# Robot Social Identity Performance Facilitates Contextually-Driven Trust Calibration and Accurate Human Assessments of Robot Capabilities

Maria P. Stull,[1], Clare Lohrmann[1], and Bradley Hayes[1]

***Abstract*— People struggle to form accurate expectations of robots because we typically associate behavior (and capability) with the physical entity even when there are clear indicators of different software programs dictating behavior at different times. This is a harmful prior, as commercially available, visibly similar robots do not necessarily share any common ground in terms of capability, safety, or behavior. Prior efforts to calibrate people's expectations of robots have not extended to anchoring on the robot's control software rather than its embodiment. In this work, we leverage social participation and flexible identity presentation to facilitate coworkers' associations of robot capability with the currently running software rather than physical entity itself. By linking each of a robot's controllers to a social identity, we enable collaborators to more easily differentiate between them. In a human subjects study ($n = 30$), participants who experienced our social identity signal understood differences between the robot's two controllers and prevented an unreliable controller from harming perceptions of the robot's other controller.**

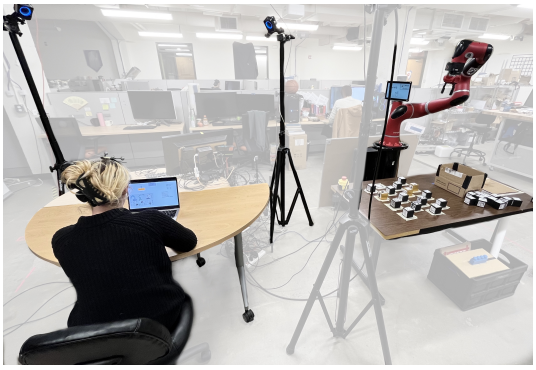

Fig. 1: A participant supervises the robot in an order-fulfillment task. Using a social identity signal to distinguish between multiple robot controllers enables participants to separately calibrate trust for each controller.

## I. Introduction

People often have inaccurate expectations of robots—whether through having a poor understanding of the robot's goals, being unable to predict its actions, or having an inappropriate level of trust [1], [2], [3]. This problem is compounded by the fact that robots are programmable and can exhibit multiple kinds of behaviors, so having a good understanding of a robot at a particular time may not translate to a different situation. User-facing computer programs, on the other hand, are packaged and branded as applications. Because of this abstraction, users form separate expectations of the capabilities of each application. This allows users to select the appropriate software to perform a particular task, and transfer their knowledge about a program from one device to another.

Crucially, due to robots' embodiment and implicit social participation, users do not not perceive them as mere computers [4], erasing assumptions of independence between software applications. Despite being capable of running multiple programs, robots do not typically communicate what logic is running at a particular time, contributing to this perception. Designers, programmers, or frequent operators—users with comprehensive understanding of the robot's behavior and capabilities—may not require explicit communication about this. Someone with less expertise, however, is unlikely to have the same depth of understanding.

* This work was supported by the Army Research Laboratory under grant number W911NF-21-2-0290.

[1]Department of Computer Science, University of Colorado, Boulder firstname.lastname@colorado.edu

For example, someone who regularly works in a warehouse with a robotic arm will become familiar with its range of motion, speed, and the tasks it usually performs, allowing them to collaborate and share space efficiently. What happens, then, when the robot begins using updated control software that changes its behavior? Even if the coworker is told about the update, their familiarity with the robot may become a liability, as they are overly confident in their ability to safely share space with it but unable to predict its new behavior. If, however, the coworker can separate their familiarity with the robot from their familiarity with its software, updates or other software changes can occur without posing the safety risks of overconfidence.

Robots have considerable flexibility in how they present their identity [5], and may use attributes such as names, speech, behavior, and their physical form to do so [6]. Robots are also fluid in their identity performance [7], meaning they can alter the identity they are performing as the situation requires it. In this work we take advantage of this flexible identity presentation and the fact that robots are social actors to communicate changes in running control software to a human co-worker. We investigate expressions of identity as an independent experimental variable for signaling a robot's capability and reliability at a suite of manipulation tasks. Participants in an in-person, between-subjects ($n = 30$) human subjects study work with an autonomous robotic arm alternating between two control programs. The robot displays either an algorithm name (unique for each controller) or one of two identity signals. We show that *people who experience a socially-engaged identity signal can better differentiate*

*between controllers and appropriately calibrate trust in each.*

## II. Related Work

Work in robotics has demonstrated that users often have an inappropriate level of trust in robots or automated systems [1], [8], [2], [3]. Both over- and under- trust in autonomous systems can cause problems ranging from inefficient allocation of resources to safety-critical failures [9]. Since mistargeting trust levels in either direction is problematic, the goal is *trust appropriateness*: a user's trust in the system matching the system's capabilities [10]. *Trust calibration* is the process of altering a user's trust level in order to reach an appropriate trust level [11]. Efforts in trust calibration focus on *trust repair*: attempts to increase a user's trust in the system, typically after a failure [12], [13], and *trust dampening*: lowering a user's expectations when they may be too high [14], [15].

