# Effects of Proactive Explanations by Robots on Human-Robot Trust

Lixiao Zhu and Thomas Williams

Colorado School of Mines, Golden CO 80401, USA
lizhu@mines.edu, twilliams@mines.edu

**Abstract.** The performance of human-robot teams depends on human-robot trust, which in turn depends on appropriate robot-to-human transparency. A key way for robots to build trust through transparency is by providing appropriate explanations for their actions. While most previous work on robot explanation generation has focused on robots' ability to provide post-hoc explanations upon request, in this paper we instead examine *proactive* explanations generated *before* actions are taken, and the effect this has on human-robot trust. Our results suggest a positive relationship between proactive explanations and human-robot trust, and reveal fundamental new questions into the effects of proactive explanations on the nature of humans' mental models and the fundamental nature of human-robot trust.

**Keywords:** Human-Robot Interaction · Human-Robot Trust · Transparency · Explanation

## 1 Introduction and Motivation

For human-robot teams to achieve high levels of team performance, appropriate levels of trust must be established between teammates [2]. Human-robot teams tend to have undesirable performance when human-robot trust is either too low or too high, which means that human-robot trust must be maintained at an appropriate level rather than directly maximized [17]. One key factor in establishing an appropriate level of human-robot trust is the transparency of robots' internal beliefs, desires, and intentions [12]. Robot transparency allows users to become aware of the robots' capabilities (thus helping to build capability-based trust) [6, 14], and helps to ensure accurate human mental models of the robot's behavior, which ensures that an appropriate level of trust is established [21].

Robot transparency is typically enabled through verbal communicative behaviors, such as explanation generation [13, 14]. While there has been a significant body of work on explanation generation in human-robot interaction, this work has largely focused only on reactive explanations: post-hoc explanations generated in response to a request from a human teammate to explain a previous behavior[16, 4]. In contrast, little work has explored proactive explanations: explanations that are generated before an action is performed.

One challenge of generating proactive explanations is that because explanations are provided on the initiative of the robot rather than the human, robots must carefully tailor their explanations to avoid communicating too much information, which may overload users [18] and violate communicative norms such as Grice's Maxim of Quanitity [5]. This presents a fundamental tension between design goals of transparency and trust-sensitivity vs. cooperativity and workload-sensitivity. In this work, we thus explore two types of proactive explanatory behavior that can be taken before an unexpected action is performed, each of which places different weight on these competing factors.

- **Proactive Announcement**: Before taking an unexpected action, a robot may perform proactive announcement by stating the action it is going to take. This may serve to enhance predictability-based trust by reducing the user's sense that the robots' actions were unexpected. However, this is not a true explanation as it does not actually reveal the beliefs and desires underlying the robots' intentions [4].
- **Proactive Explanation**: In contrast, the robot may instead perform proactive explanation by stating the action it is going to take *and why*. This functions as a true explanation, revealing the dispositions behind the robot's actions. This allows the robot's teammate to verify that these dispositions are suitable, leading to appropriate levels of deeper, understanding-based trust [3].

We believe that while both proactive announcement and proactive explanation will help to build human-robot trust, proactive explanations will help to build deeper trust by shifting humans' mental models of robots from one in which they can only predict the robot's behavior and assess its reliability, to one in which they can also understand that behavior and assess the suitability of its dispositions. In this paper we present the results of a human subject experiment designed to test this hypothesis.

## 2 Human-Subject Experiment Design

### 2.1 Research Goal

The main research goal of our study is to understand the fundamental relationship between human-robot trust of robots and proactive explanatory behaviors. Specifically, we seek to assess the following two hypotheses.

**Hypothesis 1:** Robots that generate proactive explanatory behavior will be more trusted than robots that do not.

**Hypothesis 2:** Robots that generate proactive explanations will build greater human-robot trust than those that perform proactive announcements.

