Forming and Updating Implicit Impressions of Robot Competence

Minae Kwon*, Melissa Ferguson†, Thomas Mann†, and Ross Knepper*
*Department of Computer Science, Cornell University, Ithaca, NY, USA
†Department of Psychology, Cornell University, Ithaca, NY, USA

***Abstract*—Given the importance of setting accurate expectations of robot capabilities in humans, we explore how people form implicit judgments toward a robot over a prolonged interaction. In our study, we created a live interaction between a robot deli cashier and a human customer. The robot displays a competent behavior followed by another competent behavior (Condition A), an incompetent behavior (Condition B), or an incompetent behavior with a warning (Condition C). We measured implicit judgments of competence toward the robot using an affect misattribution procedure over time and by condition. Contrary to what many studies on implicit updating predict, results show that participants updated implicit impressions of competence over time with minimal pieces of information. However, we found no significant effect of warnings on guarding against a drop in competence judgments.**

I. Introduction

Seamless human-robot collaboration necessitates accurate expectations of collaborator capabilities. Socially intelligent robots exacerbate this problem, as their socially adept behaviors can increase the robot's perceived intelligence [1, 8]. Therefore, it is likely that in many situations, people's perceptions of a robot's capabilities and their actual capabilities are not aligned, creating an expectations gap. Incorrectly generalizing capabilities creates false expectations, setting people up for disappointment and eventually mistrust [7]. A lack of trust has been shown to impair team performance [2, 9].

The first step in solving the expectations gap problem is to understand how humans form impressions of robots over time. Our aim in this work is to examine how people implicitly judge robot capabilities during an interaction. We measure implicit judgments because they are not prone to self-presentation biases that we might find with explicit measurements like Likert scales. Wilson et al. define implicit attitudes as attitudes that have an unknown origin, are activated automatically, and influence uncontrollable responses [16]. Implicit measures differ from explicit measures in that they assess automatic evaluations indirectly, or without asking the participant to report his or her attitude [3]. We assess implicit judgments with the Affect Misattribution Procedure (AMP) [10], a widely-used implicit measure in psychology.

We contribute three findings. First, given that many studies in psychology show that it is difficult for humans to update implicit attitudes toward other humans [6, 15, 13], we provide evidence that robots can unintentionally update human judgments of their competence with ease through their actions. Second, despite this ease of updating, we also find that *intentionally* manipulating these judgments may be less straightforward than previously thought. Third, we find that observations of one robot affect implicit attitudes towards similar but unseen robots as well.

Fig. 1. Setup of deli with Baxter the cashier robot.

II. Experiment

In our study, we recruited N=217 participants to participate in a Wizard-of-Oz live interaction with Baxter (Fig. 1). The study was set in a deli where participants were the customer and Baxter was a robot deli worker. Each participant was asked to place two orders. Throughout the study, we used the AMP to measure participants' implicit judgments of Baxter's competence three times: once before the first order, once after the first order, and once after the second order.

A. Procedure

Participants completed the first AMP prior to coming into the study (Time 1). In each AMP, we presented sixty trials. In each trial, a prime was first presented (shown for 75ms), followed by a neutral face (shown for 100ms), and then a backward mask. We then asked participants to rate whether the neutral target image was more or less competent than average. The idea behind the measure is that participants will unintentionally misattribute trait judgments from the prime to the neutral target image [10]. We used four primes in our study: a cash register, a human, Baxter, and PR2, another humanoid robot. Baxter was our main prime. Our goal was to measure how participants primed with Baxter rated the target relative to the three other primes. The cash register

and human served as lower and upper bound comparisons of competence judgments respectively while PR2 was included to see if competence judgments of Baxter generalized to similar humanoid robots. We randomly alternated usage of several human primes to avoid overfitting on a specific type of face. We used a computer generated human face as our neutral target image.

A couple of days later participants came into the lab to complete the second portion of the study. Participants interacted with Baxter twice where they ordered a granola bar in the first interaction (Time 2) and ordered a sandwich with a condiment during the second interaction (Time 3). During the orders, Baxter greeted the participant and asked the participant what they would like to order. After an order was placed and delivered, Baxter wished the participant a nice day.

During Time 2, Baxter displayed competent behavior by successfully delivering the granola bar in all conditions. The purpose of Time 2 was to set an initial expectation of Baxter's capabilities. The experimenter then led the participant to the computer to complete a manipulation check and the second AMP.