Several strategies for targeted trust calibration have shown promise. Trust-dampening messages from a virtual agent, provided before low-reliability periods, increased a user's trust appropriateness [15]. A study on adaptive trust calibration [16] found that when users showed signs of over-trust in a simulated quadcopter, alerting them before periods of unreliability was more likely to change their behavior than continuously reporting the likelihood of success. In another study of virtual agents, participants were able to differentiate trustworthiness between an agent that gave mostly correct and mostly incorrect answers to general knowledge answers [17]. These studies show that there are successful strategies to encourage people to differentiate trust between the same agent at different times, or between multiple virtual agents; our work expands this line of inquiry to differentiating between different agents within the same physical embodiment.

Robot control code is difficult to interpret and predict, even for expert users. Increasing the transparency of a robot collaborator—by, for example communicating system limitations, reliability, or task information—makes human-robot teams more effective [18]. We use identity cues to increase robot transparency by giving more information about the robot's current state and programming. While previous work has developed methods to give expert users insight into control code at a granular level [19], [20], our work provides a method for increasing transparency by giving non-expert users an "at-a-glance" understanding of the robot's behavior.

We use the flexible social capabilities afforded by robots to associate the robot's social identities with specific elements of its behavior, such as different policies [21] or controllers for task execution. Luria et al. [22] observed that robots are able to express social presence in ways humans cannot; a social agent is not bound to a single form, and a robotic form can host multiple social agents. This flexibility of social presence can be useful: for example, a single presence across multiple embodiments can create a unified customer service experience [22], or a social presence that suffers a hardware failure can move to a new body while maintaining its relationship with its human collaborator [23]. These techniques must be deployed carefully; if used improperly, non-humanlike social presences can reduce trust, cause privacy concerns, and raise questions about robot ability [24]. Flexible social presence strategies are highly context-dependent; small changes in how an agent or robot is presented may strongly impact how a user views them.

Luria et al. [22] described a type of flexible social presence similar to the one we investigate as *co-embodiment*: a single form hosting multiple identities at once. People were uncomfortable with the co-embodiment scenario presented: an autonomous vehicle, where the "driving" agent had a tense discussion with a different agent. In this situation, the safety concerns of a distracted driver may be more salient than the general acceptability of co-embodiment. In our work, one robot hosts multiple agents, but the agents are not active at the same time and do not interact with each other. We expect that this method of presenting agents and the less safety-critical situation will not cause the same level of discomfort.

Williams et al. [25] proposed *Deconstructed Trustee Theory*, positing that trust in a robot can be separated into trust in its physical form and in the agent or persona the robot displays. They compared trust-building and -damaging statements from robots with a consistent identity/embodiment pairing, and robots where an agent "migrated" to another body. They found evidence that participants had different levels of trust in the body and the identity.

Our work investigates elements of this theory further, though rather than differentiating between the robot body and identity, we ask whether participants can differentiate between two different social agents presented as "controlling" a single robot form. We investigate two forms of identity cue to determine the minimum degree of social performance that is salient to participants. We conduct our evaluation in-person, because a social robot's physical embodiment has a significant impact on perceptions of it [3] [26]. If a user can employ social identity signals to explain the robot's behavior, they can adjust their level of trust in the system according to the robot's capabilities at a particular time.

## III. Methods

### A. Hypotheses

Our key insight is that identity signals should enable users to differentiate between robot behaviors. To test this, participants worked with a robot that used two controllers with noticeably different levels of reliability. The robot used one of three methods to signal which of its controllers was active: a social identity signal, a "weak" identity signal, and a baseline signal of an algorithm name. We expected that the identity signals would improve participants' awareness of which controller was active at a particular time, and differences in the behavior of each controller. Informed by literature on robot identity and trust, we hypothesize that:

$H_1$ : Participants in the identity-based groups will differentiate between the controllers on measures of trust and competence more than those in the algorithm-name group.

