Artificial Intelligence and Social Computing, Vol. 28, 2022, 92–100
https://doi.org/10.54941/ahfe1001451

Supradyadic Trust in Artificial Intelligence

Stephen L. Dorton

Human-Autonomy Interaction Laboratory, Sonalysts, Inc. Waterford, CT, USA

ABSTRACT

There is a considerable body of research on trust in Artificial Intelligence (AI). Trust has been viewed almost exclusively as a dyadic construct, where it is a function of various factors between the user and the agent, mediated by the context of the environment. A recent study has found several cases of supradyadic trust interactions, where a user's trust in the AI is affected by how other people interact with the agent, above and beyond endorsements or reputation. An analysis of these surpradyadic interactions is presented, along with a discussion of practical considerations for AI developers, and an argument for more complex representations of trust in AI.

Keywords: Trust, Artificial intelligence, Human-machine teaming

INTRODUCTION

Artificial Intelligence in Intelligence Analysis

Intelligence analysis is the continuous cycle of planning, collecting, processing, exploiting, and disseminating information to support decision making (Clark, 2014). There has long been a trend where intelligence analysts have been unable to keep up with an already substantial, and increasing quantity of data to analyze (Menthe et al., 2012). Artificial Intelligence (AI) has been viewed as the means by which analysts will cope with extreme data quantities (Symon and Tarapore, 2015); however, there are many factors that affect whether intelligence analysts trust, and therefore adopt AI systems into their workflow (Dorton and Harper, 2021; Dorton and Harper, 2022). Therefore it is important to develop a rich understanding of trust dynamics in AI, which may ultimately "make or break" our ability to cope with increased volumes of data.

Trust: Multifaceted, Dynamic, and Dyadic

Trust has been studied considerably and can be summarized as the degree to which one is willing to be vulnerable by putting themselves in the hands of another agent (human, AI, etc.) (Lee and See, 2004). Trust is critically important since it plays a substantial role in whether the agent is adopted into the larger sociotechnical system (Lyons et al., 2016; Schaefer et al., 2016). Regarding AI or autonomy more broadly, trust must be calibrated based on the fitness of the AI to complete tasks in the context of the environment, as too much or too little trust can be problematic (Dorton and Harper, 2022; Lee and See, 2004).

Trust is a complex and multifaceted phenomenon. Previous research has identified dozens of factors that affect trust in autonomy, which can be attributed to the individual, the autonomous agent/system, or the environment (Schaefer et al., 2016; Siau and Wang, 2018; Hoff and Bashir, 2015; Hancock et al., 2011). Recent research has explored whether these factors were present when intelligence analysts gained or lost trust in AI (Dorton and Harper, 2021; Dorton and Harper, 2022). Another important aspect of trust is its dynamic nature. That is, the level of trust in an AI system is calibrated, or incremented and decremented, as the user interacts with it to accomplish work (Yang et al., 2021; Stevens et al., 2015).

Finally, trust has nearly exclusively been viewed as a dyadic construct, where trust is framed as a relationship (comprised of various factors) between a dyad of a human and an agent, mediated by the context of the environment (e.g. the resources, processes, and constraints affecting the human-agent dyad) (Schaefer et al., 2016; Hancock et al., 2011). This dyadic model or representation of trust is not exclusive to autonomy, robotics, or AI, and is widespread throughout the management and organizational literature. Some such research has focused exclusively on the human-agent dyad, ignoring the environment (Stevens et al., 2015); or merely acknowledged the environment as a factor (Kessler et al., 2017). This dyadic construct persists regardless of the two agents comprising the dyad, where trust has been framed in terms of manager-subordinate, salesperson-customer, or supplier-buyer (Moustafa-Leonard, 2007; Lussier et al., 2017; Dahwa et al., 2013).

Research Objective

In a recent study on factors affecting trust in AI (Dorton and Harper, 2021; Dorton and Harper, 2022), there were several instances where the trust in AI was affected by how *other* humans outside of the user-agent dyad interacted with the AI agent. These interactions were above-and-beyond mere endorsements or the reputation of the AI (Siau and Wang, 2018), and included how other humans in the larger sociotechnical work system developed, maintained, or otherwise used, misused, or abused the outputs of the AI. Based on these findings, the objective of this work is to identify and characterize various supradyadic interactions (i.e. how other humans interact with the AI) that affect trust in an AI system. More specifically, the objective is to identify how people outside of the user-AI dyad can affect the user's trust in the AI, to generate actionable findings for developers of such systems.

