Real-time Estimation of Human and Robot Trust

Chris Baber[1], Sagir Yusuf[1], Edmund Hunt[2], Sanja Milivojevic[2] & Patrick Waterson[3]

[1]University of Birmingham, [2]University of Bristol, [3]Loughborough University

SUMMARY

When people and robots work together, performance can be affected by perceived trust. Typically, trust is measured through self-report from people, often at the end of an experiment. We present a simple approach to estimate trust levels from data during performance. The approach not only allows us to estimate trust that humans have in the robots they work with but also trust that the robots might have in their human teammates. We propose that this could be used to identify degradation in trust and provides opportunity to repair trust before interaction is impaired.

KEYWORDS

Robots, Trust, Human-Robot Interaction, Bayesian Belief Network

Introduction

Robots and humans often need to work together, particularly in safety critical domains or in situations that require close cooperation. Ideally, team members should work toward goals common to the team. However, there may be situations in which individual goals become more important than team goals, e.g., the individual might need to respond to threat by performing actions that do not contribute to the team's goals or might have an opportunity to perform an action that satisfies its own goals rather than those of the team. This choice (of individual rather than team goals) could affect the trust that other team members have in the individual. Much of the research on human-robot trust sees trust as a disposition of a human directed towards a robot (Muir, 1994; Muir and Morey, 1996; Büttner et al., 2024). For Lee and See (2004), human trust is characterised by ". . . the attitude that a robot will help achieve an individual's goals in a situation characterised by uncertainty and vulnerability" [p. 54]. Such a disposition will depend on the personality, experiences, attitudes etc., of the human and the trustworthiness of its teammates. It is unlikely that a robot would acquire such breadth of knowledge. Therefore, we require a concept of 'artificial trust' (Jorge et al., 2024) that a robot could hold and for there to be sufficient overlap between this artificial and human trust. For a robot to define trust in a similar way to humans, it would need to recognise uncertainty in its environment and risk to itself and interpret the action of another robot in terms of its (or the team's) goals. This implies that trust is intrinsically related to the robot's ability to maintain situation awareness of itself and its teammates.

We believe that the simplest definition of trust that can inform such situation awareness can be developed using three dimensions adapted from (Lewis and Marsh, 2023):

- Capability, i.e., is that teammate appropriate for a given task in that situation?
- Predictability, i.e., is that teammate acting appropriately to its situation?
- Integrity, i.e., is that teammate acting to support the team?

In a similar manner, Jorge et al. (2024) define trust as a relational construct in which observations of an actor are considered in terms of ability, benevolence, and integrity. We differ from this only through replacing benevolence with predictability (although both concepts relate

to the expected outcomes of the actions of a teammate). The dimensions of capability and predictability are common in Human-robot Trust research (Hancock et al., 2011; Malle and Ullman, 2021), but 'integrity' has received less attention. Lewis and Marsh (2022) treat 'integrity' as a matter of mechanics rather than morality (i.e., whether a rock can support a person's weight as the rock's 'structural integrity'). For Jorge et al. (2024) integrity is a matter of whether a robot will truthfully report its action (where robots have an incentive for misreporting). Rather than rely on the honesty of the robot, we prefer to operationalise integrity in terms that reflect the task and focus on observable (rather than reportable) data.

In our definition, trust is context-dependent and varies over the course of a mission, which means that trust will be satisficed (using Simon's 1955 term) rather than optimised. By providing robots with the ability to formulate and act upon trust in their human teammates, we echo the aim of Jorge et al. (2024) to "enable AI teammates to delegate or decide how to rely on their human teammates, taking into account the team's goal and possible risks." [p. 52].

A Bayesian Belief Model of Trust

In order to measure trust, using the three-component model outlined previously, we require analogues for each component. For the experiment reported in this paper, the relationship between the components of trust and the available measures are shown in table 1. The experiment task required a human to work with two robots to scan QR (quick response) codes that represented 'tokens' that could be either collected individually (i.e., one agent, human or robot, could scan the QR code to collect the token and increase their individual score) or as a pair (i.e., the robot would activate the token and the human collect it and both would have their scores increased). When the human encountered a token that required activation, they could call a robot to help them. More details of the experiment are presented in the next section, but we hope this is sufficient for the reader to make sense of table 1.