$H_2$ : Participants in the identity-based groups will more accurately predict the capability of the robot than the

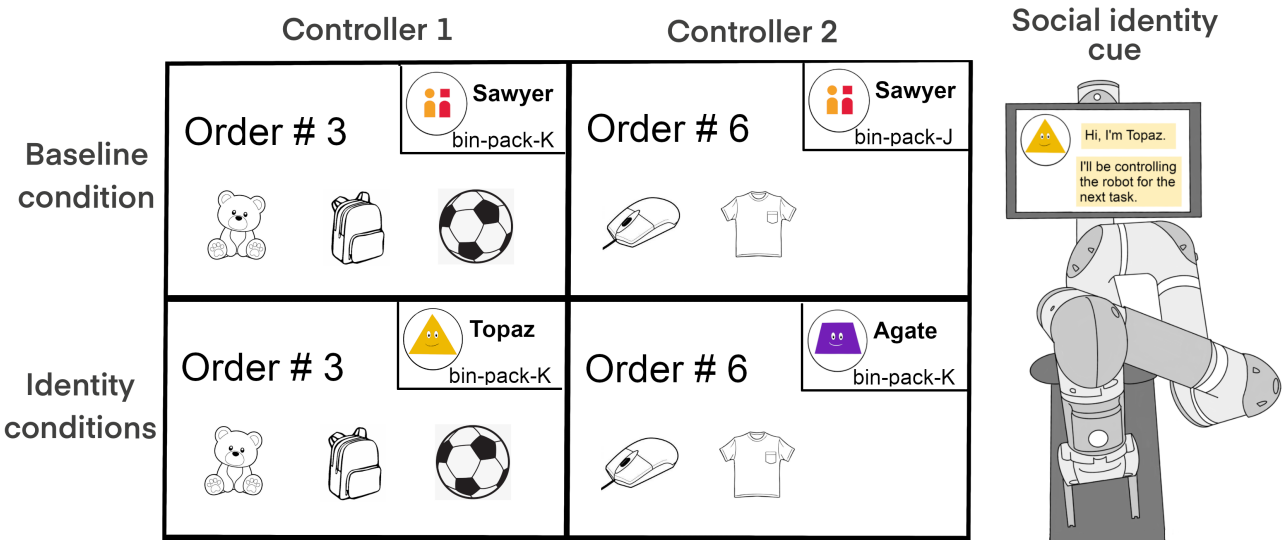


Fig. 2: Diagram of the robot's screen display in each condition, using either an algorithm name or an identity signal to differentiate between each controller. The additional introduction cue used in the social identity condition is on the right.

algorithm-name group by incorporating the context of the actively running controller.

We further expected that the social relevance of a humanlike identity and the increased understandability of the robot's behavior would make participants feel more positively about the robot, thus we hypothesize:

$H_3$ : Participants in the identity-based groups will have a more positive perception of the robot than participants in the algorithm-name group.

### B. Experimental Design

In the study, participants were randomly sorted into one of three conditions via block randomization. Participants worked as a team with the robot, supervising the robot performing an order-fulfillment task while they worked on a distractor task. Throughout the order-fulfillment task the robot switched between two types of control software. One controller was very reliable, but only grasped a subset of the items used in the task; the other could grasp any item but sometimes made mistakes. In each group, the robot had a different method of communicating which controller was operating. The robot displayed an algorithm name (in the baseline group), a weak identity signal, or a social identity signal.

*1) Robot Signals:* The three conditions differed by how the robot displayed which controller it was using. For the duration of each order, the robot displayed an agent name, a profile picture, and an algorithm name on its screen. An example of this display is pictured in Figure 2.

**Baseline**: In the baseline condition, the robot used an algorithm name: bin-pack-J for the unreliable controller, and bin-pack-K for the reliable one. Both controllers had the same agent name ("Sawyer") and profile picture.

**Weak identity**: In the weak identity signal condition, the robot used a more human-like name and a profile picture. The agent name "Agate" was associated with the profile picture of a purple trapezoid with a smiling face. The agent name "Topaz" was associated with a yellow triangle with a smiling face. For each participant, either Agate or Topaz was randomly chosen to be associated with the reliable controller, and the other associated with the unreliable controller. The algorithm name displayed with both agents was "bin-pack-K."

**Social identity**: In the social identity signal condition, the display of names, profile pictures, and algorithm name was identical to the weak condition. In this condition, before every order the robot's screen displayed an introductory message for each agent which read "Hi, I'm [agent name]. I'll be controlling the robot for the next task."