METHODS

This research built upon an existing dataset from a previous study, so I will discuss relatively few particulars here. Readers can find a more exhaustive explanation of the data collection procedure, demographics of participants, and characteristics of the dataset in other sources (Dorton and Harper, 2021; Dorton and Harper, 2022). The critical incident technique was used to collect data about incidents where intelligence professionals (collectors, analysts, etc.) gained or lost trust in an AI system in the context of their work in intelligence (Flanagan, 1954). Each participant provided one incident where they gained or lost trust in AI, with the exception of one participant who provided two, resulting in a sample of 30 incidents. Participants had expertise in various intelligence disciplines, and had careers across numerous organizations in the US Military and Intelligence Community. I used an iterative thematic analysis process to identify high-level themes in the dataset, and then identified more granular themes within them (Sherwood et al., 2020).

RESULTS

Thematic analysis resulted in the identification of 27 separate supradyadic interactions where participant trust in the AI was affected by how other humans interacted with the AI. These interactions occurred at three relative points in time: Before use, during use, and after use of the AI. The following subsections provide a more detailed analysis of the results that are summarized in Table 1.

Before Use Themes

The majority of supradyadic interactions ($n = 17$, 63%) occurred prior to the participant's use of the AI, such as the development of the AI system, or the process of curating or annotating data to train the AI model. There were two themes for interactions occurring prior to participant use of the AI.

Users in Development

As reported in Reference (Dorton and Harper, 2022), there were 10 cases where trust in the AI was gained or lost based on whether other end users or subject matter experts were involved in the development of the AI. Participants gained trust in the AI knowing that end users or domain experts were involved in the development of the AI system (e.g. design of algorithms or logic, feature engineering, or designing model outputs and user interfaces). Conversely, trust was lost when the participants were aware that end users were not involved in the conceptualization or design of the AI. Participants cited problems in successful employment of AI systems because the developers did not understand what was needed, "They knew the math behind it... they couldn't translate it for the [users]... if somehow the developers knew what we were trying to achieve... [it would have succeeded]" (#4).

System Inputs

There were seven cases where participants gained or lost trust in the AI based on how other people annotated data, or otherwise curated datasets for training the model(s). Most commonly, participants lost trust in the AI because they did not trust the other people who were entering data into the system. This was more general, but participants also cited specifically that the people tagging the data were not experienced enough, "I don't think they understood the need for quality. They should have made teams with senior people so they could make sure it was good" (#50). Conversely, in one case, a participant trusted the AI more because they knew that another human was verifying

Table 1. Summary of thematic analysis results.

Theme	*n*	Example Quote(s)
		Before AI Use ($n = 17$)
Users in Development	10	"They missed a huge part… and they did not include the [users] enough… Sometimes it needs to include less technical focus and more involvement of the people who make these decisions every day without computers." (#10) "They are working with industry and younger [users]... not what they've always done. I have more trust in it." (#36)
System Inputs	7	"Some analysts are good, some are OK, and some are bad… by the time they got it all [annotated] they didn't have uniform quality in their training set… They didn't have ground truth, [but] they trained the model on it…" (#50) "Depends on how other units do this process-whether or not they do [Quality Control]…" (#22)
		During AI Use ($n = 7$)
Confirmation	4	"I talked to [another intelligence cell] and they confirmed the threat [the AI identified] was in the area." (#39) "We got feedback from [analysts at organization] that it must not have been accurate, but it turned out that it was, in fact, accurate." (#53)
Requisite Knowledge	2	"The abilities of the human specialist in the loop is decreasing… they aren't able to pick out errors. The experts don't really have the same expertise; they don't have to think critically about doing it." (#6)
Technical Support	1	"When there is a big failure we know we can get it fixed… the fact that I can yell across the hallway and get answers and fixes quickly is a big factor in my trust." (#24)
		After AI Use ($n = 3$)
(Mis)Use of Outputs	2	"We take a month or longer to find the actual [answer], so when our number comes out, the difference with the AI was off. We will [conduct analysis] with a range that is less satisfying to people that already latched on to the AI output. We essentially had to tell people that they had to tell their bosses that they were wrong. (#4)
Feedback Loop	1	"I only talked to people in my branch- they basically just commiserated… I didn't have any insight on [how to give feedback to] the developers or maintainers." (#31)

