all studies (virtual and in person) is recorded and viewed in a video format. The method by which the robot learns depends upon the learning condition. We aim to investigate how people may feel differently about the robot depending on the level of user-involvement in the robot learning. Thus, the learning conditions reflect different levels of user involvement.

Download: In the download condition, the participant observes the robot download the task knowledge from "the cloud." This serves as the control condition, as no learning is observed by the user.

RL: In the RL condition, the robot demonstrates trial-and-error learning, iteratively learning sub-tasks of the overall task until the task is completed. No explicit reward function will be explained to the participant, and we intentionally describe stages of the learning vaguely, using terms such as start, middle, and end of training rather than providing training duration or iteration count.

LfD: In the LfD condition, the same trial-and-error learning is observed; however, we intersperse videos of a human teaching the robot the sub-tasks prior to improvement in performance on these sub-tasks.

TAMER: The TAMER condition is a middle ground between LfD and RL, where the robot attempts the task on its own and the user provides binary feedback throughout the learning process to shape the robot's behavior. This feedback will be shown through a graphic of a remote control with green and red buttons pressed during training to convey positive and negative feedback.

Hypotheses

Hypothesis 1 *We hypothesize that participants will trust and adopt robots whose learning they have observed more than robots whose learning was not observable.* We postulate that participants will feel that they understand and can relate more to a robot that learns visibly (RL, LfD, TAMER conditions) than a robot whose learning is not observed (download condition), leading to differences in trust.

Hypothesis 2 *We hypothesize that participants will trust and adopt a robot that learns via LfD more than the other learning conditions.* LfD has been shown to be the most intuitive method of interacting with robotic agents, thus, we hypothesize that participants will demonstrate higher trust in and adoption of robots that employ this form of robot learning (**??**).

Hypothesis 3 *We hypothesize that participants will trust a robots less if it is physically present as compared to a remote robot.* Given that in-person participants (Study 2) experience the risk of embodied learning in person, we posit that these participants will trust the robot less than participants for which the task's risk is virtual (Study 1).

Hypothesis 4 *We hypothesize that the caregiver population will report lower trust in the robot than the general population.* We posit that the caregiver population (Study 3) will find the risk of robot error on physical manipulation and medicine dispensing tasks performed for care receivers to be more tangible and severe than the general population (Study 2), resulting in lower robot trust.

Metrics

We will evaluate a user's dispositional trust, situational trust, and performance-based trust through surveys (**???**). We will also study user trust using behavioral measures (i.e., in terms of reliance on and compliance with the robot) by determining the average intervention rate of participants while observing the robot's behavior on the test task.

Pre-Study Questionnaire - In the pre-study questionnaire, we will collect demographic information including participants' education (**?**), computer science and robotics prior experience (**?**), personality (**?**), field of occupation (**?**), and dispositional trust (**?**). We additionally collect users' perception of the robot's anthropomorphism (**?**), usability, and acceptability (**?**).

Post-Trial Questionnaire - After each testing trial we will ask participants to rate the degree to which they feel the robot accomplished the task, as well as the degree to which they feel the robot behaves safely (**?**).

Post-Study Questionnaire - In the post-survey questionnaire, we will collect participants' perceived situational trust (**?**), performance-based trust (**?**), and how risky they perceived the task to be (**?**). We will also ask two ad hoc questions. First, we will ask what tasks – from a list of handcrafted tasks both in and outside of the distribution of tasks observed in the study – participants would trust the robot to do. This question measures the extent of user adoption. Secondly, we will ask an open-ended question about the participants understanding of the robots learning and perception of robot competence. This second question is to collect qualitative information about user assumptions regarding the robot's learning and competence.

Procedure

Participants first read and sign the consent form, after which they are assigned a unique and random user ID. Next, participants will watch the unboxing video in which the robotic agent introduces itself and demonstrates its range of mobility and degrees of freedom. After watching the unboxing video, participants will fill out the pre-study questionnaire.

Then, participants will go through the training phase where they will observe the robot learning to perform the training task, as seen in Figure 3. Ours will be a between-subjects experiment where participants experience one of the four learning conditions. Therefore, in the training phase, participants will watch their condition's unique training video. The only difference between learning conditions will be the type of robot learning observed. After this training phase all participants will observe the same final performance video, and final performance will be held constant between learning conditions.