Before each order, the robot would not move until the participant had identified the name of the agent (identity conditions) or algorithm (baseline condition) displayed on the robot's screen. Their answer did not have to be correct for the robot to begin the task.

The weak identity signal was used as an intermediate signal between the social identity signal and the baseline, utilizing identity cues more passively than the social identity condition. The weak identity condition serves to unambiguously measure whether the social *performance* of an agent, rather than simply the presentation of name and picture, would allow participants to differentiate between controllers.

*2) Tasks:* Participants had two tasks to complete simultaneously: supervising the robot performing an order-fulfillment task, and doing a sorting task on a computer.

The supervision task was designed to simulate a warehouse order fulfillment environment. The team would sequentially receive 14 orders, each consisting of one to three items to be placed into a box. Each item was a wooden cube with an image of the product it represents printed on four sides, as pictured in Figure 3. The robot would pick up the items from a table and place them into a box. When they observed the robot make a mistake, participants were to press

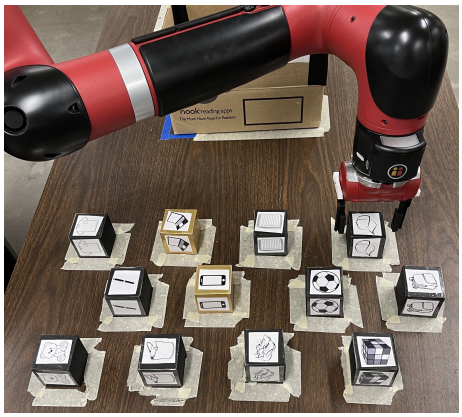

Fig. 3: The robot picking up an item from its workspace to be placed into the box for the order-fulfillment task. The cubes with gold backgrounds (e.g., laptop in the top row) are high value items.

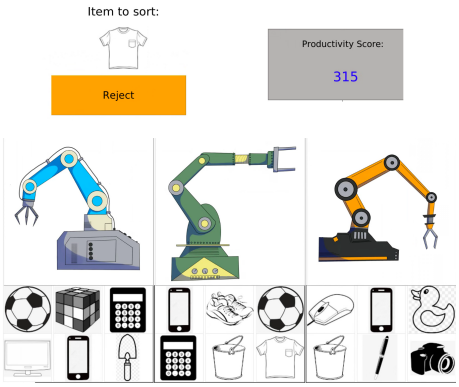


Fig. 4: The participants' distractor task. They matched each item that appeared above the "reject" button to one of the banks of items beneath the robotic arms, or rejected the item if there was no match.

a button indicating this (stopping the robot's operation) and manually fix the order by placing the correct items into the box.

Four of the 14 orders were designated high value orders, which were composed of a distinct set of objects from the other orders (the gold blocks pictured in Figure 3). In these situations, participants were asked to choose ahead of time whether the robot would fulfill the order, or if they would pack it themselves. If the robot failed, participants were not given the opportunity to fix these orders.

In the sorting (distractor) task, pictured in Figure 4, participants would see items to be sorted. They had to match items to one of three banks of allowable items (pictured beneath the robotic arms) or indicate that the item was not currently allowable. The allowable items were periodically randomized. An actual worker would have multiple responsibilities and not be able to watch a robot at all times, so participants were asked to divide their attention between supervising the robot and the sorting task.

For both tasks, participants were motivated by points. Participants could always see the team's point total on their screen, shown at the top right of Figure 4. When the robot completed an order correctly, or the participant intervened to correct an order, the team earned points; if the order was not fulfilled correctly, the team lost points. The participant could also earn points for the team by sorting items correctly on the distractor task. Participants were instructed to attempt to reach a target number of points.

*3) Robot controllers:* For all participants, the robot would run one of two software controllers. When using one controller (reliable), the robot was only able to pick up a subset of the items that appeared in orders, but made no mistakes. With the other controller (unreliable), the robot could pick up every item, but sometimes made mistakes. The reliable controller was 100% reliable, and the unreliable one was 70% reliable on a per-item basis. The robot would use the unreliable controller only when an order contained one of the items the reliable controller could not pick up. Participants were not told that the robot was using two controllers, as we were measuring whether they noticed and recalled differences between them.

The unreliable controller exhibited 4-5 failures throughout the game. All failures were problems with manipulation or perception: for example, dropping an item or grasping an unwanted item. The first two failures occurred at the same point for all participants, within the first four orders attempted. High value orders, where participants chose to assign the task to the robot or themselves, occurred at specific points throughout the tasks. The sequence of the remaining orders (including failures) was randomized for each participant.