data being fed into the system. Some participants also cited lost trust because of poor practices in data annotation, "Analysts were not allowed to rate their confidence for [annotation], or say 'no' or 'I don't know,' they had to make a call" (#50).

During Use Themes

There was a plurality of supradyadic interactions ($n = 7, 26\%$) that occurred while the participant, or their colleagues were using the AI. There were three themes for these interactions.

Confirmation

There were four cases where participants gained or lost trust in the AI based on confirmation or disconfirmation of the AI's outputs from other people who were not using the AI (i.e. arriving at conclusions independent of the AI). As one may expect, trust was gained when the AI outputs were confirmed by external sources. While this dataset does not provide specific examples, it stands to reason that other people disconfirming the AI's outputs would decrease the user's trust in the AI.

Requisite Knowledge

There were two cases where participants lost trust in AI because other colleagues using the AI were unable to recognize, and therefore resolve, any issues or errors in the AI. In both cases participants commented that the introduction of AI to the domain created a dependency on the AI. Because of this dependency, colleagues no longer learned or maintained the requisite knowledge required to understand the domain in which the AI operated. Participants lost trust in the AI because many of their colleagues could not recognize when the AI was misperforming.

Technical Support

There was one case where a participant gained trust in the AI because there was another person (a technical expert or developer) available to provide technical support and troubleshoot issues with the AI while the participant was interacting with it.

After Use Themes

There were few supradyadic interactions ($n = 3$, 11%) that took place after the participant used the AI. There were two themes for these interactions.

(Mis)Use of Outputs

There were two cases where trust was lost because colleagues misused the outputs of the AI (i.e. using verifiably incorrect outputs), which caused rework or performance issues in the larger work system in which the participant worked. Further, trust was also lost in the AI when it was not readily apparent how human colleagues were using the AI outputs, "[I don't know] whether [the AI] didn't see it, or it saw it, flagged it, and a human disregarded it… A human could have saw it and said well [criteria wasn't met] so I won't take action yet" (#44).

Feedback Loop

There was one case where a participant lost trust in the AI because there was no means to provide feedback for system improvement to the developers of the system. This is similar to the technical support theme, although it was different in that the participant was seeking technical support after using the AI (i.e. not in situ).

DISCUSSION

This research has identified supradyadic interactions with AI that affect user trust. Further, it provides insights on the nature of such interactions and when they may occur (i.e. before, during, or after the AI was used to complete work). These results support the notion that trust in AI is a complex phenomenon, and may require a supradyadic representation. These findings provide several implications for practitioners- those who conceptualize, design, build, deploy, or otherwise maintain intelligent systems. Given the importance of context in designing AI-based systems, it is unwise to write overly prescriptive guidelines. Thus, I offer the following higher-level considerations for AI practitioners to ask themselves, end users, and other stakeholders, as early as possible in development of AI systems:

1. Have end users or domain experts been sufficiently involved in the development process? Have they been involved early enough? Often enough?
2. How do users currently make decisions without AI? What information and heuristics do they use? Have these been adequately codified in the AI?
3. How will the users employ the AI in operational contexts? Does the AI provide outputs that facilitate operational decision making and workflows?
4. How will AI input quality be managed? Who will annotate data? What knowledge, skills, or abilities should or will they have?
5. Are those providing inputs able to capture their level of confidence, or refuse to provide an annotation when they are unsure? Where might they go wrong?
6. What other sources of information (e.g. people or sensors) might users rely on to confirm AI outputs? Are outputs in a format that easily facilitates this?
7. How might the introduction of the AI drive skill decay in previously manual tasks? What adverse second-order effects might require mitigation?
8. Will users be able to identify when the AI is misperforming? How?
9. How will users receive technical support during operational use of the AI?
10. How might users misuse the outputs of the AI? Will they be aware of the capabilities, limitations, or the conditions for which its use is validated?
11. How will users or other third parties choose to accept or reject AI outputs? Will those criteria or thresholds be personally held by individuals or collectively known by others within the broader sociotechnical system?

