability condition, conversation as each CDRS event (i.e. startup, venting, shutdown), and comparison groups based on the reliability condition. The ENA algorithm uses a *moving window* to construct a network model for each line in the data, showing how codes in the current line are connected to codes that occur within the recent temporal context [17]. Codes that occurred outside of this window were not considered connected. We defined moving window as 2 lines (each line plus the previous line) within a given conversation. The resulting weighted networks are aggregated for all lines for each unit of analysis in the model. In this model, we aggregated networks using a binary summation in which the networks for a given line reflect the presence or absence of the co-occurrence of each pair of codes.

4 RESULTS AND DISCUSSION

Figure 2 and Figure 3 contain: (1) a plotted point, which represents the location of that unit's network in the low-dimensional projected space, and (2) a weighted network graph. The positions of the network graph nodes are fixed, which are determined by an optimization routine that minimizes the difference between the plotted points and their corresponding network centroids. Because of this co-registration of network graphs and projected space, the positions of the network graph nodes—and the connections they define—can be used to interpret the dimensions of the projected space and explain the positions of plotted points in the space. Our model had co-registration correlations of 0.8 (Pearson)

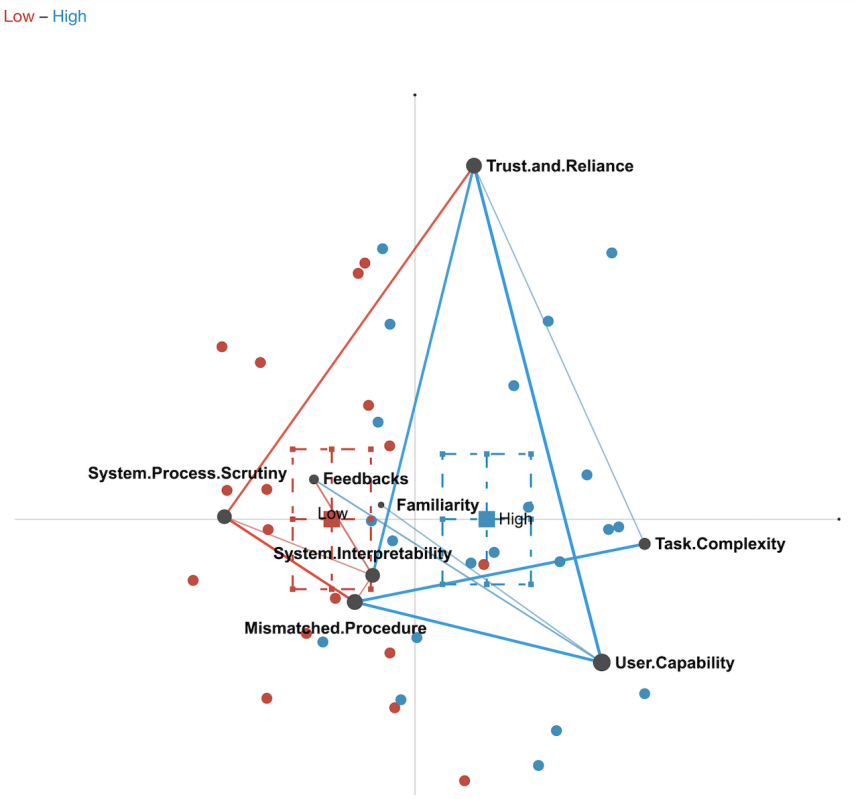

Fig. 2. ENA network of subtracted connections for high reliability (blue) versus low reliability (red)