### C. Experimental Procedure

Participants gave written consent on a form explaining the experimental procedure, then filled out a questionnaire with demographic questions and their initial impressions of the robot. After this, an experimenter verbally explained the task and demonstrated the robot using both the reliable and unreliable controller, and the identity or algorithm-name signal for the participant's group.

An experimenter then seated the participant at a table adjacent to the robot's workspace (pictured in Figure 1). Their table contained a laptop that they used to complete their portion of the task. The workspaces were situated so that participants had to turn their head toward to robot to see what it was doing. Participants were given a motion-tracking headband to wear while supervising the robot and performing their sorting task. The task took 15-20 minutes.

After completing the task, participants filled out questionnaires with questions about their views of the robot, and of each algorithm or agent they had seen. Then an experimenter conducted a brief interview with each participant, and participants were debriefed about the purpose of the study.

### D. Metrics

*1) Attention and Prediction Metrics:* Participants were seated in the field of view of motion-tracking cameras, and

wore a headband containing motion-capture markers. Their chair faced the computer displaying the user interface, and the robot was at an adjacent table. The robot's workstation was positioned so participants could not see which objects the robot was grasping without turning their heads to look at it.

To measure participant attention, we calculated whether they were looking at the computer, the robot, or elsewhere during each portion of the task. We logged when participants were actively performing the sorting task during any time of the experiment, and whether each "click" on this task was correct. We recorded whether participants identified each time the robot made a mistake in the order fulfillment task.

For the four high-value orders where the participant decided ahead of time who had responsibility for packing the order, we measured whether they assigned responsibility to themselves or to the robot.

*2) Subjective Metrics:*

**Attitudes about robot.** Prior to working with the robot, participants had answered five questions about their opinion of the robot, which they then answered again after doing the task with the robot. The post-experiment questionnaire also contained more extensive questions about their opinions of the robot. Questions were based on the Multi-Dimensional Measure of Trust scale [27] and Hoffman's fluency measures for Human-Robot Collaboration [28]. All questions were in the form of a statement, with responses falling on a 7-point scale with the endpoints "strongly agree" and "strongly disagree." Intermediate points were numbered but not labeled.

**Attitudes about controllers.** Each participant had been shown two agents or algorithms; after working with the robot, they answered whether they had seen each of four agent or algorithm names. For each name they remembered seeing, they answered questions about it. The wording by necessity varied slightly between conditions, but differed as little as possible: "I trusted the robot to complete the task correctly when the [algorithm being used/agent was active]"

**Interview** After the experiment, we asked each participant if they had noticed the agents or algorithms as they were performing the task. If they had, we asked if they had noticed any differences between them, and if so, what they were.

## IV. Results

We conducted an IRB-approved between-subjects study with 31 participants recruited from our university campus. One participant's data was discarded due to robot malfunction. The data of 30 participants were analyzed, 10 in each group. 20 participants identified as male, nine as female, and one did not say. The median age range reported was 18-24.

We found that the social identity group was significantly more likely to differentiate between the robot's controllers than the other groups. After participants completed the task with the robot and the post-experiment questionnaires, experimenters conducted a brief verbal interview. During the interview, we asked whether participants had noticed a difference between the agents or algorithms (i.e., the robot's two controllers), and if so, what the differences were. 60% of participants from the social identity group verbally described the difference between the two controllers, compared to 30% in the baseline group and 20% in the weak identity group, as visualized in Figure 5. Comments about how the controllers differed included descriptions such as one being "more reliable," "more accurate," or "less trustworthy." Participants in the social identity group were more likely to recall both the names of the controllers and the behavior of each. We did not count inaccurate comments: those which named agents or algorithms that did not appear, noted that the reliable agent or algorithm made errors, or claimed that one agent was active for the high value orders and the other completed the remaining orders.

