Interactive Human-Robot Teaching Recovers and Builds Trust, Even With Imperfect Learners

Vivienne Bihe Chi  
vivienne_chi@brown.edu  
Brown University  
Providence, RI, USA  

Bertram F. Malle  
bfmalle@brown.edu  
Brown University  
Providence, RI, USA  

ABSTRACT
Building and maintaining trust is critically important for continued human-robot teaching and the prospect of robots learning social skills from natural environments. Whereas previous work often explored strategies to reduce system errors, mitigate trust loss, or enhance learning by interactive teaching, few studies have investigated the possible benefits of fully engaged, interactive teaching on human trust. Motivated by a pair of discrepant previous investigations, the present studies for the first time directly tested the causal impact of interactivity on the loss and recovery of trust in a human–robot social skills training context. Building on a previously developed experimental paradigm, we randomly assigned participants to one of two modes of interaction: interactive teacher vs. supervisor of an experimentally controlled virtual robot. The robot was engaged in learning norm-appropriate behavior in a healthcare setting and improved from mistake-prone to nearly-flawless performance. Participants indicated their changing trust during the 15-trial training session and how much they attributed the robot’s improvement to their own training contributions. Interactive teachers were more resilient to initial trust loss, showed increased trust in the robot’s performance on additional tasks, and attributed more of the robot’s improvement to themselves than did supervisors, even when the robots were slow learners.

CCS CONCEPTS
• Applied computing → Psychology; • Human-centered computing → Empirical studies in HCI

KEYWORDS
Interactivity, Trust, Human-Robot Interaction, Teaching

ACM Reference Format:

1 INTRODUCTION
Human teaching is critically important for the evolution of artificial autonomous agents in social environments, allowing agents to develop knowledge, acquire skills, and productively engage with their human partners [5, 56]. Social skills, in particular, such as the ability to learn and abide by social norms [4, 41], could be learned through continued interaction in natural environments [10, 42]. Norm teaching never ends, however, if people want robots to apply their skills to new tasks, operate successfully in novel contexts, and interact appropriately with different people. In such expanding deployment, humans must be willing to continuously teach the robot new norms and refine preexisting norms. Moreover, robots must combine autonomy and interactive capabilities to benefit from this teaching and co-evolve with their human user [45].

Humans are natural pedagogues [61] and are experienced in not only learning norms but teaching norms to one another [57]. However, when humans teach a robot, they may not have the same interest, patience, or enjoyment as they have when teaching a child or new team member. Moreover, humans are also users of machines, who expect the robot to already know certain things and perform tasks for them. Robot designers therefore need to carefully balance the demand for human teaching input and the payoff that humans receive from the robot, which motivates them to keep engaging with the agent. Designers must also ensure that a social robot credibly improves, and that human teachers have trust in the robot’s ability to learn, especially in the complex domain of social norms.

Building trust is necessary for successful learning interactions, so people continue to teach the robot and rely on it to actually perform intended tasks in daily life [19, 67]. Human-machine trust is influenced by many factors when deployment environments become complex and social [16, 30, 49]. When people interact with flawed but steadily evolving machines, they are bound to feel uncertain [40] and lose trust [3, 20, 52]. However, if humans take on a more interactive teaching role to an evolving agent, their trust in it might strengthen even when the agent is imperfect—because they are involved in the machine’s improvements [1, 2].

Previous work showed that various forms of enhanced interactivity can benefit a machine’s learning success [33, 53, 65], but few studies have directly investigated the possible benefits of fully engaged, interactive teaching on human trust. Without maintenance of trust, however, teachers may give up on the robot too soon, and the prospect of learning success remains out of reach.

In a pair of previous investigations [16], researchers examined how human trust changed during robot training and suggested a significant difference in the trust trajectories of more engaged vs. less engaged human teachers. In one study, participants were directly teaching and actively intervening (by instructing or evaluating the robot) and were able to overcome initial trust loss and expand trust over the course of the training. In the other study, participants followed the robot learner passively through its learning
journey but struggled to ever recover from their initial trust loss. Both kinds of teachers were exposed to the same robot improvement trajectories, and both were sensitive to the same measured and experimentally manipulated variables. So the failure to recover from the early trust loss may be attributed to the fact that participants were actively teaching the robot in one study but only supervising its learning process in the other study. However, the studies differed in multiple features and the critical feature of interactivity was not systematically manipulated.