### 2.2 Experimental Context

To assess these hypotheses, we had participants collaboratively engage with robots in a novel resource management task, in which participants spent different types of resources while exploring an environment, while a robot positioned behind the player was responsible for "collecting" these resources. Through the course of this task, human teammates must spend different types of resources to explore different regions of their environment. The user's thus has an implicit need for resources to be collected that can be expected to be needed in the future, so that they avoid circumstances in which the type of resource required has run out. The type of resource that is actually collected at any given point is determined in two ways. First, the user can manually instruct the robot to collect a particular type of resource. Second, the robot can periodically decide on its own to collect a different type of resource than it is currently collecting, based on what it believes will be most needed, as assessed by the ratio of resources of a particular type revealed to be needed to explore the current exploration frontier, to the amount of resources available of that type:

$$resourceToCollect = argmin_{r \in R} \frac{stored_r}{needed_r} \tag{1}$$

When changing to collect a different resource, either at human direction or of its own volition, the robot rotated to face one of four placards signifying the to-be-collected resource type. The only way for participants to determine which resource was currently being collected by the robot was to physically turn their body to inspect the robot and observe which of these four placards the robot was facing.

### 2.3 Experiment Design

Our experiment used a within-subjects Latin Square design in which each participant engaged in three randomly and procedurally generated resource management tasks, in each of which the robot used a different order-counterbalanced explanatory behavior. Specifically, in each of the three experimental conditions, if the robot decided of its own volition to collect a new resource type, before turning to face the corresponding placard, it used the proactive explanatory behaviors dictated by its experimental condition:

1. **Proactive Announcement (PA)**: In this condition, a robot autonomously switching to a different resource type informed participants of *what* resource it planned to collect, e.g., by saying "I am going to collect blue resources."
2. **Proactive Explanation (PE)**: In this condition, a robot autonomously switching to a different resource type informed participants of both *what* resource it planned to collect, and *why*, e.g. by saying "I am going to collect red resources because you are low on red resources, but it seems that you may need a lot of them."
3. **No explanations (NE)**: In this condition, no proactive explanatory behavior was used.

### 2.4 Experiment Procedure and Measures

Upon arriving at our laboratory, participants provided informed consent, were introduced to the resource management task and the turtlebot robot used in the experiment, and were given time to familiarize themselves with the task. Participants were guided to sit in front of a desktop computer, behind which were located the turtlebot robot and resource extraction points. Fig 1 represents the general setup of the human-subject experiment. Participants then participated in each of the three experiment blocks according to their Latin Square condition.

During each experiment block, participant actions were monitored using cameras mounted in the corners of the experimental space. Camera data was used to calculate an objective measure of human-robot trust, operationalized as the frequency and duration of humans' monitoring of the robots' behavior, with more frequent and/or higher-duration turns to observe the robot taken as evidence of lower trust in the robot (cp.[11, 15]). At the beginning of the experiment and after each experimental block, participants completed the 14-item human-robot trust scale presented by Schaefer et al. [19] to self-report their trust in their robot teammate. Gain scores between baseline trust scores and post-condition trust scores were then used as a subjective measure of trust to supplement our observational measure.

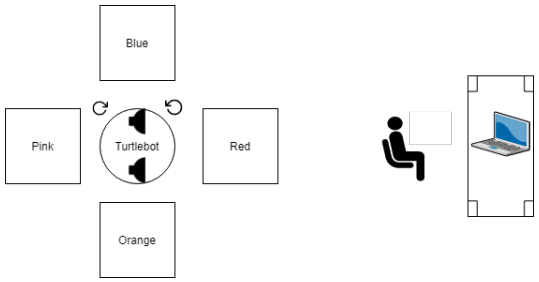



Fig. 1: Turtlebot Robot Setup for the Human-subject Experiment

### 2.5 Participants

32 participants (primarily university students) were recruited for the experiments from the Colorado School of Mines campus. While we initially intended to collect data from a greater number of participants, recruitment was cut short due to the COVID-19 global pandemic. The data from 21 participants were retained after removing data from participants who performed actions that required experimenter intervention (e.g., accidentally closing the testbed window). This human-subject experiment was approved by the Human Subjects Research (HSR) board at the Colorado School of Mines.