During Time 3, participants placed a more complicated order–a sandwich and a condiment. In our control condition (Condition A), Baxter successfully delivered the sandwich and condiment. In our experimental conditions, Baxter delivered the incorrect sandwich and condiment without a warning (Condition B), or with a warning (Condition C). In the warning, Baxter stated, "I sometimes have difficulty with more complex orders." In both Conditions B and C, Baxter showed confusion by moving his grippers back and forth between two similar sandwiches and condiments. (Videos of Baxter's actions can be found *here*.)

Finally, participants were led back to the computer to complete the third AMP. Afterwards, participants were compensated and debriefed.

B. Results

Our experiment was a 3 (Time: 1, 2, 3) x 4 (Prime: Baxter, PR2, Control faces, Cash register) x 3 (Condition: Always succeeds (1), fails with no warning (2), fails with warning (3)) study. We created a measure of participants' implicit competence judgments by calculating the proportion of times participants indicated that the neutral face target image was more competent than average for each time, condition, and prime. In our analyses, we first looked at whether participants updated implicit judgments over time.

1) Updating from Time 1 to Time 2: In order to test whether updating of competence judgments happened from Time 1 to Time 2, we tested interactions of curtailed time (Time 1, Time 2) and prime type within each condition. In Condition A, there was no interaction of time and prime. This suggests that there was no updating across Time 1 and Time 2, perhaps because participants' implicit judgments of Baxter were unusually high at Time 1 in this condition, leaving less room for updating in Time 2. In Condition B, there was an interaction between curtailed time and prime $F(2.75,\ 209) = 9.23$ $p < .001$. Looking at simple effects, implicit judgments toward Baxter marginally significantly *increased* from Time 1 to Time 2, $F(1,\ 76) = 3.34$ $p = .07$, (Time 1 $M = .54$, Time 2 $M = .59$). This suggests that participants thought Baxter was significantly more competent after watching Baxter succeed at delivering the granola bar (Time 2). In Condition C, there was again an interaction between time and prime, $F(2.84,\ 165) = 4.14$, $p = .009$. Implicit judgments toward Baxter again significantly *increased* from Time 1 to Time 2, $F(1,\ 58) = 14.23$, $p < .001$ (Time 1 $M = .49$, Time 2 $M = .61$). This similarly indicates that participants thought Baxter was significantly more competent after observing Baxter's behavior during Time 2. Although there were no differences in Baxter's movements across conditions from Time 1 to Time 2, we suspect results differ slightly among conditions because of noise.

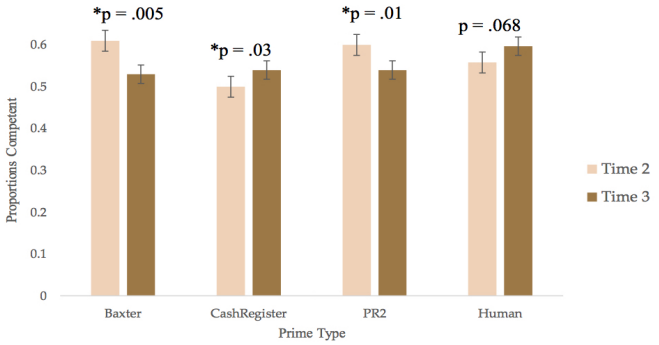

Fig. 2. Implicit updating of competence judgments from Time 2 to Time 3. Updating was significant for all four prime types. Competence judgments toward Baxter dropped despite the warning.

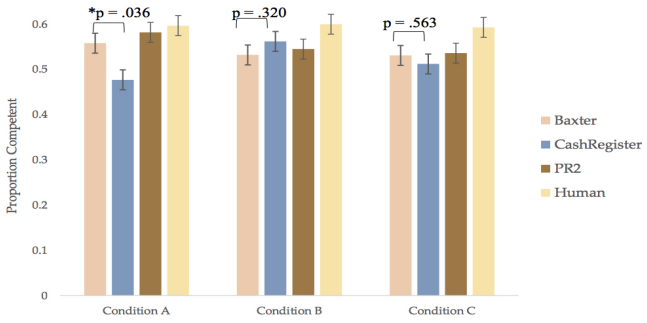

Fig. 3. During Time 3, competence judgments toward Baxter were significantly higher than the cash register in Condition A, but there was no significant difference between the two in Conditions B and C.