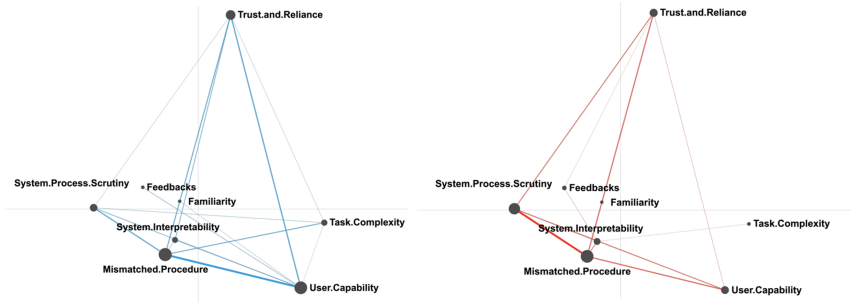

Fig. 3. ENA network for high reliability (left) and low reliability (right

and 0.79 (Spearman) for the first dimension and co-registration correlations of 0.81 (Pearson) and 0.83 (Spearman) for the second.

Figure 2 shows subtracted network graphs depicting the discourse differences between high reliability and low reliability and 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 or nodes connection within each conversation between participants and conversational agent. Thus, the thicker the edges, the stronger the node connection is observed in the human-agent conversation.

To interpret the results of ENA network, the x-axis and y-axis should be defined based on the code structure and domain knowledge. The X-axis reflects the codes that capture conversations related to the degree of **system processes**. These include task complexity, mismatched procedure, system interpretability, feedback on system performance, and system scrutiny. The left end of the axis shows the detailed processing of the system whereas the right end shows the general comments on the system complexity. The Y-axis shows the codes that reflect the **user's state** of trust in the system, self-efficacy (belief in their own capability) and familiarity.

The centroids presented in Figure 2 summarized the dimension of each network. Centroids indicated by boxes and confidence intervals (dotted lines) enable comparisons of networks statistically as well as visually. To test the differences between the reliability conditions, we applied two-sample t-test. Along the X axis, a two-sample t test showed that the high reliability condition (M=0.20, SD=0.28, N=22) was statistically significantly different at the α=0.05 level from the low reliability condition (M=-0.24, SD=0.32, N=19); $t(36.16) = 4.69$, p=0.00, Cohen's d=1.48). The significant results indicate that the conversation between high-reliability versus low-reliability condition differed along the system level conversation codes. When in high reliability, conversation around the system level process was more general (e.g. the task is easy). When in low reliability, the conversation was more centered on the system scrutiny (e.g. The CO2 is supposed to be at lower level). This difference indicates that when people in lower level of trust, the conversation would be more detailed-orientated around the system processing. Along the Y axis, a two sample t test assuming showed high reliability (M=0.00, SD=0.76, N=22 was not statistically significantly different at the α=0.05 level from low reliability (M=0.00, SD=0.75, N=19; t(38.27)= 0.00, p=1.00, Cohen's d=0.00). The result suggested that people's conversation around user capability does not change between hign and low reliability.

ENA subtracted network also provides the visual representation to explain the reasons for the statistical difference between nodes connections in high and low reliability conditions(shown in Figure 2). The `Trust and Reliance` node is important in both the high and low reliability group . The connected lines represents the subtracted connections or co-occurrences of two codes. Both groups connected `Trust and Reliance` with the `Mismatched Procedure` node. However, based on Figure 3, in the high reliability condition, the strongest connection is `Mismatched Procedures` and `User capability`, indicating that when rating their trust in the conversational agent, participants frequently discussed the procedure matching and their own self-efficacy together. In the low reliability condition, we noticed a stronger connection between `Trust and Reliance`, the `Mismatched Procedure`, and `System Process Scrutiny`. This means that in the low reliability condition, subjects expressed their low level of trust by comparing the `Mismatched Procedure` in the study and 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`)

5 CONCLUSION

To build better human-AI teaming, the AI needs to monitor and manage trust in real-time. Conversational data provides a novel approach to measure real-time trust. Prior approaches using quantitative analysis (e.g. machine learning, text analysis) or qualitative analysis (e.g. grounded theory), cannot provide *meaningful connections* between the trust indicators. We employed epistemic network analysis, a quantitative ethnographic approach that can systematically identify the patterns in the data while providing interpretable construct connections, on our human-agent conversational data. Results suggested that trust indicators significantly differed in conversation codes related to the system processes. Compared to high level of trust, people focused on scrutinizing the system process and misaligned information when

they interact with a low reliability agent. These results provided an initial step in identifying relevant trust indicators in the conversation. To manage the trust, the identified trust indicators can be further used to guide the conversation agent design.