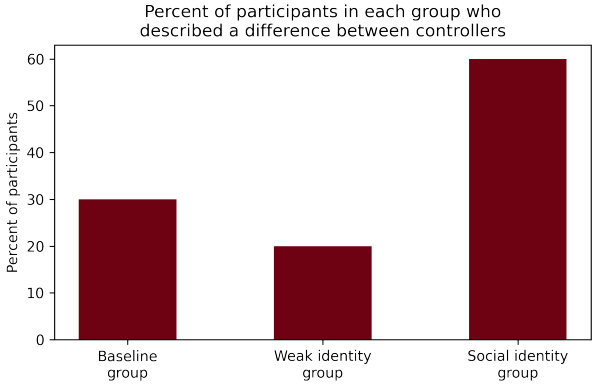


Fig. 5: After completing the task, participants were asked whether they recalled a difference between the two controllers presented to them. Each participant reported the name of the active controller 14 times while working with the robot. Despite this, **only participants in the social identity group reliably recalled the difference.**

When asked in the questionnaire about the agents or algorithms they had seen, the social identity group was the only one to record a significant difference in their confidence level between the robot's two controllers. Participants indicated their agreement on a 7-point scale with 7 being "strongly agree." Participants in the social-identity group responded with a median answer of 5 for the name associated with the unreliable controller and 6 for the name associated with the reliable controller. Median responses are shown in Figure 6. A Wilcoxon signed-rank test (used for comparison of paired non-parametric data) shows that this difference is statistically significant ($W = 0.0, p = .008$). The other groups' responses were not significantly different between controllers. This indicates that **participants in the social identity group calibrated their trust appropriately between the controllers.**

The social identity group's confidence in the unreliable controller was also significantly lower than the baseline group's confidence in the same controller. Participants in the social identity group responded with a median answer of 5 (as discussed above), while participants in the baseline group had a median response of 7. A Kruskall-Wallis test (used for comparisons of non-parametric data between more than 2

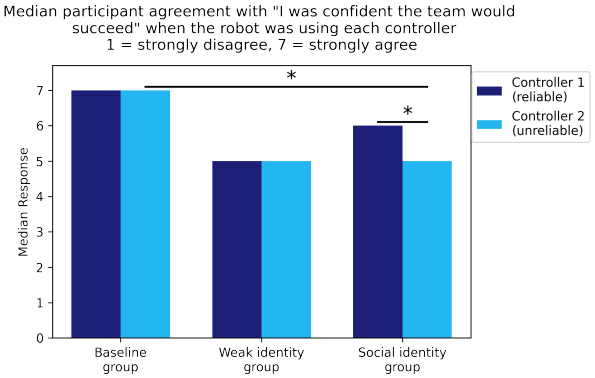


Fig. 6: After working with the robot using each controller several times, participants were asked about their confidence in each controller. * indicates that social identity group participants were significantly more confident in the reliable controller than the unreliable one ($p < .05$), and the social identity group was significantly less confident in the unreliable controller than the baseline group ($p < .05$).

groups) of all three groups suggests that there is a difference between groups. A Mann-Whitney U Test (used for comparisons of non-parametric data between independent groups) between the baseline group and the social identity group shows that the difference is statistically significant ($U = 62.0, p = .01$). This suggests that the trust-calibrating effects of experiencing the social identity signals were stronger than those of the baseline algorithm name signal.

### A. Attitudes About the Robot

**Changes in attitude** To determine the social impact of the robot's behavior, participants answered five questions about the robot before and after working with it. Participants in the social identity group were the only ones whose answers differed significantly after working with the robot.

Participants in the social identity group thought the robot was more likable after working with it. For the statement "The robot is likable," the median answer (on a 7-point scale) for participants in the social-identity group was 5 before working with the robot, and 6 after. A Wilcoxon signed-rank test shows the difference between before and after responses is statistically significant ($W = 0.0, p = .008$).

Participants in the social identity group were also less likely to think the robot was uncooperative after working with it. For the statement "The robot is uncooperative," the median answer for participants in the social-identity group was 3 before working with the robot and 1.5 after working with the robot. A Wilcoxon signed-rank test shows the difference between before and after responses in this group is statistically significant ($W = 0.0, p = .03$). Participants in the other groups did not show a statistically significant change for any of the questions after working with the robot. These results suggest that the social cues displayed by the robot gives users a positive social experience with the robot.

**General attitudes:** A Kruskal-Wallis test indicates that there is a difference between the groups in responding to "The robot had an important contribution to the success of the team." The median response was 7 for participants in the baseline group, and 6 for the social identity group. A Mann-Whitney U test shows that the difference between the baseline and social identity groups is statistically significant ($U = 81.5, p = .02$). This difference somewhat contradicts our expectations, but may be caused by the identity groups seeing both the robot and agents as entities with apparent agency.