To examine the role of interactivity in a more rigorous way, the present studies experimentally manipulated the mode of interaction that participants were able to engage in, and we measured the teachers’ changing trust in nearly continuous assessments. We also assessed teachers’ intention to use this robot in other future tasks, beyond the ones they trained it for. We closely matched the experimental paradigm originally developed by [14–16] that places participants in a robot training setting and allows the researchers to assess trust at multiple levels. We manipulated interactivity by inviting human teachers to actively engage in acts of teaching or to supervise the robot at each learning step. All other features in the two conditions were held constant: Teachers were exposed to exactly the same robot performance trajectories (witnessing numerous mistakes at first but reaching perfect performance at the end), experienced the same communicative interface with the robot, and expressed their trust in the robot in exactly the same way and at the same points of the learning process. This design enabled us to test most directly the causal impact of interactivity on the loss and recovery of trust in a human-robot training context, centering on the critical social skills of norm-appropriate behavior.

2 RELATED WORK

2.1 Measuring Human-Robot Trust

Trust is an important element in successful human-machine interactions [30, 40, 50]. However, trust has typically been treated as a response to machine reliability and competence [12, 52, 55, 58], whereas the domain of social interaction and the learning of social norms introduces a richer set of feelings and judgments of trust [25, 38, 43, 69]. Previous work has explored a variety of possible antecedents to trust that range from robot- and human-related to environment-related [30]. These factors shift in their relative contributions from early stages of interaction, when trust is determined more by features of the human tutor, to later stages, when trust is determined increasingly by the trustee’s actions [35].

One important determinant of human-robot trust is of course the robot’s task performance [39]. Accordingly, our studies experimentally manipulate trial-by-trial performance variations to assess the resulting trial-by-trial trust variations. Moreover, trust accumulates over multiple pieces of evidence [36] and people must “update” their trust. Such repeated trust updating is rarely measured in HRI studies, where instead a single trust violation is followed by a repair attempt [64]. In real-world contexts, by contrast, performance fluctuations and trust responses covary over time and across multiple tasks [12, 22]. The human-robot teaching setting presents a complex case of such evolving trust. Therefore, our studies capture these variations by measuring local trust 15 times for 15 different tasks, but all in a common context of a healthcare setting.

2.2 Who is Responsible

A teacher’s trust varies according to the logic of causal attribution [11], in which the trustor attributes the trustee’s performance to both the trustee’s dispositions and the task’s difficulty [31]. But there is a third factor: Teachers also attribute the learner’s success or failure to themselves, as teachers. In human-human studies, researchers have found a consistent tendency for individuals to attribute success to themselves [44], whereas attributions of failure show considerable variability [6, 29, 54, 68]. Similarly, in human-robot interactions, humans are more consistently found to credit robots for successes compared to assigning blame in case of a task failure [21, 32, 47]. Whether teachers consider themselves responsible for the success or failure of their learners is reflected in the construct of teaching efficacy [46], and some would argue it is merely an “illusion of control” [51]. Whether illusory or not, an actively engaged teacher is more likely to experience and believe that they have helped a learner improve. In our studies, we therefore assessed people’s self-perceived contribution to their robot learner’s improvement. This measure was intended to both verify that even a supervising (low-engagement) teacher experiences some impact on the learner’s success and also to demonstrate that an interactive teacher has a greater sense of contribution.

2.3 Interactivity and User Engagement

In human-machine interaction, trust loss is often due to system error [52, 70], and researchers have explored ways to mitigate such trust loss: improving learning algorithms [12, 13], deploying trust repair strategies [18, 23, 26], and treating robots as ingroup members to make people more error-tolerant [60].

Trust loss can also arise when the user has high expectations that the robot cannot meet, which often occurs in teaching settings. Researchers have examined the mitigating power of rich interactive teaching experiences [9, 27, 28, 62], which can increase a human’s positive emotions and willingness to engage with a robot [28, 62]. These benefits could foster further engagement and user trust, although interaction could also decrease trust if the robot resists the teacher’s efforts and continues to fail.

Research has examined ways to optimize robot learners’ interactivity for better user experience. Some explored the positive effect of a robot verbally asking proactive questions [8, 17], others used modalities such as eye-gaze as a means to communicate understanding or the robot’s need for guidance [48, 63]. A few studies explored whether the teacher’s interactive engagement with the learner affects trust. [24] found that people experienced more trust after giving active feedback to (rather than merely observing) a robot arm’s self-learning process. However,[59] did not find evidence that participants trust a robot more if they taught it themselves, compared to a robot they believed someone else had taught. Few studies have examined interactive teaching outside the physical performance (e.g., kinematics) domain, so we presented our teachers with the goal of teaching a robot social norms in a healthcare setting—something that is both personally relevant to many people and vital to society’s well-being. Thus, even the supervising interaction mode should engage participants, but not in the same specifically interactive way that the teacher role does.
3 RESEARCH OBJECTIVES