6 ACKNOWLEDGEMENT

This work was supported by NASA Human Research Program No.80NSSC19K0654. We also thank members of the University of Wisconsin-Madison Cognitive Systems La-boratory for their insightful discussions and comments

REFERENCES

[1] Areen Alsaid, Mengyao Li, John Lee, and Erin Chiou. n.a.. *Understanding similarities and differences across trust questionnaires: A text analysis approach.*

[2] Erin K Chiou and John D Lee. 2016. Cooperation in human-agent systems to support resilience: A microworld experiment. *Human factors* 58, 6 (2016), 846–863.

[3] Erin K Chiou and John D Lee. 2021. Trusting automation: Designing for responsivity and resilience. *Human factors* (2021), 00187208211009995.

[4] Matteo Fuoli and Carita Paradis. 2014. A model of trust-repair discourse. *Journal of Pragmatics* 74 (2014), 52–69.

[5] C Hinojosa, AL Siebert-Evenstone, BR Eagan, Z Swiecki, M Gleicher, and C Marquart. 2019. nCoder.

[6] Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. 2000. Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics* 4, 1 (2000), 53–71.

[7] Matthew Johnson and Alonso Vera. 2019. No AI is an island: the case for teaming intelligence. *AI magazine* 40, 1 (2019), 16–28.

[8] Spencer C Kohn, Ewart J De Visser, Eva Wiese, Yi-Ching Lee, and Tyler H Shaw. 2021. Measurement of Trust in Automation: A Narrative Review and Reference Guide. *Frontiers in psychology* 12 (2021).

[9] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[10] Mengyao Li, Areen Alsaid, Sofia I Noejovich, Ernest V Cross, and John D Lee. 2020. Towards a Conversational Measure of Trust. *arXiv preprint arXiv:2010.04885* (2020).

[11] Mengyao Li, Erickson1 Isabel, Ernest V Cross, and John D Lee. 2022. Measuring Trust in Conversational Agent with Lexical and Acoustic Features. In *Proceedings of the Human Factors and Ergonomics Society.*

[12] Cody L Marquart, Zachari Swiecki, Brendan Eagan, and David Williamson Shaffer. 2019. *ncodeR: Techniques for Automated Classifiers.* https://CRAN.R-project.org/package=ncodeR R package version 0.2.0.1.

[13] Julianne S Oktay. 2012. *Grounded theory.* Oxford University Press.

[14] D Shaffer and A Ruis. 2017. Epistemic network analysis: A worked example of theory-based learning analytics. *Handbook of learning analytics* (2017).

[15] David Williamson Shaffer. 2017. *Quantitative ethnography.* Lulu. com.

[16] David Williamson Shaffer, David Hatfield, Gina Navoa Svarovsky, Padraig Nash, Aran Nulty, Elizabeth Bagley, Ken Frank, André A Rupp, and Robert Mislevy. 2009. Epistemic network analysis: A prototype for 21st-century assessment of learning. *International Journal of Learning and Media* 1, 2 (2009).

[17] Amanda Lee Siebert-Evenstone, Golnaz Arastoopour Irgens, Wesley Collier, Zachari Swiecki, Andrew R Ruis, and David Williamson Shaffer. 2017. In search of conversational grain size: Modeling semantic structure using moving stanza windows. *Journal of Learning Analytics* 4, 3 (2017), 123–139.

[18] Daniel Ullman and Bertram F. Malle. 2018. What Does It Mean to Trust a Robot? Steps Toward a Multidimensional Measure of Trust. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (Chicago, IL, USA) *(HRI '18).* Association for Computing Machinery, New York, NY, USA, 263–264. https://doi.org/10.1145/3173386.3176991