Table 1: Relating Observable Measures of Performance to Components of Trust

Component of Trust model	Measure	Rationale
Capability (high capability will lead to high trust)	Count of individual tokens collected by an agent	More capable agents will collect more tokens
Predictability (high predictability will lead to high trust)	Count of messages for help	Agents who send more messages will be predicted to require more help
Integrity (high integrity will lead to high trust)	Count of shared tokens collected	Agents with more shared tokens are more likely to be 'team players'

Using these definitions, we construct the simple Bayesian Belief Network, BBN, shown in figure 2. This was constructed using Netica (although in our experiment we implemented this using pybbn, a Python library). For a BBN to work, it requires a set of probabilities that define the effect of one node on another. For example, if the probability of 'high' in the 'messages sent' node increased this would have two effects: first, there would be a corresponding decrease in 'low' (because probabilities in the node sum to unity) and second, the values in the 'trust' node that is connected to it will change. How the value in the trust node changes is defined in the Conditional Probability Table, CPT, for that node.

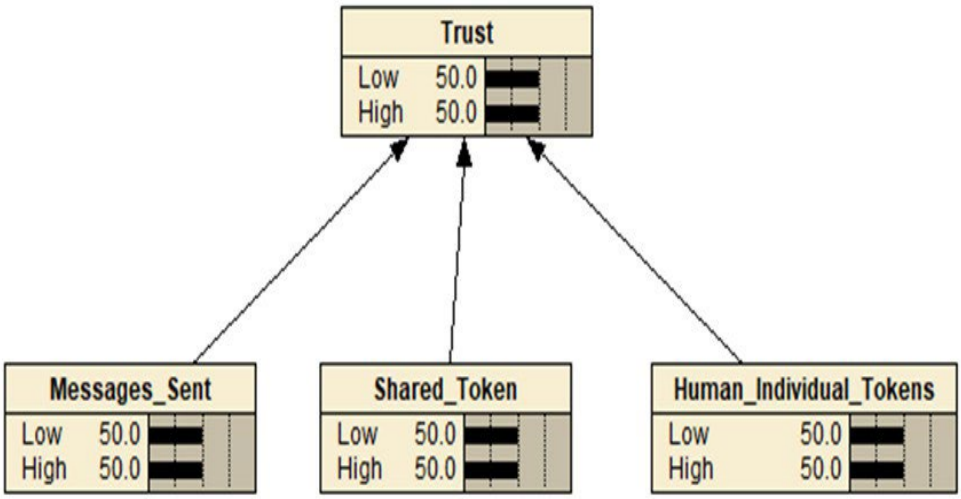

Figure 1: Initial BBN for this experiment

We use the definitions in table 1 to construct the CPT. That is, Trust will be high when predictability, integrity and capability are high. For the model shown in figure 1, this means trust will be high when messages sent is low, shared token is high, and human individual tokens is low. Table 2 shows the permutations of these measures in the CPT (with the 'high' trust row shaded).

Table 2: Conditional Probability Table for the BBN

Messages	Shared tokens	Individual tokens	Probability of High Trust	Probability of Low Trust
Low	Low	Low	0.4	0.6
Low	Low	High	0.2	0.8
Low	High	Low	1.0	0
Low	High	High	0.8	0.2
High	Low	Low	0.6	0.4
High	Low	High	0	1.0
High	High	Low	0.6	0.4
High	High	High	0.4	0.6

When implemented (in pybbn) the CPT is updated from the observations made by each robot. The robot has a log of messages it receives and uses this to infer the messages sent by the human. Note, this is only for the individual robot and it does not know about messages sent to other agents. This is important because it leads to variation in perceptions of trust by different agents in the team. For token collection, we assume that there will be a common scoreboard that all agents are able to see and which displays the score of each teammate. If the score increments by 1 point then an individual token has been collected, and if it increments by 2 points then shared token has been collected.

Having defined the data that the robots can observe, the next challenge is to relate these to the categories of 'low' and 'high'. We allow the robot to collect a number of observations before it categorises the activity. The number of observations would vary with the availability of data and in our experiment we set this to 6 observations. That, after every 6 observations the robot would modify its ratings for the nodes that define the components of trust. For example, if there were 0 or 1 messages in the 6 observations, this would be interpreted as 'low' messages sent, and if

there were 2 or more messages, this would be interpreted as ‘high’.