We found an unexpected difference between groups for the statement "The robot is intelligent." The median answer for the baseline group was 6, and for the weak identity group was 4. A Mann-Whitney U test indicates that the difference between groups is significant ($U = 92, p = .01$). In general, the weak identity group's opinions of the robot were slightly more negative than the other groups.

### B. Attention and prediction

26% of participants had very limited or no activity on the distractor task, even when reminded. We could not meaningfully analyze these participants' distractor task or head-turning behavior, because they were not splitting their attention between the robot and the distractor task. We suspect that this noncompliance was due to the robot being more interesting than the sorting task.

Rates of assigning responsibility to the robot for high-value orders were nearly identical between groups.

Among participants who engaged with the distractor task, participants in the social identity group had a different supervision strategy than those in the baseline group. A Tukey's HSD test shows that social identity group participants had significantly fewer clicks on the distractor task at times when the robot was moving items than participants in the baseline group (mean clicks social identity: 36.6, baseline: 71.6, $p = .005$). There was no significant difference in activity when the robot was moving but not handling items. This finding indicates that social identity group participants watched the robot more at critical points in the task than baseline participants.

## V. Discussion and Conclusions

Despite each participant identifying the active controller **14 times** throughout the task, and overseeing at least four failures of the unreliable controller, **only participants in the social identity group could reliably differentiate between the controllers** after working with the robot. Participants in the social identity group were also more likely to report a **different level of confidence** in each controller. This difference in confidence level indicates that participants in this group calibrated their trust for each controller, not the system as a whole. Participants in the baseline and weak identity groups largely did not differentiate between the controllers. This finding **supports $\mathbf{H_1}$**, suggesting that the social identity signal is more effective than the other cues for a robot to demonstrate differences in its programming that will affect its behavior.

We did not find evidence to support $H_2$.

The social identity cues led to improved social perceptions of the robot: participants in the social identity group found the robot more likable and less uncooperative after working with it. This finding supports $H_3$. This is likely due to the social engagement of the agents' greeting messages. Since the agents introduced themselves via the robot, the sociability of the agents may have transferred to perceptions of the robot as a whole. We did not, however, observe these positive social perceptions with the weak identity cue.

A key finding from this study was that not all performances of identity are equal. While the social identity cue successfully allowed users to develop a separate understanding of each controller, the weak identity cue was not a strong enough signal to get participants to recall a difference between them. This finding is in line with previous results, which have found that there is a great deal of variability in acceptance of robot social presence cues [22], [24]. Subtle robot social cues that are subconsciously followed in human-human interactions have previously been found to be inadequate in communicating robot intent [29]. **Expressions of robot social identity likewise need to be conspicuous to be effective**, as shown through our study.

Though the small sample size was a limitation in this study, we found that treating each controller as an entity separate from other controllers or the robot itself can be a useful strategy for a robot's users. This conceptualization brings nuance to a user's understanding of a robot, enabling them to evaluate distinct elements of its behavior separately. We find that a performance of identity is an effective way to signal users of software differences, and that robot identity performances must contain social cues to be salient to human collaborators.

## REFERENCES


[1] P. Robinette, W. Li, R. Allen, A. M. Howard, and A. R. Wagner, "Overtrust of robots in emergency evacuation scenarios," in *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*. IEEE, 2016, pp. 101–108.

[2] S. Booth, J. Tompkin, H. Pfister, J. Waldo, K. Gajos, and R. Nagpal, "Piggybacking robots: Human-robot overtrust in university dormitory security," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 2017, pp. 426–434.

[3] W. A. Bainbridge, J. W. Hart, E. S. Kim, and B. Scassellati, "The benefits of interactions with physically present robots over video-displayed agents," *International Journal of Social Robotics*, vol. 3, pp. 41–52, 2011.

[4] P. H. Kahn Jr, A. L. Reichert, H. E. Gary, T. Kanda, H. Ishiguro, S. Shen, J. H. Ruckert, and B. Gill, "The new ontological category hypothesis in human-robot interaction," in *Proceedings of the 6th international conference on Human-robot interaction*, 2011, pp. 159–160.

[5] M. P. Stull and B. Hayes, "The utility of non-anthropomorphic robot identity," in *HRI Workshop on Robo-Identity*, 2024.

[6] R. B. Jackson, A. Bejarano, K. Winkle, and T. Williams, "Design, performance, and perception of robot identity," in *Workshop on Robo-Identity: Artificial identity and multi-embodiment at HRI*, vol. 2021, 2021.