Previous work suggested that interactive teaching, compared to supervision, may contribute to a teacher’s recovery from early trust loss (caused by the robot’s flawed initial performance) [16]. However, the evidence was indirect—relying on different samples and methodologies. To rigorously test the impact of interactivity, the present two studies experimentally manipulated the mode of interaction that participant-teachers are able to engage in: interactive teaching vs. supervision as a control condition. We measured the trajectory of fast-changing task-specific trust feelings during the training session and also assessed changes of more general trust in the robot’s future performance on additional tasks. Moreover, we added a new variable to elucidate the potentially distinct trajectories of trust development in the two modes of interaction. After the training session, participants had an opportunity to indicate the causal contribution they thought they made to the robot’s improvement. We reasoned that this self-attribution would be sensitive to the experimental manipulation of interactivity and through the mediation of the perception of robot improvement.

Thus, based on theory and empirical evidence discussed above, as well as our initial findings [15], we formulate the following preregistered hypotheses:

Hypothesis 1: Interactivity and trust trajectory. Interactive teachers, compared to supervisors, of a learning robot (a) exhibit less net trust loss from the robot’s flawed beginning of the training session to the improved end, and (b) recover more of their lost trust by the end of the training.

Hypothesis 2: Interactivity and teacher’s self-attribution. Interactive teachers, compared to supervisors, of a learning robot perceive themselves as making a greater causal contribution to the robot’s displayed performance improvement.

Hypothesis 3: Interactivity and teacher’s future trust. Interactive teachers, compared to supervisors, of a learning robot trust the robot on future tasks more after the training session than before.

4 OVERVIEW: METHODOLOGY AND STUDIES

In a set of two online studies, participants used a chat app to engage with a virtual, fully experimenter-controlled robot in a teaching setting. They were asked to help the robot learn how to act norm-appropriately in several healthcare tasks, and the robot’s learning path was experimentally controlled to improve at a certain learning rate. Participants were randomly assigned to either an interactive teacher role or a supervisor role. In both roles, people expressed their feelings of trust in the robot before each of the 15 tasks. This trust feeling measure aimed to capture local, trial-by-trial variations. In addition, we measured people’s future trust in the robot’s performance in other task contexts, both at the beginning and at the end of the training session.

Study 1 investigated the impact of teaching interactivity on people’s early loss of trust (due to the robot’s flawed performance) and subsequent recovery, as well as their change in future trust, and their self-perceived causal contribution to the robot’s improvement. In Study 2, we experimentally manipulated not only the teaching interactivity but also the robot’s learning rate (slow vs. fast performance improvement), as trust loss and recovery may be critically moderated by the learner’s objective improvement.

5 STUDY 1

5.1 Methods

5.1.1 Participants. We aimed at recruiting 100 participants (for power ≥ .80 at d ≥ .40, α < .05) in each of two conditions. To prepare for invalid data, we enrolled 221 U.S. participants on Prolific and compensated them $3.00 for the 15-minute study, which was approved by Brown University’s Institutional Review Board. No participant had invalid data, so 221 cases were available for analysis.

5.1.2 Procedure. We largely followed the procedures of [15]. Participants adopted the role of a nurse responsible for training a healthcare robot assistant (see Supplementary Material [SM] Sec. 1 for verbatim text). Unlike [15], before the training session, we established participants’ baseline trust in this minimally described robot to perform four future tasks (“future trust”). Then the training session began, consisting of five blocks (“training days”), each featuring three tasks that a robot nurse assistant would realistically perform—for example, politely knocking before entering a room or handling a patient complaint. For each task, the robot initiated the interaction by describing its current environment (e.g., elevator), its goal (e.g., to go downstairs), and action-relevant context (e.g., other people are waiting ahead). (See SM Sec. 6 for all task scenarios.)

Next, the critical condition of interaction mode was randomly assigned. Interactive teachers had a choice, for each of the 15 trials, to either (a) instruct the robot or (b) ask for and evaluate its action proposal. If the participant chose to instruct, the instructed action was transmitted and the robot replied with “Noted!” along with its own original plan of action (“I was planning to...”). If the teacher chose to evaluate, the robot offered its planned action for the current task and the participant evaluated the proposal on a rating scale. By contrast, supervisors learned about the robot’s action plan straightaway. Thus, in both interaction mode conditions, participants always knew about the quality of the action that the robot was prepared to take, irrespective of the participant’s assigned role of interactive teacher or supervisor.