12. What feedback loops exist for users to report issues or desired improvements? Is funding secured and is management committed to act on these requests and continuously improve the AI through its expected lifecycle?

Looking forward, I argue for a supradyadic representation of trust in human-AI teams. Although this is relatively specific in nature, others have argued more generally for increased complexity in models of trust (Hoffman, 2017), or at least for a more dynamic view of trust (Hancock, 2017). Given these calls for more complicated models of trust, and the results of this study, I offer a "folk theory" of supradyadic trust for consideration (Dorton and Thirey, 2017). Figure 1 provides a notional supradyadic representation of trust in contrast to the current one.

The current representation frames trust (T) as a function of different factors ($F = \{v_1, v_2, ..., v_m\}$) between a user ($U$) and an agent ($A$), in the context of the environment (E) that they are working in. Denoting v_i as the level present of trust factor i, for the m factors present, then this dyadic model of trust can be expressed as

$$T_u^E(A) \sim f(v_1, v_2, \ldots, v_m) \tag{1}$$

A notional supradyadic representation builds upon Eq.1 by explicitly accounting for the various interactions (I) described herein, that other humans (H) have with the agent. Denoting the level of the j^{th} interaction as h_{j}, for the n interactions present, a supradyadic model of trust can be expressed as

$$T_u^E(A) \sim f(v_1, v_2, \ldots, v_m, h_1, h_2, \ldots, h_n) \tag{2}$$

In conclusion, this study provides notional support for a supradyadic representation of trust. To apply a more critical view, this is still likely a grossly oversimplified representation of trust, and more work is required to validate these claims and further build our collective understanding of trust in AI. Further, it is possible that these supradyadic interactions are not actually related to trust; one could readily argue that concepts such as *fitness* may be more appropriate to describe these findings (Rayo et al., 2020). Further,

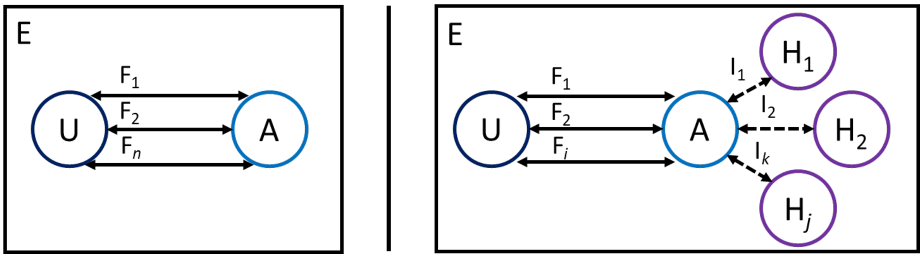

Figure 1: Dyadic (Left) and supradyadic (Right) representations of trust.

the interaction of others feeding the AI bad data may not be a supradyadic *interaction*, but simply signaling of a performance issue (a well-known factor of trust in dyadic representations). I hope that additional research and critical discussion are to follow this initial contribution.

ACKNOWLEDGMENT

This work was supported in part by the US Army Combat Capabilities Development Command (DEVCOM) under Contract No. W56KGU-19-C-0062. The views, opinions, and/or findings contained in this report are those of the author and should not be construed as an official Department of the Army position, policy, or decision unless so designated by other documentation. This document was approved for public release on 17 February 2022, Item No. A562. I would like to thank Ian Frommer and Angelique Wetzel for their helpful inputs on this manuscript.

REFERENCES

Clark, R. M. (2014) Intelligence collection. SAGE.

Dahwa, M.P., Al-Hakim, L., & Ng, E. (2013) The importance of trust in procurement practices and its impact on business performance: An empirical investigation from the perspective of the buyer-supplier dyad. Journ. of Rel. Mark., 12(4), 280–300.

Dorton, S. & Thirey, M. (2017) Effective variety? For whom (or what)? A folk theory on interface complexity and situation awareness. Proc. of 2017 IEEE CogSIMA.