# 3 Results

## 3.1 Analysis

Our results were analyzed under a Bayesian analysis framework using JASP [9]. Results were analyzed using Bayesian analyses of variance (with experimental condition as the independent variable and subjective and objective trust measures as dependent variables), followed by Bayes Factor analyses and pairwise post-hoc Bayesian t-tests.

Bayes factors can roughly be interpreted as ratios of evidence in favor of alternative hypotheses relative to competing (e.g., null) hypotheses [8]. Bayes factors between 0.33 and 3 are generally taken as anecdotal evidence [10] insufficient to confirm or refute a hypothesis. Bayes factors between 0.33-0.10 or 3-10 provide substantial evidence against or for the hypothesis; Bayes factors between 0.03-0.10 and 10-30 provide strong evidence; and Bayes factor less than 0.01 or greater than 100 provide decisive evidence.

## 3.2 Subjective Measures

A Bayesian ANOVA provided moderate evidence against any effect of proactive explanatory behavior on self-reported human-robot trust (Bf 0.143). This Bayes Factor indicates that the collected data were approximately seven times less likely to have been generated under a model accounting for proactive explanatory behavior than under one that does not.

## 3.3 Objective Measures

Four video recordings were removed due to camera system failure, yielding 17 remaining video recordings, from which we analyzed the frequency and duration of human teammates' physical turns to monitor their robot teammate.

Table 1: Post Hoc Comparisons of Monitoring Duration

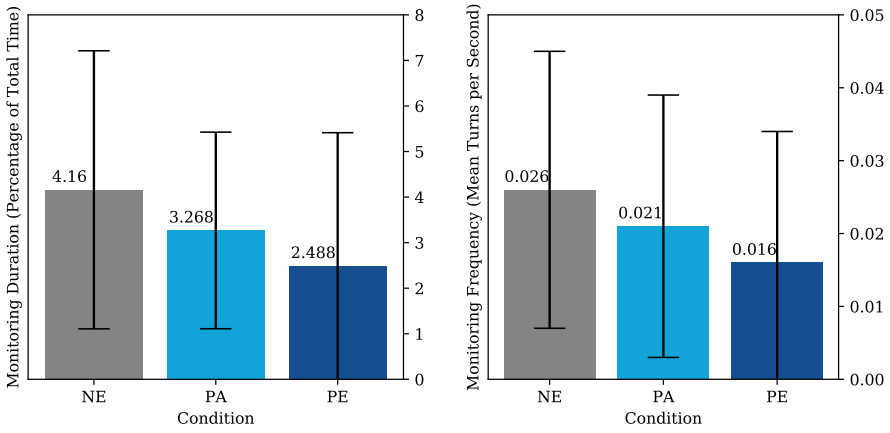
| | | Prior Odds | Posterior Odds | $BF_{10,U}$ | error % |
|---|---|---|---|---|---|
| NE | PA | 0.587 | 0.258 | 0.440 | 6.640e-4 |
| NE | PE | 0.587 | 13.766 | 23.436 | 5.425e-5 |
| PA | PE | 0.587 | 0.215 | 0.366 | 0.003 |

Bayesian ANOVAs provided indecisive evidence, neither supporting nor refuting an effect of proactive explanation condition on duration (Bf 1.090) and frequency (Bf 1.313) of human-robot monitoring. To interrogate these inconclusive effects, we performed post-hoc pairwise comparisons between experimental conditions for both monitoring duration and frequency. Tables 1 and 2 presents the results of these pairwise post-hoc analyses. As visualized in Figures 2a and 2b,

Table 2: Post Hoc Comparisons of Monitoring Frequency

| | | Prior Odds | Posterior Odds | $BF_{10,U}$ | error % |
|---|---|---|---|---|---|
| NE | PA | 0.587 | 0.343 | 0.583 | 0.002 |
| NE | PE | 0.587 | 3.130 | 5.328 | 1.591e-4 |
| PA | PE | 0.587 | 0.222 | 0.377 | 0.002 |

our results suggest that human-robot monitoring was less frequent (Bf 5.328) and of lower duration (Bf 23.436) when proactive explanations were used than when no explanatory behavior was used, but provided indecisive evidence insufficient to either confirm or refute any difference in duration or frequency of human-robot monitoring between proactive announcements and either of the two other behaviors. Overall, these results generally support our subjective findings, but suggest that more observational data would be necessary to fully confirm them.