After the robot described its task and communicated its planned action, participants expressed their momentary trust feelings. In addition, after each block of three tasks, participants completed a multi-item measure of how trustworthy they judged the robot to be (see SM Sec. 4). Because of space constraints, we do not report the results of this measure, but they are consistent with the results in [15].

At the end of the entire training session, participants again expressed their future trust, rated how much they thought the robot had improved over the course of their training session, and indicated how much they felt their training session played a part in the robot’s improvement (see 5.1.3).

5.1.3 Materials and Measures. We used a smartphone-based interface (designed by [14, 15] and programmed in jsPsych) in which participants have the experience of engaging with a virtual robot agent in a training session. In this interface, participants communicate with the robot via short messages, make their teaching strategy
choices, give instructions or feedback (in the interactive teaching condition), and complete all dependent measures. Specifically, between trials, participants indicated their feelings of trust in the robot, answering the question, “At this moment, how much do you feel you are trusting this robot to act appropriately?,” on a slider scale from 0 (“Not at all”) to 100 (“ Completely”). To go beyond trust for actions in the training session we assessed people’s future trust. Participants indicated whether they trusted the robot assistant to complete each of four tasks in the future (see SM Sec. 5), rated on a slider scale from 0 ("Absolutely No") to 100 ("Absolutely Yes"). The four tasks were ordered from closest to farthest from the healthcare learning domain: (1) a familiar in-domain (healthcare) task (distributing medication, which also occurred in the training); (2) a novel in-domain task (being a companion to an older hospital patient); (3) a task in the out-domain context of education (handling out props, which resembled a familiar in-domain task); and (4) a novel out-domain task (helping children with homework). We compared these four tasks with a set of reverse Helmert contrasts, comparing the in-domain novel against the in-domain familiar task (2 vs. 1), the familiar out-domain task against the two in-domain tasks (3 vs. 1/2), and the out-domain novel task against the other three (4 vs. 1/2/3).

Two new measures of perceived improvement ("How much do you think the robot has improved over the course of your training session?"") and self-attribution ("How much do you feel like your training session with the robot played a part in its improvement?”) were both measured on 0-100 rating scales.

5.1.4 Experimental Manipulations. Instructions for the interactive teacher vs. supervisor conditions were highly similar except for the explanation to the interactive teacher of how to choose between instructing and evaluating (see SM Sec. 1).

Evidence of learning progress (steady improvement after flawed initial performance) was experimentally controlled by a fixed scheme (see SM Sec. 2) of the robot’s action proposals that, over the course of the five blocks, increased in quality. Each action proposal had a level of appropriateness, established in a pretest in which participants (N = 70) rated the quality of all action proposals on a 1-5 scale of appropriateness. The resulting mean ratings were treated as a continuous predictor in statistical analyses.

The difficulty of the tasks the robot faced was also experimentally controlled, and a given block of three tasks was always ordered from easy to difficult. The specific levels of difficulty were established in another pretest, in which participants (N = 30) rated all tasks on a 0 ("easy") to 100 ("difficult") scale. The resulting mean ratings were treated as a continuous predictor in statistical analyses.

5.2 Results

5.2.1 Trial-by-trial Feelings of Trust. To examine people’s continuously updated trust we analyzed the series of 15 trial-by-trial ratings of trust feelings that participants expressed immediately after learning about the quality of the robot’s proposal for the current task. We predicted the trajectory of these trust feelings from a mixed-effects model with participant as a random effect and these fixed effects: interactivity (interactive teaching vs. supervising), difficulty level of the current task, robot’s current performance (inappropriate, acceptable or appropriate proposal quality on the current task) and previous performance (proposal quality on the previous task), as well as the cumulative performance score (sum of performance quality across all previous tasks), and their interactions.

The analysis revealed that people’s feelings of trust in the robot increased considerably in response to the robot’s current performance (t = 22.38, p < .001). In addition, interactive teachers felt overall higher trust than did supervisors (t = 2.03, p = .04), especially when the robot’s proposal was not appropriate (t = 7.26, p < .001). Further, people tracked the robot’s improving performance over the course of the training session: The higher the robot’s cumulative performance score up to and including the previous task, the more they trusted the robot on the current task (t = 3.5, p < .001). And when a robot had been accumulating particularly high-quality performance, a better action proposal for the current task elicited even stronger feelings of trust (t = 10.0, p < .001).
increase in trust from Day 4 to 5. This latter pattern disconfirms Hypothesis 1-b, according to which interactive teachers should recover more of their lost trust by the end of the training session. In reality, it was supervisors who recovered more of their lost trust, because they had greater trust loss to begin with; at the end, supervisors had even slightly higher trust than interactive teachers.