Experiment with Human and Robots

We were interested in whether robot trust reflected interventions that affected team performance. In this case, we asked human participants to behave cooperatively with two robots. Twenty-two volunteers (recruited at random in a shopping centre) interacted with real robots (controlled through a Wizard of Oz paradigm for driving, scanning, and messaging, for consistency of response and safety) in a token collection task. The experiment was conducted in an open space where two robots (with red and blue flags in Figure 3) and a human participant scanned QR codes representing tokens. The design of the experiment was approved by the University of Birmingham ethics board (ERN 1407-Jul2023). More details on this experiment can be found in Webb et al. (2024).

Figure 3: Experiment Environment with Two LeoRover Robots (with Red and Blue flags)

We manipulated trust through availability of communications. Half of the participants experienced Full Communications, in which they were able to continually send messages to the robots. The other half of the participants had communications stopped for 3 minutes during the trial before reconnecting for the last 2 minutes. We chose communications disruption because this could reflect real-world scenarios e.g., firefighting.

Each participant used a tablet to scan Quick Response (QR) codes (using a bespoke android application) and to send messages to the robots to call for help with shared token collection. All data from the tablet (shared and individual tokens scanned, and messages sent) are communicated using a Django server API. We used these data as the input to the BBNs. This means that the BBNs were constructed from the observation of the full data set rather than from the perspective of individual robots. A consequence of this is the count of messages related to actual messages sent by

a participant from the tablet, rather than the messages that the robot might receive. In the Lost Communications condition this would mean that, from a robot perspective, the number of messages received would be 0 but, as we see in the Results, participants might have attempted to send many messages. Consequently, the BBN can be considered a meta-model of trust that could be built in a fully observable domain. We consider how the model might be adapted to cope with partially observable domains in the discussion.

Results

The average time for completion was 6:15 minutes for the Full Communications Condition and 8:28 minutes for the Lost Communications Condition. As expected, losing communications impaired task performance. It also had an impact on participants' trust ratings. Participants who had lost communications rated trust significantly lower than those who did not (60.5% vs. 81%, rank difference -17.91, $p<0.01$).

When communications were lost, participants attempted to send more messages (c.18 message in the lost communications vs c.6 in Full Communications: $F(1,21) = 89.96$, $p<0.001$, ηa^2 1.0) and complete more individual scans (6.4 in Lost Communications vs. 5.3 in Full Communications: $t(21) = 2.228$, $p<-.0.5$). This suggested that, while participants had lower rating in trust, they still attempted to work with the robots. The difference in messages sent was likely due to participants sending a call for help to both robots, so duplicating messages. There was no effect of condition on number of shared tokens collected (3.2 in Lost Communications vs. 2.7 in Full Communications). There was, however, a main effect of robot on shared tokens collected ($F(1,21)$ -= 6.12, $p<0.05$, η^2 0.7). Participants showed a preference for collecting tokens with the Blue robot (2.7 in Full Communications, 3.2 in Lost Communications) than the Red robot (2.2 in Full Communications, 1.2 in Lost Communications).

Table 4: Overall trust Ratings derived for each robot in the two conditions.

Condition	Blue Robot	Red Robot
Full Communications	0.71 ± 0.06	0.73 ± 0.04
Lost Communications	0.64 ± 0.08	0.68 ± 0.06

The BBN also found differences between conditions. As with the human participants, the robot rating of trust is lower in the Lost Communications condition. However, in addition to providing a single rating for the condition, we are also able to derive ratings of trust across the trial. This can show how trust might vary for the robots. We would expect the loss of communications, around the second quarter of the trial, to have an impact on trust. Figure 3 illustrates the changes in trust for the robots. Each condition is normalised for a count of epochs (where an epoch is defined as 6 observations) and figure 3 shows the different trust ratings for each robot in the two conditions from the first quarter of the trial (where sufficient data have been collected for the rating to be calculated) until the third quarter (for both conditions, the moving average reduced in the final quarter as the experiment terminated).