2) Updating from Time 2 to Time 3: We collapsed Conditions B and C and tested simple effects of prime from Time 2 to Time 3 (Fig. 2). We excluded Condition A from our analysis because we did not predict implicit updating for that condition (we introduced new evidence, a failure interaction, only in Conditions B and C). Implicit judgments toward Baxter *decreased* significantly from Time 2 to Time 3, $F(1,\ 135) = 8.21$, $p = .005$ (Time 2 $M = .61$, Time 3 $M = .53$). This result suggests that participants implicitly thought Baxter was significantly less competent after observing Baxter fail during Time 3.

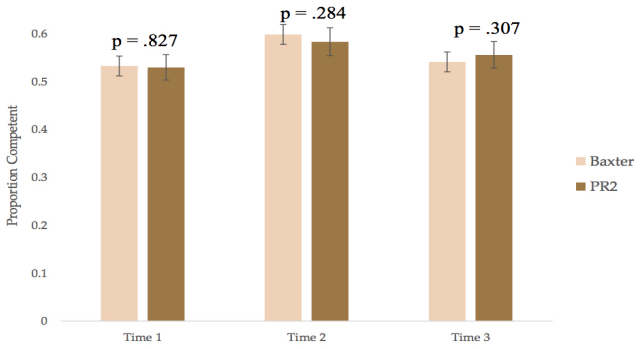

Fig. 4. Competence judgments toward Baxter and PR2 did not significantly differ across time.

3) No effect of warning: We next looked at whether the warning in Condition C guarded against a drop in competence judgments. At Time 3, there was no significant difference in competence judgments toward Baxter between Conditions B and C, p = .991 (Cond. B M = .532, Cond. C M = .531). As Fig. 3 shows, at Time 3, there is no significant difference between Baxter and the cash register in Conditions B and C while there was a difference in Condition A. This implies that the warning did not help guard against a drop in competence judgments in Condition C.

4) Grouping of Baxter and PR2: Lastly, implicit judgments of PR2, another robot which participants did not interact with, mirror those of Baxter. Participants not only formed similar initial impressions of competence but also persistently updated impressions of competence of the two robots in the same manner across time, (Time 1 p = .83, Time 2 p = .28, Time 3 p = .31) (Fig. 4).

III. Discussion

We found that people were quick to update their implicit impressions of competence after observing new and inconsistent displays of competence. People implicitly increased competence judgments of Baxter to human levels after a single display of competent behavior and then proceeded to decrease competence judgments to cash register levels after a single display of incompetent behavior.

There are many possible reasons why people update implicit impressions so easily toward robots. One possible contributing explanation for implicit updating in the present work could be the novelty effect of robots. Novel stimuli are thought to be more informative, and thus carry more weight when forming impressions of a target [4].

Implicit updating of judgments toward robots implies that robots can unintentionally manipulate impressions of competence. Not only will impressions of competence be influenced when a robot malfunctions, but also a robot that is simply doing its job could inaccurately raise expectations of competence. In future work, it will be important for robots to take advantage of people's capacity for implicit updating and learn behaviors that can accurately set expectations of its capabilities.

Surprisingly, there were no effects of warning on competence judgments. Baxter in Condition C was judged as implicitly competent as Baxter in Condition B, who failed without warning. There are several possible explanations for this. First, the failure to deliver the sandwiches and condiments could have been so great that a warning was not enough to change's people's impressions of Baxter's decreased competence. Second, Baxter's role as a service robot could have increased participants' expectations of Baxter's capabilities and his ability to address customers' needs.

A closely-related study by Lee et al. [8] investigates the effect of forewarning strategies on perceptions of a robot and its service in service robots. The authors experiment with verbal and nonverbal forewarning strategies that lower expectations before the service robot fails a task. The verbal forewarning was very similarly worded to the warning in our study. They find that forewarning strategies generally increased explicit ratings of robot competence. This finding differs from our results. A possible explanation for this discrepancy is that a person's implicit and explicit attitudes can often be inconsistent [16, 12, 14]. Rydell & McConnell [12] suggest that there are separate cognitive processes that form implicit and explicit attitudes, making it possible for someone to hold inconsistent explicit and implicit attitudes. More research is needed to understand how the discrepancy between implicit and explicit judgments toward robot competence will affect human behavior, and whether or not it will be important to reconcile differences between implicit and explicit attitudes. Another possible explanation for the difference between our results and Lee et al.'s results is that participants in Lee et al.'s study observed the robot's failure from a third-person's perspective whereas in our study, participants were directly affected by the robot's incompetence. This difference in perspective could have also made the warning in our study less effective in mitigating negative competence judgments.