5.2.2 Future Trust. We measured future trust before and after the training session to detect changes in trusting the robot to perform tasks that go beyond the training. We analyzed change scores in a mixed-effects model with a participant random effect and fixed effects of interaction mode, future task (reverse Helmert contrasts as described earlier), and their interactions.

Overall, people in both interaction mode conditions increased their future trust in the robot from pre-training ($M = 60.44$) to post-training ($M = 67.55$), $t = −5.52, p < .001$. People were also sensitive to task type (see Fig. 2). In particular, they displayed smaller trust increases for the out-domain novel task ($M = 2.72$) than for the other three tasks ($M = 8.59$), $t = −2.3, p = .02$. The only difference between interactive teachers and supervisors was in their responses to the first three tasks ($t = −2.21, p = .03$); interactive teachers indicated greater trust for the two in-domain tasks than for the familiar out-domain task whereas supervisors showed moderate trust throughout.

5.2.3 Perceived Improvement and Self-Attribution. One might worry whether supervisors experienced any teaching at all during the training session, given that they only observed but did not directly guide the robot’s behavior like the interactive teachers did. Yet, participants assigned to be supervisors still attributed a considerable amount of the robot improvement to their own training1 ($M = 57.73$, $SD = 29.61$ on the 0-100 scale), although less than the interactive teachers ($M = 66.38$, $SD = 25.12$), $t(205) = −2.33, p = .02$. These differential experiences did not, however, alter teachers’ perceptions of the objective performance improvement that (by experimental control) the robot displayed in both conditions (also on a 0-100 scale): for supervisors, $M = 73.3$, $SD = 22.85$; for interactive teachers, $M = 74.55$, $SD = 22.00$), $t(215) = −0.41, p = .68$.

In sum, everybody saw the robot improve, but interactive teachers attributed more of this improvement to themselves.

5.3 Discussion

Hypothesis 1 proposed that interactive teaching would positively affect people’s trust trajectory in a robot training session. Hypothesis 1-a had presupposed an overall loss of trust, and in that sense was disconfirmed. However, we found that interactive teaching mitigated an early trust loss, compared to the steep loss for supervisors. The additional measures of future trust and self-attribution of robot improvement indicate a possible drawback of interactive teaching, however. Because interactive teachers attributed more of the robot’s improvement to their own training, they might have believed that teaching was necessary for the robot to improve and that the robot would not be able to generalize across domains without more mentorship. As a result, interactive teachers had confidence in the robot’s ability to master additional in-domain tasks but less so out-domain tasks. By contrast, robot supervisors had somewhat less confidence in the robot for in-domain tasks but no less so for a familiar out-domain task (see Fig. 2).

Because the robot was programmed to show a fast learning rate over the course of the training session, both interactive teachers and supervisors were able to recover their trust and further grow it beyond their starting trust. Such quick performance improvement is difficult to achieve in most current real-world robotic systems. Working with realistic systems, users who are supervising a flawed robot might quickly disengage after their early trust loss and miss the opportunity for later trust recovery. To explore this possible disengagement, we added in Study 2 a measure of people’s desire to switch to and train a different robot (assessed right before the fast improvement of robot behavior occurs).

Hypothesis 1-b was disconfirmed, as interactive teachers did not show more trust recovery than supervisors; on the contrary, supervisors had a larger trust loss to recover from and still reached comparable—by the end even slightly higher—trust levels as interactive teachers did. We conjecture that the fast-learning robot greatly impressed participants in both conditions, and perhaps impressed supervisors even more because they observed the robot improve so quickly on its own. To examine this possibility, Study 2 included a slower-learning robot (as designed by [16]) that may be closer to reality. Because it does not quite achieve perfection by the end of the training session, it may not foster as high levels of trust as the faster-learning robot did in Study 1. As a result, supervisors may be less confident that the robot will achieve flawless performance, whereas interactive teachers may build such confidence because they themselves mentored the robot onto a promising path toward perfection.

6 STUDY 2

Whereas theory and previous findings allowed us to formulate specific hypotheses for Study 1, Study 2 tries to answer questions raised by Study 1 and aims to identify boundary conditions for interactive teaching vs. supervision of a slow-learning robot.

---

1In fact, only one participant out of the 105 assigned to the supervisor role expressed concern (in the open-ended feedback at the end of the study) that their interaction was too limited to have helped the robot to improve.
Specifically, for the case of training a slow-learning rather than fast-learning robot, we posed the following five research questions:

1. Will early trust loss be mitigated for the interactive teacher?
2. Will supervisors still recover from their steeper early trust loss and exceed initial trust?
3. Will supervisors (having less evidence for the robot learning “on its own”) no longer increase their future trust from before to after training?
4. Will interactive teaching (more so than supervision) continue to increase people’s perceived own contribution to a slow-learning robot’s improvement?
5. Will interactive teachers tolerate a slow-learning robot’s errors and not opt out of training whereas supervisors will?