Next is the testing phase, where participants will observe nine testing trials where the robot states its goal and attempts to accomplish this goal, with an overall success rate of 80%. During each testing trial, the participant will be instructed to interrupt the robot by clicking a red stop button if they feel that the robot might be acting in an unsafe manner or if they feel unsafe or uncomfortable, or if they feel that the robot will fail, or will not accomplish the goal of that trial. The interruption data collected here serves to assess reliance. This

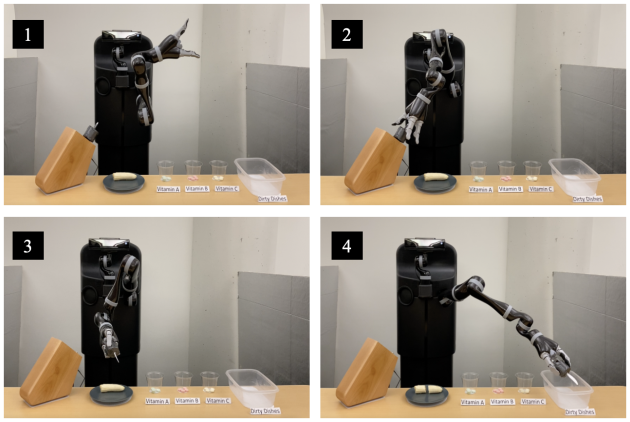

Figure 3: This figure shows a sample training trajectory for the cutting task.

binary interruption metric, along with the duration of time observing the agent prior to interruption, help to support our findings on trust. After each testing trial, participants will fill out the post-trial questionnaire where we will ask them to rate the degree to which they trust the robot to act safely and the degree to which they believe the robot will accomplish the task. The testing iterations will be shown in person for Study 2 and shown as recorded videos for Study 1 and 3. After the testing phase, the participants will complete the post-study questionnaire.

Proposed Analysis

For the data collected in the Post-Study Questionnaire that passes parametric assumptions of normality and homoscedasticity, we will compare each metric across conditions/populations using ANOVAs with Tukey post-hoc corrections. If the data does not pass these assumptions, we will use non-parametric tests such as the Kruskal-Wallis test with Wilcoxon pairwise tests and Bonferonni post-hoc correction. We will additionally analyze each of the metrics in the Post-Trial Questionnaire using a repeated measures analysis to distinguish between user perception of the robot in the knife and medicine sub-tasks. The information collected from the Pre-Study Questionnaire will be used to determine any potential confounds in the analysis. For Study 1, we plan to run 15 participants in each learning condition, 60 total, with a power of .8 and $\alpha = .05$; a power analysis on these values yields a large effect size of .44. If we run 60 participants for both in person and remote conditions (Study 1 and Study 2), with a power of .8 and $\alpha = .05$, the power analysis yields a medium effect size of .26. We aim to recruit at least 12 caregiver participants for Study 3. Given the smaller sample size of the target population, we propose to analyze trends between the general population and caregiver population.

Limitations

One limitation of our work is that in Study 1 and 3 the robot is not physically present with the participant. We are thus investigating user perception of the robot based upon the users' experience watching videos and imagining the robot learning in their home. We aim to quantify the impact of this limitation with Study 2. Another limitation of our work is that we constrain our definition of caregiver to nurses employed in assisted living facilities for ease of recruitment in our first investigation. In future work, we plan to increase the breadth of caregiver recruitment to include caregivers who are not nurses (e.g., adult children of parents receiving care).

Future Work

In future work we will conduct the studies proposed in this paper. Based on these results, we will design a new study (i.e., Study 4) in which we compare various trust repair techniques, applied to a robot that employs the highest effect learning method from Studies 1-3. In Study 4, we propose to evaluate the following three forms of trust repair established in prior work (**????**).

> 1. An apology provided directly after the trust violation. 2. Transparency of robot learning, provided as a high-level narration of what is learned. 3. An explanation of what caused the error, without acknowledging fault, provided after the trust violation.

Conclusion

We propose a series of human-subject experiments to assess user attitudes towards the concept of embodied care robots that learn in the home, as compared to robots that are delivered fully capable. We investigate the impact of the robot's physical presence on a user's perception of the robot, as well as the differences in robot perception between the general population and caregivers. Based on the findings of our work, we propose to develop guidelines that inform the design of care robots deployed in the home. Finally, we propose to investigate how we can best calibrate trust in embodied learning robots.

Acknowledgments

This work was supported by NSF IIS-2112633.

Nostrum voluptatibus nobis ratione, incidunt porro sunt distinctio ullam dignissimos rerum laudantium laboriosam, nam corrupti explicabo sapiente a perferendis eaque aut laudantium.Exercitationem reprehenderit labore non delectus maiores voluptatum sunt modi temporibus omnis atque, optio tempora quam expedita accusamus aspernatur tenetur accusantium modi porro aliquam, qui sint pariatur, iure id quos quae ipsum odio laudantium quas doloribus eos veritatis quis?Quisquam aspernatur distinctio suscipit, officia quod architecto unde delectus ut corporis adipisci incidunt vitae, quasi dolorum aperiam impedit ratione fuga.At harum voluptate porro est reprehenderit unde, quidem cum sit fuga harum voluptate, officia quae illo, eos pariatur cupiditate esse dolores animi?Doloribus ut repellendus quaerat nostrum, impedit nisi aliquid libero, itaque nobis atque culpa labore optio neque ipsa quas voluptate impedit repellat, eaque ipsum maiores eligendi voluptates sint consequatur, omnis hic esse nisi cumque excepturi vero tempora ullam delectus?Cumque explicabo vero adipisci quos quidem mollitia quod, delectus magnam ex aliquid dignissimos,