Finally, we note that implicit judgments of PR2 are similar to judgments of Baxter across time. The coupling suggests that participants are generalizing Baxters competence levels to other similar looking robots. This finding is consistent with the literature on automatic updating between in-group and out-group members. Implicit attitudes about a target are more likely to transfer to new individuals if the new individuals are members of the same group [11], or are similar in appearance [5]. Evidence for generalization across robots lends important insights when designing robots for multi-robot human interaction. When teaming with multiple distinct robots, it may be more difficult for humans to form accurate expectations of individual robot capabilities. Furthermore, one robots performance could drastically affect how the person interacts with other robots. Thus, in future work, it is important that we design robot behaviors that allow humans to individualize robots.

References

[1] Christoph Bartneck, Takayuki Kanda, Omar Mubin, and Abdullah Al Mahmud. Does the design of a robot influence its animacy and perceived intelligence? *International Journal of Social Robotics*, 1(2):195–204, 2009.

[2] Kurt T Dirks. The effects of interpersonal trust on work

group performance. *Journal of applied psychology*, 84 (3):445, 1999.

[3] Melissa J Ferguson and Jun Fukukura. Likes and dislikes: A social cognitive perspective on attitudes. *The SAGE Handbook of Social Cognition. SAGE Publications Ltd, London*, pages 165–186, 2012.

[4] Susan T Fiske. Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of personality and Social Psychology*, 38(6):889, 1980.

[5] Bertram Gawronski and Kimberly A Quinn. Guilty by mere similarity: Assimilative effects of facial resemblance on automatic evaluation. *Journal of Experimental Social Psychology*, 49(1):120–125, 2013.

[6] Aiden P. Gregg, Beate Seibt, and Mahzarin R. Banaji. Easier done than undone: Asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology*, 90(1):1–20, 2006. doi: 10.1037/0022-3514.90.1.1.

[7] Victoria Groom and Clifford Nass. Can robots be teammates?: Benchmarks in human–robot teams. *Interaction Studies*, 8(3):483–500, 2007.

[8] Min Kyung Lee, Sara Kielser, Jodi Forlizzi, Siddhartha Srinivasa, and Paul Rybski. Gracefully mitigating breakdowns in robotic services. In *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*, pages 203–210. IEEE Press, 2010.

[9] Roger C Mayer, James H Davis, and F David Schoorman. An integrative model of organizational trust. *Academy of management review*, 20(3):709–734, 1995.

[10] B Keith Payne, Clara Michelle Cheng, Olesya Govorun, and Brandon D Stewart. An inkblot for attitudes: affect misattribution as implicit measurement. *Journal of personality and social psychology*, 89(3):277, 2005.

[11] Kate A Ratliff and Brian A Nosek. Negativity and outgroup biases in attitude formation and transfer. *Personality and Social Psychology Bulletin*, 37(12):1692–1703, 2011.

[12] Robert J Rydell and Allen R McConnell. Understanding implicit and explicit attitude change: a systems of reasoning analysis. *Journal of personality and social psychology*, 91(6):995, 2006.

[13] Robert J. Rydell and Allen R. Mcconnell. Understanding implicit and explicit attitude change: A systems of reasoning analysis. *Journal of Personality and Social Psychology*, 91(6):9951008, 2006. doi: 10.1037/0022-3514.91.6.995.

[14] Robert J Rydell, Allen R McConnell, Diane M Mackie, and Laura M Strain. Of two minds: Forming and changing valence-inconsistent implicit and explicit attitudes. *Psychological Science*, 17(11):954–958, 2006.

[15] Robert J. Rydell, Allen R. Mcconnell, Laura M. Strain, Heather M. Claypool, and Kurt Hugenberg. Implicit and explicit attitudes respond differently to increasing amounts of counterattitudinal information. *European Journal of Social Psychology*, 37(5):867878, 2007. doi: 10.1002/ejsp.393.

[16] Timothy D Wilson, Samuel Lindsey, and Tonya Y Schooler. A model of dual attitudes. *Psychological review*, 107(1):101, 2000.