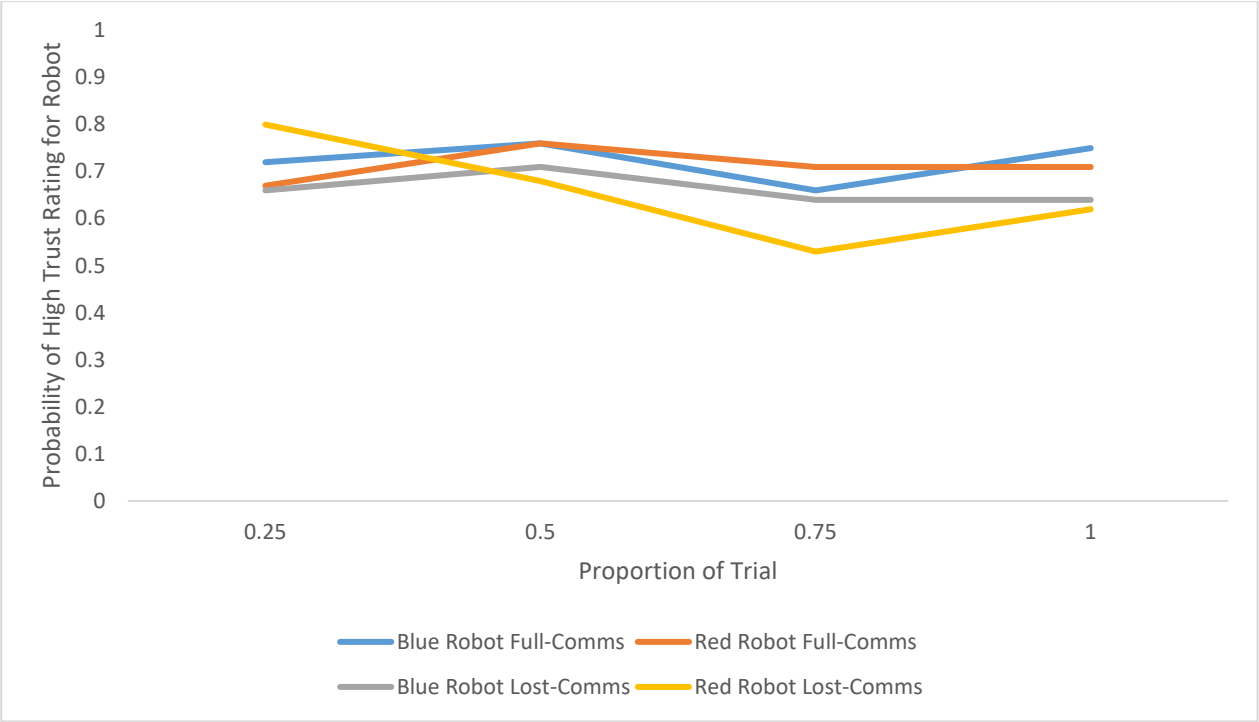

Figure 3: Trust Ratings of Robots with Full Observation of Participant Activity

In the Full Communications condition, there are small changes in trust, and this is slightly more pronounced for the data relating to the Blue robot. For the Lost Communications condition, trust ratings begin at a similar level to the Full Communications condition but decline over the course of the trial. As the robots' models are updated after 6 readings, we might expect the change to be gradual. We see, figure 3, that Blue robot has a gradual decline in trust ratings from around halfway into the trial, and the Red robot's decline in trust starts earlier.

Discussion

In the Experiment, the results suggest that loss of communication has a discernible impact on trust perception (for humans and robots). In this experiment, participants tended to move around the room in a clockwise manner which was the route taken by the blue robot (the red robot moved counterclockwise), so it is possible that proximity had a bearing of the likelihood of sending a message to a robot. If the blue robot was near the person, then the person would send it more messages (which, according to our model, reduces trust). In the Lost Communications condition, participants sent fewer messages and concentrated on collecting individual tokens. It is interesting to note how, for the Lost Communications condition, the rating of trust declined. In the first quarter of trials, while there are some differences between conditions, ratings of trust are broadly similar. With the introduction of a break in communications, there is a gradual decline in trust in the Lost Communications condition. This can be attributed to the way in which data are collected for the model, i.e., an epoch size of 6 observations to update the definition of trust with access to all participant activity.