6.1 Procedure and Design

Study 2 followed the procedures of Study 1, with two modifications:

First, we added a between-subjects factor of slow vs. fast robot learning rate, following the parameters of [16]. The slow-learning robot’s slope of improvement over the course of the training days was less steep than the fast-learning robot’s (which appeared in Study 1). (For details of this improvement scheme, see SM Sec. 2.)

Second, at the end of the third training day, we asked participants: “You will continue to work with this robot assistant for the next few training days. However, if you were given the choice, how much would you prefer to work with a different robot?” We posed the question as a hypothetical to avoid loss of experimental control, because if an unknown number of participants would actually self-select into working with a new robot, they would have different expectations, consequently interpret robot performance differently, and Study 2 results could not be compared to Study 1 results. The hypothetical can still capture an authentic sentiment, just as in many real-life circumstances when we wish we could do something different even though we know we cannot.

6.2 Participants

In a post-hoc analysis of Study 1, the size of the observed effect of interactivity on the early trust loss was $d = .51$, which suggested an intended sample size of $n = 61$ per group in Study 2’s 4-group design (statistical power $\geq .80, a < .05$). In the end, 258 U.S. participants were recruited on Prolific and compensated at $3.00 for the 15-minute study, which was approved by Brown University’s Institutional Review Board. Study 1 participants were ineligible for Study 2 recruitment. 11 participants were excluded following the preregistered criteria, which left just over 61 per cell in the design.

6.3 Results

6.3.1 Trial-by-trial Feelings of Trust. To investigate which factors made an impact on a participant’s trust feeling update trajectory, we again used a mixed-effects model to predict people’s trust feelings ratings in each trial, this time with robot learning rate and its interactions with the other factors as the additional fixed effects.

Replicating the results of Study 1, participants’ trust feelings were responsive to the robot’s task-by-task performance ($t = 19.79, p < .001$) as well as its cumulative performance ($t = 4.99, p < .001$). More important, interactive teachers expressed greater feelings of trust than did supervisors, regardless of the robot learning rate ($t = 3.99, p < .001$), and especially when current performance (the robot’s action proposal) was of lower quality ($t = -8.13, p < .001$) or when the robot had not yet displayed sufficient cumulative performance ($t = -2.7, p > .01$).

We next analyzed the more stabilized trust trajectories across the five training days. Fig. 3 shows the key patterns of trust loss and recovery for the two interaction modes and the two robot learning rates. Collapsed across interaction modes, the slow-learning robot elicited very different trust trajectories than the fast-learning robot, with significant contributions from all polynomial contrasts ($t > 5.0, ps < .001$), but the most salient ones were the slow learner’s flatter linear increase and deeper trough-like quadratic pattern. Collapsed across learning rates, interactive teachers experienced a less drastic trust loss than did supervisors, measured as the Day2-to-Day1 contrast ($t = 4.45, p < .001$).

The key analyses to answer our research questions require analyses within fast- and slow-learning robots separately. We see that, for the fast-learning robot, trajectories for interactive teacher and supervisor were very similar to those in Study 1 (see Fig. 1), differing in linear, cubic, and quartic components (all $ps < .001$). For the slow-learning robot, the trajectories for interactive teacher and supervisor were differentiated by a linear component (large early trust differences that decreased at the end, $t = -4.34, p < .001$) and a quadratic component (far deeper and longer-lasting trust loss for supervisors, $t = -4.61, p < .001$).

So the answer to question (1)—Will early trust loss be mitigated for interactive teachers even of a slow-learning robot?—is a clear Yes. The answer to question (2)—Will supervisors still recover from their steep initial trust loss and exceed early trust?—is also Yes, but their trust recovery with a slow-learning robot lagged behind the recovery for interactive teachers (see right panel of Fig. 3).

6.3.2 Future Trust. A mixed-effects model examined the change of trust in the robot’s future performance on additional tasks. Trust increased after the training session more for the fast-learning than the slow-learning robot ($t = 4.05, p < .001$). In addition, while both interaction modes showed considerable increases in trust in the fast-learning robot, interactive teachers increased their trust in slow-learning robots more than supervisors ($t = 2.10, p < .05$).
fact, answering research question (3) in the negative, supervisors of slow-learning robots no longer showed trust increases after training (see Fig. 4, right panel).