(a) Duration (b) Frequency

Fig. 2: Effects of Proactive Explanatory Behavior on Duration and Frequency of Human-Robot Monitoring

## 4 Discussion

We hypothesized that proactive explanatory behavior would increase human-robot trust (H1), especially when proactive explanations rather than proactive announcements are used (H2). While these hypotheses were not supported by our subjective measures, they *were* partially supported by our objective measures.

Specifically, our results suggest that robots that generated proactive explanations were more trusted than those that generated no proactive explanatory behaviors, but that more data must be collected before concluding precisely what effect is had by proactive announcements.

### 4.1 Participant Observations

In order to understand the discrepancy between our subjective and objective results, we begin with several interesting observations based on informal observations made by participants. Anecdotally, several times during our study, participants volunteered during the post-experiment debriefing that the task sessions with proactive explanatory behaviors made them feel that the robot was more of a teammate than a mere tool. This is a positive result given the research of Billings et al. [1], which suggests that people must perceive robots as teammates rather than tools to have effective interaction. This strongly suggests the need for future experimentation to assess whether this is a generalizable benefit of explanatory behaviors

In addition, several participants volunteered during post-experiment debriefing that they in actually preferred the robot's proactive announcements over proactive explanations, and that the proactive explanations felt wordy and unnecessary. We see two possible explanations for these observations. First, it could be the case that there is no difference between proactive announcements and proactive explanations in terms of understandability or desirability of robot behavior, similar to the observations made by Stange et al. in their examination of robot reactive explanations [20]. This explanation, however, would seem to contradict what is known about explanations in general and the benefits they provide in terms of transparency and facilitation of accurate mental modeling. Accordingly, a second, and we argue more likely explanation, is that while the robot's *first* proactive explanations may have facilitated accurate teammate modeling of the robot's belief- and desire-related dispositions, its subsequent explanations did not contribute any additional dispositional knowledge (thus leading to a violation of Grice's Maxim of Quantity [5], and human displeasure with the robot). This may explain our subjective results, and would suggest that for robots to jointly optimize human-robot trust, maximize robot likability, and minimize human workload, robots must themselves maintain sufficient models of their human teammates' beliefs about their own beliefs (second-order theory of mind [7]) in order to know *when* to generate proactive explanatory behaviors (and what kind to generate).

### 4.2 Impact of Explanatory Behaviors on "Theory of Mind"-oriented Mental Models

Similarly, second-order theory of mind effects (in this case, triggering of human teammates' own first- and second-order "Theory-of-Mind'-oriented mental models with respect to the robots) may also help to explain the discrepancy

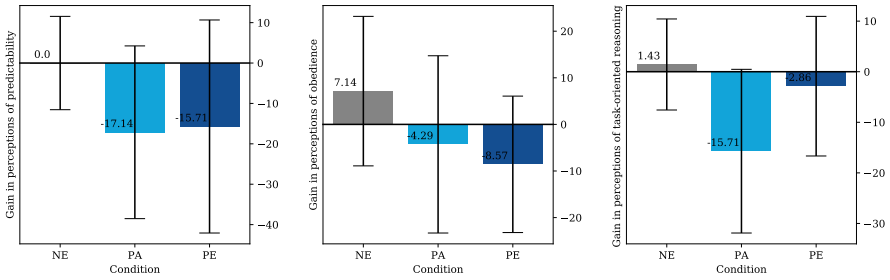
between our subjective and objective results. Specifically, by generating proactive announcements, robots implicitly demonstrated an ability to reason and communicate that had been heretofore unobserved. This may well have led to increased perceptions of agency, capability, and intelligence; first-order theory of mind modeling with significant potential for impact on human-robot trust, especially capability-based trust. Similarly, generating proactive explanations that made reference to the human-robot team and their shared task may have well have led to increased perceptions of sociality and solidarity, i.e., willingness to help to fulfill the participants' needs; second-order theory-of-mind modeling with additional potential for impact on human-robot trust, especially reliability-based trust. These observations yield a number of testable hypotheses that must be explored in future work.