[7] K. Winkle, R. B. Jackson, A. Bejarano, and T. Williams, "On the flexibility of robot social identity performance: benefits, ethical risks and open research questions for hri," in *HRI Workshop on Robo-Identity*, 2021.

[8] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "Would you trust a (faulty) robot? effects of error, task type and personality on human-robot cooperation and trust," in *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*, 2015, pp. 141–148.

[9] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," *Human factors*, vol. 39, no. 2, pp. 230–253, 1997.

[10] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human factors*, vol. 46, no. 1, pp. 50–80, 2004.

[11] B. M. Muir, "Trust between humans and machines, and the design of decision aids," *International journal of man-machine studies*, vol. 27, no. 5-6, pp. 527–539, 1987.

[12] P. Robinette, A. M. Howard, and A. R. Wagner, "Timing is key for robot trust repair," in *Social Robotics: 7th International Conference, ICSR 2015, Paris, France, October 26-30, 2015, Proceedings 7*. Springer, 2015, pp. 574–583.

[13] A. L. Baker, E. K. Phillips, D. Ullman, and J. R. Keebler, "Toward an understanding of trust repair in human-robot interaction: Current research and future directions," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 8, no. 4, pp. 1–30, 2018.

[14] E. J. De Visser, M. M. Peeters, M. F. Jung, S. Kohn, T. H. Shaw, R. Pak, and M. A. Neerincx, "Towards a theory of longitudinal trust calibration in human–robot teams," *International journal of social robotics*, vol. 12, no. 2, pp. 459–478, 2020.

[15] T. Jensen and M. M. H. Khan, "I'm only human: The effects of trust dampening by anthropomorphic agents," in *International Conference on Human-Computer Interaction*. Springer, 2022, pp. 285–306.

[16] K. Okamura and S. Yamada, "Adaptive trust calibration for human-ai collaboration," *Plos one*, vol. 15, no. 2, p. e0229132, 2020.

[17] R. Moradinezhad and E. T. Solovey, "Investigating trust in interaction with inconsistent embodied virtual agents," *International Journal of Social Robotics*, vol. 13, no. 8, pp. 2103–2118, 2021.

[18] J. B. Lyons, "Being transparent about transparency: A model for human-robot interaction," in *2013 AAAI Spring Symposium Series*, 2013.

[19] B. Hayes and J. A. Shah, "Improving robot controller transparency through autonomous policy explanation," in *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*, 2017, pp. 303–312.

[20] R. H. Wortham, A. Theodorou, and J. J. Bryson, "Robot transparency: Improving understanding of intelligent behaviour for designers and users," in *Towards Autonomous Robotic Systems: 18th Annual Conference, TAROS 2017, Guildford, UK, July 19–21, 2017, Proceedings 18*. Springer, 2017, pp. 274–289.

[21] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[22] M. Luria, S. Reig, X. Z. Tan, A. Steinfeld, J. Forlizzi, and J. Zimmerman, "Re-embodiment and co-embodiment: Exploration of social presence for robots and conversational agents," in *Proceedings of the 2019 on Designing Interactive Systems Conference*, 2019, pp. 633–644.

[23] S. Reig, E. J. Carter, T. Fong, J. Forlizzi, and A. Steinfeld, "Flailing, hailing, prevailing: Perceptions of multi-robot failure recovery strategies," in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 2021, pp. 158–167.

[24] S. Reig, M. Luria, E. Forberger, I. Won, A. Steinfeld, J. Forlizzi, and J. Zimmerman, "Social robots in service contexts: Exploring the rewards and risks of personalization and re-embodiment," in *Designing Interactive Systems Conference 2021*, 2021, pp. 1390–1402.

[25] T. Williams, D. Ayers, C. Kaufman, J. Serrano, and S. Roy, "Deconstructed trustee theory: Disentangling trust in body and identity in multi-robot distributed systems," in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. Association for Computing Machinery, 2021, p. 262–271.

[26] J. Kennedy, P. Baxter, and T. Belpaeme, "Comparing robot embodiments in a guided discovery learning interaction with children," *International Journal of Social Robotics*, vol. 7, pp. 293–308, 2015.

[27] D. Ullman and B. F. Malle, "Mdmt: multi-dimensional measure of trust," 2019.

[28] G. Hoffman, "Evaluating fluency in human–robot collaboration," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 3, pp. 209–218, 2019.

[29] H. Admoni, A. Dragan, S. S. Srinivasa, and B. Scassellati, "Deliberate delays during robot-to-human handovers improve compliance with gaze communication," in *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, 2014, pp. 49–56.