When we compare future trust across tasks, we see that participants increased trust in the fast-learning robot less for the out-domain novel task than for the three remaining tasks ($t = -2.45, p = .001$), which is consistent with Study 1.

![Figure 4: Future trust patterns in Study 2. Compared to supervising a slow-learning robot, interactively teaching it led people to increase their intention to rely on it more.](image)

6.3.3 Perceived Improvement and Self-Attribution. Overall, participants accurately differentiated between the greater improvement of the fast-learning robot ($M = 75.31, SD = 19.15$) than the slow-learning robot ($M = 59.38, SD = 24.98$), $t = 5.62, p < .001$. Directly replicating Study 1, when engaging with a fast-learning robot, interactive teachers in Study 2 perceived the same amount of improvement ($M = 76.46, SD = 17.09$) as supervisors did ($M = 74.28, SD = 20.92$), $t = 0.63, p = .53$. Also as in Study 1, interactive teaching led to stronger self-attribution ($t = 2.09, p = .039$), even when controlling for perceived objective improvement ($t = 11.06, p < .001$).

When teaching a slow-learning robot, however, this picture changed. Interactive teachers perceived more improvement in their slow learners ($M = 64.89, SD = 22.11$) than did supervisors ($M = 54.78, SD = 26.44$), $t = 2.28, p = .025$), even though the robot performed the same way. Moreover, in a multiple regression predicting self-attributed contributions to the robot’s improvement, controlling for the strong variable of perceived improvement ($t = 11.73, p < .001$), being an interactive teacher no longer independently predicted people’s self-attribution ($t = 0.72, p = .47$).

We can demonstrate these different patterns of fast and slow robot learners in a pair of mediation analyses (see Fig. 5), where perceived improvement acts as a mediator to the effect of interactive teaching (vs. supervision) on people’s self-attributed contributions. With fast learners, interactive teaching has no impact on perceived improvement but directly increases self-attribution. With slow learners, interactive teaching increases perceived improvement (as if teachers see more potential in the learner), but any impact of interactive teaching on people’s self-attributed contributions is mediated by these increases in perceived improvement.

We can now answer question (4)—Will interactive teachers (more so than supervisors) more readily attribute a slow-learning robot’s improvement to their own contributions? Yes, but solely because interactive teachers have more generous perceptions of the slow learner’s improvement, and perceptions of greater improvement lead to self-attributions of having contributed to that improvement.

6.3.4 Continued engagement. Question (5) was whether supervisors would opt out of training a slow-learning robot, whereas interactive teachers could tolerate the robot’s early errors because they feel they contribute to the robot’s improvement. When working with the fast-learning robot (like the one in Study 1), supervisors were as reluctant to opt out ($M = 47.08, SD = 28.06$) as interactive teachers ($M = 43.17, SD = 28.06$), $t = - .78, p = .44$. But when working with the slow-learning robot, supervisors were considerably more inclined to give up on the robot ($M = 59.49, SD = 33.4$) than were interactive teachers ($M = 47.0, SD = 31.78$), $t = -2.11, p = .04$.

6.4 Study 2 Discussion

Study 2 made two contributions. First, it replicated all the core results of Study 1. Interactive teachers of fast-learning robots were more resilient to early trust loss, but supervisors caught up by the end. Interactive teachers also showed increased future trust from baseline to post-training, but supervisors showed no less increase, presumably because they witnessed a robot that improved on its own. Interactive teachers attributed more of the robot’s improvement to themselves than did supervisors.

Second, Study 2 offered new insights into the power of interactive teaching by exposing teachers to a slow-learning robot and testing the effects on trust trajectories, future trust, and self-attributions. Interactive teachers of slow learners showed the same mitigated trust loss and steady trust recovery that they showed with fast learners. Supervisors showed the same substantial trust loss that they showed with fast learners, but their recovery lagged behind. Perhaps as a result, supervisors of slow-learning robots showed stagnant confidence in the robot to succeed in future tasks, and they were more inclined to give up on the slow learner. Finally, interactive teachers saw more improvement in the slow learner, which may indicate greater patience with the robot or tolerance of its early errors.

7 GENERAL DISCUSSION

We investigated whether interactivity in human-robot teaching can foster trust, especially when errors and imperfections lead to temporary loss of trust. Our overall conclusion is that interactivity
has powerful benefits for trust management, but even mere supervision of a learning robot builds trust, as long as the robot visibly improves.