### 4.3 Impact of Explanatory Behaviors on the Nature of Trust

If the robot's explanatory behaviors do indeed trigger these Theory-of-Mind-oriented changes in participants' mental models of their robot teammate, this would have dramatic effects on participants' beliefs about the dispositions of the robot and the suitability thereof; in short, it would change precisely what it would even mean for participants to trust the robot.

In particular, we now consider what may have transpired specifically for participants who first encountered the robot in the condition in which it generated no explanatory behaviors, and the dramatic change in their mental model of the robot that would have transpired when, in their second task, the robot began generating proactive explanations. Before and after this shift, there was no change in the robot's actual behavior in terms of frequency of deviation from users' commands. Before the shift, we would presume that participants would interpret robots' deviations from commands as a low-level error in the robot's programming; the robot would perhaps have been perceived as being unreliable or incapable due to *unintentional* resource targeting "drift", rendering it untrustworthy in terms of successfully fulfilling the participant's commands.

In contrast, once the participant entered the second experimental condition and the robot began to generate announcements, it would become immediately obvious that the robot's deviations were in fact intentional acts of disobedience: the robot may in fact have been fully capable of achieving the user's goal, but unreliable in terms of its motivations and drive to comply; a gain in one dimension of trust coupled with a loss in another. Finally, once the participant entered the third experimental condition and the robot began to explain *why* it was deviating from participant commands, the robot would be perceived as being occasionally disobedient in order to better achieve the team's goals: a potential source for increased trust in terms of the robot's high level motivations, while still allowing for comparable sources of distrust (a) in the robot's willingness to accede to the participants' requests, and/or (b) in the quality of the robot's capability to successfully pursue a course of action that would actually lead to greater benefit to the team's goals than strict obedience would have.




(a) Predictability (b) Obedience (c) Task-Driven

Fig. 3: Effects of Proactive Explanatory Behavior on Duration and Frequency of Human-Robot Monitoring

In order to assess the plausibility of this narrative, we re-analyzed the data from only the seven participants who saw experimental conditions in the order {NE, PA, PE}, on a question-by-question basis. While the strength of our results were quite weak due to the small sample size, even with this small sample a number of interesting results emerged.

1. First, we examined the perceived predictability of the robot (Survey item 4). Even though the robot behaved identically in all three conditions in terms of its actual decisions, a Bayesian ANOVA demonstrated that perceived predictability may have significantly dropped from the No Explanation condition to the Proactive Announcement condition (Bf 1.275, Fig. 3a), indicating that once the robot began speaking, the notion of what it even meant for the robot to be "predictable" may have changed substantially.
2. Second, we examined the perception that the robot performed "exactly as instructed" (Survey item 13). A Bayesian ANOVA demonstrated that this perception may have dropped from condition to condition, resulting in a difference from first to last condition in terms of perceived disobedience (Bf 1.341, Fig. 3b).
3. Finally, we examined perceptions that the robot acted to meet the needs of the task (Survey item 10). A Bayesian ANOVA demonstrated that beliefs to this effect may have dropped substantially once the robot began explaining its actions (Bf 2.443), but that this drop in trust may have recovered once the robot began explaining the team-driven reasoning behind its actions (Bf 0.984, Fig. 3c).

The small sample size on this analysis means that these results are ultimately inconclusive, but suggest that additional data could provide evidence for a substantial change in mental models between conditions, in which participants initially view the robot as a faulty tool, then as needlessly disobedient, and finally as disobedient for the purposes of the task.