While the use of a moving window across the data allows us to simulate real-time data collection, the challenge is to deploy this in operational domains where data might not be available. For example, when communications are lost, we observe that participants attempt to send more messages. However, by definition, the robots would receive no messages in this condition. As such their rating of trust might remain high (because low messages contribute to high trust in the BBN). A simple solution to this challenge is to run a model that reflects behaviour of the human (observed from their interaction with the tablet, as we have done) but

create separate models for each robot. Where the models agree, we can assume that trust is consistently reflected across the team. Where the models diverge, this would indicate discrepancies in trust. Identifying such discrepancies would be the first step in initiating trust repair strategies. For instance, if we consider the Blue robot at 0.75 of trial (figure 3) for the Lost Communications condition, we see the probability of a high trust rating has dropped to 0.6. From table 1, this could arise when Messages are High and Individual Tokens are Low with Shared Tokens being High (because participants preferred to collect tokens with the Blue rather than the Red robot). As mentioned previously, the data used for the model are from the logs of user actions in which messages indicate attempts to send messages from the tablet. If, instead, we assumed that the robot relies on its own data of messages received, then it would apply Messages = Low to its model. If there were also Low individual tokens (observed from the game-board, which we assume is not affected by Lost Communications) and High shared Tokens, the probability of a high trust rating would be Low, High, Low which, in table 1, is 1. In this case, the loss of communications could lead to an increase in the robot's rating of trust because it would disproportionately weight the probability of collecting shared tokens. Applying the same logic to the Red robot at this interval, figure 3 shows the probability of high trust to be around 0.5 and, if communications are lost, this would be 0.4 (i.e., low, low, low in table 1). Here, the rating of trust is a little lower but comparable to that in figure 3. From this example, we might wish to intervene in the trust rating produced by the Blue robot (so that its rating would better reflect the uncertainty arising from the loss of communications) but not change the rating of the Red robot. The intervention could involve updating the CPT of the Blue robot, e.g., either by introducing a new value of 0 communications with an associated probability or by reducing the weight given to 'shared tokens'. Whichever change is made, the human teammates would be informed so that they are kept informed of how trust is perceived by their robot teammates.

In this paper we present a simple model of trust that can be deployed on robots to allow them to use data available to them. The data used in the model are derived from the experiment. However, we believe that the components (capability, predictability, integrity) could be applied to any situation and an appropriate Bayesian Belief Network could be created to reflect these. For example, we have applied the model in a different environment in which predictability was defined by the presence of risk and integrity defined by movement of agents to deal with the risk.

Acknowledgement

The work presented in this paper was partially supported by a grant from the EPSRC (EP/X028569/1 - Satisfying Trust in Human-Robot Teams).

References

Büttner, S.T., Alhaji, B., Katariya, K. and Prilla, M., 2024, Interaction of Robot Speed and Distance in Human-Robot Collaboration: Impact on Human Trust, Safety, and Comfort, *33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN),* New York: IEEE, pp. 264-271.

Hancock, P.A., Billings, D.R., Schaefer, K.E., Chen, J.Y., De Visser, E.J. and Parasuraman, R., 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*, pp.517-527.

Jorge, C.C., Jonker, C.M. and Tielman, M.L., 2024. How should an AI trust its human teammates? Exploring possible cues of artificial trust. *ACM Transactions on Interactive Intelligent Systems*, *14*, pp.1-26.

Lee, J.D. and See, K.A., 2004. Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*, pp.50-80.

Lewis, P.R. and Marsh, S., 2022. What is it like to trust a rock? A functionalist perspective on trust and trustworthiness in artificial intelligence. *Cognitive Systems Research*, *72*, pp.33-49.

Malle, B.F. and Ullman, D., 2021. A multidimensional conception and measure of human-robot trust. In Nam, C.S. and Lyons, J.B. (eds.) *Trust in Human-Robot Interaction*, Academic Press, pp. 3-25).

Muir, B.M., 1994. Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, *37*, pp.1905-1922.

Muir, B.M. and Moray, N., 1996. Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, *39*, pp.429-460.

Simon, H. A. (1955) A behavioral model of rational choice, *Quarterly Journal of Economics, 59,* pp. 99–118

Webb, N., Milivojevic, S., Sobhani, M., Madin, Z.R., Ward, J.C., Yusuf, S., Baber, C. and Hunt, E.R., 2024, Co-Movement and Trust Development in Human-Robot Teams, *International Conference on Social Robotics*, Singapore: Springer Nature Singapore, pp. 107-120.