Interactivity protects against steep initial trust loss, both for “talented” and less talented robot learners. Why might this be? In each trial, the teacher made a choice to either instruct the robot or evaluate its action proposal. It has long been known that choices increase engagement and investment in outcomes [37]. The knowledge that one will be able to instruct, guide, and evaluate a robot provides motivation to invest in the teaching and confidence that, ultimately, the learner will improve.

Moreover, in the multiple exchanges, interactive teachers received contingent, receptive responses from the robot (“I had planned to...”), which have been shown to foster communication and perceived agency [34, 66]. The interactive teaching was more akin to human-human interaction (reminiscent of parent-child interaction [7]), than human-machine interaction, and it more readily evoked feelings of trust. By contrast, from a supervisor’s perspective, the robot learner simply reported action plans without any invitation for further interaction, which may have led people to a distant, critical perspective and less investments in its improvement.

Supervisors, without the opportunity to directly intervene, are more affected by a learner’s early mistakes. However, as the robot improves (even the slow-learning one), even supervisors are keenly aware of progress and recover their lost trust and gain even more. The recovery process takes longer for supervisors than interactive teachers; and it takes longer when working with slow-rather than fast-learning robots. But one of the important insights is that participants—in all interaction modes and for all robots we examined—are tolerant of initial errors, patient in awaiting improvements, and accurate in recognizing the improvements when they do arrive.

In addition to investigating the dynamics of trust in human–robot teaching, we also examined a new variable: the teacher’s judgment of their own causal contribution to a learner’s success. We found that all trainers, even the more passive supervisors, experienced making considerable contributions, though the interactive teachers more so. The notion of teaching efficacy is well-studied in the education literature [46], but its value for the design of future robots may have been underestimated. If ordinary community members are to become social partners with robots in the near future, they must be ready to teach, correct, and help the robots around them. However, if humans do not believe they can teach robots or if they do not see evidence that their teaching makes a difference, they will quickly rescind their teaching role. Therefore, we must better understand how people’s natural teaching abilities can be sparked and how their teaching efficacy can be nourished, so that they make contributions to robots’ improvements through patient, persistent teaching.

8 LIMITATIONS

Our investigation did not introduce interactive teaching as a novel concept; it is well-known in the education literature but has rarely been studied in the HRI literature. Building on recent work in which we fortuitously found evidence for the effects of interactive teaching on human–robot trust [15], we set out to experimentally manipulate this variable, replicate our findings, and gather new insights into the boundary conditions of different forms of teaching (e.g., slow learners, trust in future tasks) and the role of self-attributed contributions to a robot’s improvement.

Our studies used a Wizard-of-Oz paradigm in which participants interacted with a virtual robot; we do not know how generalizable the findings are to live human–robot teaching. To increase realism we created a rich narrative context, described a physical robot’s actions (e.g., take the elevator), and mimicked remote robot control. These attempts seemed to have a positive impact, since people showed a high level of engagement, even as supervisors.

The communication channel was solely text-based. Such channels are second nature to people in modern life, and they are used in real human–robot interactions (e.g., human-to-Mars-rover communication), in addition to countless chatbots; but they limit information about the robot and the precise environment it is in. Various enrichments might increase realism but would introduce confounding variables: voice-based communication would elicit perceptions of robot gender or personality; robot images would introduce the powerful impact of humanlike appearance; scene images would introduce features of patients and the hospital staff. To retain experimental control of the central variables (interactivity, robot learning capacity), compromises were necessary.

9 CONCLUSIONS

Trust is needed both to engage humans as robot teachers and to encourage users to rely on robots for daily tasks.

We hope to have shown that interactive teaching has powerful benefits for trust recovery and growth, even trust for future tasks. But people’s natural pedagogy goes further, as even supervisors experience engagement and are tolerant of imperfect robots as long as the robots improve. In our studies, people had high expectations and lost trust when the robot made several mistakes early on. Trust recovery was aided by either direct teaching or a fast-learning robot, or by time—as most participants exceeded their initial trust after 15 trials of training. People were more hesitant to generalize their newly gained trust to other, future tasks. Here it takes again either interactive teaching or a fast learner to build such future trust, but for truly novel tasks, people did not generalize—which may demonstrate an important fence against overtrust.

If robot designers want to maximize the chances that human teachers become significantly invested in a learning robot’s success and maintain trust and patience even when the robot makes mistakes, then the best option is to create engaged teaching interactions and give people autonomy to make teaching choices. If robot learning algorithms are built to take advantage of people’s natural teaching ability, then robots can become truly better and will be met with trust and patience to get better yet.

ACKNOWLEDGMENTS

This project was supported in part by a grant from the Air Force Office of Scientific Research, Award No. FA9550-21-1-0359

REFERENCES