Predictive Dynamics of Trust in Human-Robot Interaction: A Recurrent Neural Network Approach

Xiaoyun Yin[1,2], Shiwen Zhou[1,2], Matthew J. Scalia[1,2], Ruihao Zhang[1,2] and Jamie C. Gorman[1,2]

[1]*Human Systems Engineering, Ira A. Fulton Schools of Engineering at Arizona State University, Mesa, AZ, USA*
[2]*Center for Human, Artificial Intelligence, and Robot Teaming, Arizona State University, Tempe, AZ, USA*

ABSTRACT

Over the years, trust in Human-Robot Interaction (HRI) has been extensively analyzed by researchers, not just as a static measure but also as a dynamic process. This paper proposes an approach using Recurrent Neural Networks (RNNs) to predict the dynamics of trust in Human-Robot interactions. To apply RNNs in HRI trust prediction, we propose segmenting time series into smaller windows and using Long Short-Term Memory (LSTM) cells to account for the temporal dynamics of trust.

KEYWORDS

Dynamic Trust, Human Robot Interaction, Trust Modeling, Recurrent Neural Network

1 Measurement of Trust in the Context of HRI

Human-Robot Interaction (HRI) extends beyond teleoperation, representing a dynamic field where robotics technology is designed to interact with humans in a range from simple responses to precise command execution and complex, semi- autonomous tasks [17]. This evolution in HRI highlights the significance of robots' autonomous decision-making, requiring interfaces that are efficient and intuitive for human users. As robots gain autonomy, the interaction naturally evolves into a partnership model, where trust becomes pivotal [5]. Establishing trust in HRI is crucial for seamless cooperation, ensuring that humans can rely on robots to perform tasks accurately and safely, thereby facilitating more integrated and productive human-robot collaborations.

Trust in HRI is defined in terms of reliance as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [8, p. 54]. The factors influencing this metric of trust include reliability [19], performance [16], and predictability [9], aligning with Lee and See's performance attribute of automation [8].

Affective- and cognitive based-trust have been identified as the foundation for interpersonal cooperation which extends to HRI [12]. Affective-based trust is rooted in the interpersonal aspects of the trusting attitude, related to reciprocal care and concern [12], and represents the emotional dimension of trust [2]. In contrast, cognitive-based trust is based on an individual's beliefs in another's reliability and dependability [12]. Cognitive trust, therefore, focuses on reliance as well as the performance, process, and ability attributes of trust in automation identified by Lee and See [8]. This paper primarily addresses cognitive trust.

Various methodologies have been developed to measure trust and trust dynamics. Moving beyond treating measurements of trust as a static "snapshot" at the end of an experiment, trust has been conceptualized and demonstrated as a variable that fluctuates over time, influenced by various factors including both positive and negative experiences [14, 18, 20]. Trust has been measured and validated in real-time situations, incorporating both subjective and objective assessments, using a layered dynamic model [3, 4] and a moving window procedure [22].

2 Related Studies

Recent studies, such as those by Li and colleagues [10], have utilized machine learning models to predict trust by analyzing communication content and flow, as well as conversational cues within people's conversations, combined with audio and text data. In this study, a random forest model that served as a partial mediator for predicting trust was built. However, their model did not account for the influence of time series data on trust detection. Given that trust is measured as a dynamic variable [20], machine learning models capable of processing time series, such as recurrent neural networks (RNNs), should be able to predict trust.

In this case, an RNN could be used to measure trust by segmenting the event and incorporating environmental information as input. It is anticipated that the RNN would output the categorization of trust/distrust or even quantify the level of change in trust/distrust. The trust events in HRI have an advantage with this strategy, as data from autonomous agents can be collected as one source of the environmental dataset.

RNNs have been widely used in natural language processing due to their ability to consider the sequence of words in a sentence [13]. This capability can be applied to analyzing sequences of events as well. Krishnan and colleagues [6] utilized an RNN to predict whether a URL is suspicious based on queries and sequential datasets. RNNs are also effective in predicting human behaviors; Li and colleagues [11] developed an RNN tree that categorizes various types of human actions and can adapt to new action classes. [15] used sequential actions as dynamic data for modeling and predicting human sequential design decisions. Additionally, RNNs are useful for cleaning high-noise datasets and categorizing different signals. Kuanar and colleagues [7] employed an RNN to process EEG data, which is inherently noisy, to produce four-class predictions. However, this discussion transcends the

simple increasing or decreasing of trust levels to focus instead on the dynamics of trust—its increase or decrease over time. Future research might explore the dynamics of distrust in comparison to trust.

To predict trust in real-time, we need to segment the time series into smaller windows. For instance, consider a scenario where three teammates, including one robotic teammate, collaborate to photograph specific targets. Here, the time period associated with each target can constitute a segment. In a specific sequence, the robotic teammate communicates task-related information, sending a message to teammate A, who then responds. This prompts the participant to inquire further about the task. The tasks assigned to the robotic teammate, the message they send, teammate A's response, and the participant's question all follow a sequence of events, and any alteration in this sequence could impact the level of trust.

Given that trust propensity is important for inspire trust [1], people's initial level of trust needs to be considered. Therefore, the use of Long Short-Term Memory (LSTM; [21]) seems well-suited for RNN cells in this context as LSTM cells contain self-connected memory cells that can store information even at the beginning of the mission. Starting with a single LSTM layer can serve as a baseline to gauge how the model predicts trust. The system will then incorporate data from various events that have been previously defined; thus, the input data will be the values of the events from different segments, ensuring the model's simplicity. Nonetheless, exploring the potential of adding second or third layers may enhance trust prediction.