## 5 Conclusion

In this work, we conducted a human-subject study to better understand the relationship between human-robot trust and robots' proactive explanations. Our results suggested that proactive explanations lead to increased human-robot trust as assessed through objective observational means. Our results were inconclusive, however, with respect to proactive announcements and the precise effects of this form of proactive explanatory behavior on human-robot trust. As discussed above, our results raise a number of interesting further questions pertaining to the effects that proactive explanatory behaviors might have on users' mental models of robots, and the impact this might have on the fundamental nature of human-robot trust. In future work, additional investigation is needed (1) to collect sufficient data to confirm or refute the inconclusive findings presented in this paper, (2) to identify the optimal policies for navigating the tradeoff between trust and workload that is presented during explanation generation; and (3) to interrogate the new research questions that have been identified regarding theory of mind and the fundamental nature of human-robot trust.

## Acknowledgments

This work was supported by an Early Career Faculty grant from NASA's Space Technology Research Grants Program.

## References

1. Billings, D.R., Schaefer, K.E., Chen, J.Y.C., Hancock, P.A.: Human-robot interaction: Developing trust in robots. In: Proc. Int'l Conf. on HRI (2012)
2. Billings, D., Schaefer, K., Llorens, N., Hancock, P.: What is trust? defining the construct across domains. In: Proc. Conf. of the American Psych. Assoc. (2012)
3. Danks, D.: The value of trustworthy AI. In: Proc. AIES (2019)
4. De Graaf, M.M., Malle, B.F.: How people explain action (and autonomous intelligent systems should too). In: 2017 AAAI Fall Symposium Series (2017)
5. Grice, H.P.: Logic and conversation. In: Syntax and Semantics 3: Speech acts (1975)
6. Helldin, T.: Transparency for Future Semi-Automated Systems : Effects of transparency on operator performance, workload and trust. Ph.D. thesis, University of Skövde (2014)
7. Hiatt, L.M., Trafton, J.G.: Understanding second-order theory of mind. In: Comp. ACM/IEEE International Conference on Human-Robot Interaction (2015)
8. Jarosz, A.F., Wiley, J.: What are the odds? a practical guide to computing and reporting bayes factors. The Journal of Problem Solving **7**(1), 2 (2014)
9. JASP Team: JASP (version 0.12.2)[computer software] (2020)
10. Jeffreys, H.: The theory of probability. OUP Oxford (1998)
11. Lee, J.D., See, K.A.: Trust in automation: Designing for appropriate reliance. Human Factors **46**(1) (2004)
12. Lyons, J.B.: Being transparent about transparency: A model for human-robot interaction. In: 2013 AAAI Spring Symposium Series (2013)

13. McManus, T., Holtzman, Y., Lazarus, H., Anderberg, J., Ucok, O.: Transparency, communication and mindfulness. Journal of Management Development (2006)
14. Mercado, J.E., Rupp, M.A., Chen, J.Y.C., Barnes, M.J., Barber, D., Procci, K.: Intelligent agent transparency in human–agent teaming for multi-uxv management. Human Factors **58**(3), 401–415 (2016)
15. Muir, B.M., Moray, N.: Trust in automation. part ii. experimental studies of trust and human intervention in a process control simulation. Ergonomics **39**(3) (1996)
16. Neerincx, M.A., van der Waa, J., Kaptein, F., van Diggelen, J.: Using perceptual and cognitive explanations for enhanced human-agent team performance. In: Int'l Conf. on Engineering Psychology and Cognitive Ergonomics (2018)
17. Ososky, S., Schuster, D., Phillips, E., Jentsch, F.G.: Building appropriate trust in human-robot teams. In: 2013 AAAI Spring Symposium Series (2013)
18. Rieser, V., Lemon, O.: Natural language generation as planning under uncertainty for spoken dialogue systems. In: Emp. Meth. in natural language generation (2009)
19. Schaefer, K.: The perception and measurement of human-robot trust. Ph.D. thesis, University of Central Florida (2013)
20. Stange, S., Kopp, S.: Effects of a social robot's self-explanations on how humans understand and evaluate its behavior. In: Proc. Int'l Conf. on HRI (2020)
21. Wang, N., Pynadath, D.V., Hill, S.G.: Trust calibration within a human-robot team: Comparing automatically generated explanations. In: Proc. HRI (2016)