Dorton, S.L. & Harper, S. (2021) Trustable AI: A critical challenge for naval intelligence. CIMSEC.

Dorton, S.L. & Harper, S.B. (2022) A naturalistic investigation of trust, AI, and intelligence work. Journ. of Cog. Eng. and Dec. Mak. doi: https://doi.org/10.1177/15553434221103718.

Flanagan, J. C. (1954) The Critical Incident Technique. Psych. Bull., 5, 327-358.

Hancock, P.A. (2017) Imposing limits on autonomous systems. Ergo., 60(2), 284–291.

Hancock, P.A., Billings, D.R., Schaefer, K.E., Chen, J.Y., de Visser, E.J., Parasuraman, R. (2011) A meta-analysis of factors affecting trust in human-robot interaction. Hum. Fac., 53(5), 517–527.

Hoff, K.A. & Bashir, M. (2015) Trust in automation: Integrating empirical evidence on factors that influence trust. Hum. Fac., 57(3), 407-434.

Hoffman, R.R. (2017) A Taxonomy of Emergent Trusting in the Human-Machine Relationship. In: Smith, P.J., Hoffman, R.R. (eds). Cognitive Systems Engineering: The Future for a Changing World, pp. 137-163. Taylor & Francis.

Kessler, T.T., Larios, C., Walker, T, Yerdon, V., & Hancock, P.A. (2017) A comparison of trust measures in human-robot interaction scenarios. In P. Savage-Knepshield & J. Chen (Eds.), Advances in Human Factors in Robots and Unmanned Systems, Advances in Intelligent Systems and Computing, 499, 353–364.

Lee, J., & See, K. A. (2004) Trust in automation: Designing for appropriate reliance. Hum. Fac., 46(1), 50–80.

Lussier, B., Gregoire, Y, & Vachon, M.A. (2017) The role of humor usage on creativity, trust and performance in business relationships: An analysis of the salesperson-customer dyad. Ind. Mark. Mgmt., 65, 168-181.

Lyons, J.B., Ho., N.T., Koltai, K.S., Masequesmay, G., Skoog, M., Cacanindin, A., & John-son, W.J. (2016)Trust-based analysis of an Air Force collision avoidance system. Erg. in Des., 24(1), 9–12.

Menthe, L., Cordova, A., Rhodes, C., Costello, R., & Sullivan, J. (2012) The Future of Air Force Motion Imagery Exploitation: Lessons from the Commercial World. Technical Report, RAND Corporation, Santa Monica, CA.

Moustafa-Leonard, K. (2007) Trust and the manager-suboordinate dyad: Virtual work as a unique context. Journ. of Beh. and App. Mgmt., 8(3), 197-201.

Rayo, M.F., Fitzgerald, M.C., Gifford, R.C., Morey, D.A., Reynolds, M.E., D'Annolfo, K., & Jeffries, C.M. (2020) The need for machine fitness assessment: Enabling joint human-machine performance in consumer health technologies. Proceedings of the 2020 Int. Symp. on Hum. Fact. and Ergo. in Health Care, 9(1), 40–42.

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016) A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. Hum. Fac., 58(3), 377–400.

Sherwood, S. M., Neville, K. J., McLean, A. L. M. T., Walwanis, M. M., & Bolton, A. E. (2020) Integrating new technology into the complex system of air combat training. In H. A. H. Handley & A. Tolk (Eds). A framework of human systems engineering: Applications and case studies, (pp. 185–204). Wiley.

Siau, K. & Wang, W. (2018) Building trust in artificial intelligence, machine learning, and robotics. Cutter Bus. Tech. Journ., 31(2), 47–53.

Stevens, M., MacDuffie, J.P., & Helper, S. (2015) Reorienting and recalibrating inter-organizational relationships: Strategies for achieving optimal trust. Org. Stud., 36(9), 1237–1264.

Symon, P. B. & Tarapore, A. (2015) Defense intelligence analysis in the age of big data. Joint Forces Quart., 79(4), 4–11.

Yang, X.J., Schemanske, C., & Searle, C. (2021) Toward quantifying trust dynamics: how people adjust their trust after moment-to-moment interaction with automation. Hum. Fac., 1–17.