The challenge that remains with this approach is the ability to differentiate event types from the data collected in the environment, such as distinguishing between changes in the robotic teammates' behavior and the robotic teammates' response.

3 Conclusion

The research conducted to date demonstrates that RNNs can process sequential data as input and generate categorical data as output. This aligns with the objective of this paper: to input a series of actions within the system and output the changes in trust level, either as an increase or decrease. The binary categorical output will benefit future studies on dynamic measurements of trust. Practically, predicting trust can lead to faster responses for trust repair, thereby enhancing team performance. It is important to note that this RNN model is designed solely as a classifier for the increase or decrease of trust levels, not for analyzing the level of trust. Future research should aim to evolve this model into a more predictive tool to test if it can detect the degree of change in trust.

REFERENCES

[1] Jason A Colquitt, Brent A Scott, and Jeffery A LePine. 2007. Trust, trustworthiness, and trust propensity: a meta-analytic test of their unique relationships with risk taking and job performance. *Journal of applied psychology* 92, 4 (2007), 909.

[2] David Dowell, Mark Morrison, and Troy Heffernan. 2015. The changing importance of affective trust and cognitive trust across the relationship lifecycle: A study of business-to-business relationships. *Industrial Marketing Management* 44 (2015), 119–130.

[3] Jamie C Gorman, Mustafa Demir, Nancy J Cooke, and David A Grimm. 2019. Evaluating sociotechnical dynamics in a simulated remotely- piloted aircraft system: A layered dynamics approach. *Ergonomics* 62, 5 (2019), 629–643.

[4] David AP Grimm, Jamie C Gorman, Nancy J Cooke, Mustafa Demir, and Nathan J McNeese. 2023. Dynamical Measurement of Team Resilience. *Journal of Cognitive Engineering and Decision Making* 17, 4 (2023), 351–382.

[5] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors* 53, 5 (2011), 517–527.

[6] N Krishnan and Gerard Deepak. 2021. Towards a novel framework for trust driven web URL recommendation incorporating semantic alignment and recurrent neural network. In *2021 7th International Conference on Web Research (ICWR)*. IEEE, 232–237.

[7] Shiba Kuanar, Vassilis Athitsos, Nityananda Pradhan, Arabinda Mishra, and Kamisetty R Rao. 2018. Cognitive analysis of working memory load from EEG, by a deep recurrent neural network. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2576–2580.

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

[9] Stephan Lewandowsky, Michael Mundy, and Gerard Tan. 2000. The dynamics of trust: comparing humans to automation. *Journal of Experimental Psychology: Applied* 6, 2 (2000), 104.

[10] Mengyao Li, Isabel M Erickson, Ernest V Cross, and John D Lee. 2023. It's Not Only What You Say, But Also How You Say It: Machine Learning Approach to Estimate Trust from Conversation. *Human Factors* (2023), 00187208231166624.

[11] Wenbo Li, Longyin Wen, Ming-Ching Chang, Ser Nam Lim, and Siwei Lyu. 2017. Adaptive RNN tree for large-scale human action recognition. In *Proceedings of the IEEE international conference on computer vision*. 1444–1452.

[12] Daniel J McAllister. 1995. Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of management journal* 38, 1 (1995), 24–59.

[13] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernocky`, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Interspeech*, Vol. 2. Makuhari, 1045–1048.

[14] Surya Nepal, Wanita Sherchan, and Athman Bouguettaya. 2010. A behaviour-based trust model for service web. In *2010 IEEE International Conference on Service-Oriented Computing and Applications (SOCA)*. IEEE, 1–4.

[15] Molla Hafizur Rahman, Shuhan Yuan, Charles Xie, and Zhenghui Sha. 2020. Predicting human design decisions with deep recurrent neural network combining static and dynamic data. *Design Science* 6 (2020), e15.

[16] Alessandra Rossi, Kerstin Dautenhahn, Kheng Lee Koay, and Michael L Walters. 2018. The impact of peoples' personal dispositions and per- sonalities on their trust of robots in an emergency scenario. *Paladyn, Journal of Behavioral Robotics* 9, 1 (2018), 137–154.

[17] Jean Scholtz. 2003. Theory and evaluation of human robot interactions. In *36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of the*. IEEE, 10–pp.

[18] Steffen Staab, Bharat Bhargava, L Leszek, Arnon Rosenthal, Marianne Winslett, Morris Sloman, Tharam S Dillon, Elizabeth Chang, Farookh Hussain, Wolfgang Nejdl, et al. 2004. The pudding of trust. *IEEE Intelligent Systems* 19, 5 (2004), 74–88.

[19] Julia L Wright, Jessie YC Chen, and Shan G Lakhmani. 2019. Agent transparency and reliability in human–robot interaction: The influence on user confidence and perceived reliability. *IEEE Transactions on Human-Machine Systems* 50, 3 (2019), 254–263.

[20] X Jessie Yang, Christopher Schemanske, and Christine Searle. 2023. Toward quantifying trust dynamics: How people adjust their trust after moment-to-moment interaction with automation. *Human Factors* 65, 5 (2023), 862–878.

[21] Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. 2019. A review of recurrent neural networks: LSTM cells and network architectures. *Neural computation* 31, 7 (2019), 1235–1270.

[22] Shiwen Zhou, Xioyun Yin, Matthew J Scalia, Ruihao Zhang, Jamie C Gorman, and Nathan J McNeese. 2023. Development of a Real-Time Trust/Distrust Metric Using Interactive Hybrid Cognitive Task Analy- sis. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 67. SAGE Publications Sage CA: Los Angeles, CA, 2128– 2136