An Exploration of Implicit Attitudes Towards Robots Using Implicit Measures

Minae Kwon
Cornell University
Department of Computer Science

Melissa Ferguson
Cornell University
Department of Psychology

Thomas Mann
Cornell University
Department of Psychology

Ross Knepper
Cornell University
Department of Computer Science

ABSTRACT

Given the importance of setting accurate expectations of robot capabilities in humans, we explore how people form implicit competence judgments toward a robot over a prolonged interaction, and how durable those implicit judgments are. 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 the affect misattribution procedure over time and by condition. We then measured the durability of these implicit impressions after several months. Contrary previous work, results show that there is no immediate effect of warning on guarding against a drop in implicit competence judgments. We find that the effect of warning emerge only months later.

1 INTRODUCTION

As robots with diverse and complex capabilities emerge, they will need to predict how people perceive and generalize their capabilities. In order to build that capacity in robots, we need to understand how people form and change their impressions of robot capabilities in response to robot actions.

However, research is still ongoing concerning how and when humans update their impressions of other *humans*, much less of robots [3]. Impression formation is complex: impressions can be implicit as well as explicit [4, 5]. Implicit impressions are those measured indirectly – that is, without asking the person directly about what they think of another person (or object). They are instead measured by inferring how someone responds to a given stimulus based on how exposure to that stimulus affects their responses to other targets [12]. On the other hand, explicit impressions are deliberately formed evaluations in response to being asked to report a judgment [6]. They are evaluated through traditional self-report measures.

According to past work in psychology, people can form complex and even conflicting implicit and explicit impressions toward a target object. These conflicting impressions can predict behavior differently depending on a mix of personal and contextual factors [1, 4, 5]. Currently, explicit measures are overwhelmingly the dominant form of assessment in human-robot interaction [2]. We argue that explicit measures alone will give us an incomplete understanding of how people form and update impressions.

Our goal is to create a more complete understanding of how people form and update impressions of robot capabilities. In this work, we take a step toward that direction by measuring implicit impressions of robot *competence*, an aggregate measure of people's perceptions of a robot's capabilities.

We conducted a large Wizard-of-Oz user study that examines how humans form and update implicit impressions of robot competence in response to different robot behaviors: success, failure, and warning with failure. We measured implicit impressions using the Affect Misattribution Procedure (AMP) [11], a well-established measure in psychology. Finally, we measured the durability of these implicit impressions over time. Ultimately, we show that there is a delayed effect of warnings on implicit competence judgments. Our results differ with past work examining the effect of warnings on explicit impressions in HRI [8, 10], and highlight the importance of studying implicit impressions.

2 FORMING AND UPDATING IMPLICIT IMPRESSIONS

We examined how people develop implicit impressions of robot competence in response to robot success and failure. In particular, we also look at responses to warning before failure to see if simple expectation setting tactics such as warnings are effective in changing implicit impressions. N=217 participants took part in a Wizard-of-Oz interaction with the Baxter robot. The study was set in a deli where participants were the customer and Baxter was a robot deli worker.

Procedure. Participants interacted with Baxter twice. They ordered a granola bar in the first interaction and ordered a sandwich with a condiment during the second interaction. During Interaction 1, Baxter displayed competent behavior by successfully delivering the granola bar. The purpose of Interaction 1 was to set an initial expectation of Baxter's capabilities. Interaction 2 represents a more complicated order.

In the first condition (Condition A), Baxter successfully delivered the sandwich and condiment. In our experimental conditions, Baxter delivered the incorrect sandwich and condiment without 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, to give the illusion that participants were able to choose their own order, we asked them to randomly pick an order out of a cup. Unbeknownst to the participants, the cup contained identical orders.

Materials. After each interaction, we employed a manipulation check that asked participants to indicate whether Baxter correctly delivered the order in a "yes" or "no" questionnaire. If participants answered "no", they were asked to describe what was incorrect about the order.

Implicit evaluations of competence were measured three times, once before the experiment (Time 1), once after the first interaction (Time 2), and once after the second interaction (Time 3).

The AMP was used to measure implicit judgments. In each AMP, we presented 60 trials. On each trial, a prime was first presented (shown for 75ms), followed by a neutral face (shown for 100ms), and then a backward mask. On each trial, we 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 [11].

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. In order to create a measure of participants' implicit competence judgments, we calculated the proportion of times participants indicated that the neutral face target image was more competent than average for each time, condition, and prime. We used a computer generated human face as our neutral target image. Although AMPs have generally been used to measure affect, the procedure has been shown to be effective in measuring other trait judgments as well [7, 9].

2.1 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. Among our findings, we present a result that shows no effect of warning on implicit competence judgments.

We looked at whether the warning in Condition C guarded against a drop in competence judgments. At Time 3, implicit judgments toward Baxter decreased significantly from Time 2 to Time 3 for both Conditions B and C, $F(1, 135) = 8.21$, $p = .005$. This means that participants rated Baxter with the warning and Baxter without the warning as implicitly just as incompetent. There was no significant difference in competence judgments toward Baxter between Conditions B and C, $p = .991$ (Cond. B $M = .532$, Cond. C $M = .531$). Finally, there was 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.

3 DURABILITY OF IMPLICIT IMPRESSIONS

We wanted to see whether the implicit impressions formed during the first study were durable. We contacted participants from the first study and asked them to complete a fourth AMP, designated Time 4, and answer survey questions. Participants were not given any new information about Baxter. Of the 217 participants, we received N = 108 responses. There was no effect of condition on whether participants returned to the study, chi-square(2) = .347, $p = .841$. Participants were aware of the hypotheses we made from the first study because they were debriefed after the experiment. However, we do not think the debriefing affected the results for our second study because their explicit responses did not reflect our hypotheses from the first study (explicit measures usually reflect demand biases, but we found none).

Procedure. Participants received a link to the study via email. The link then directed them to a form where they provided informed consent. Participants then completed an AMP followed by a series of survey questions. Finally, participants were debriefed and compensated with $5 Amazon gift cards.

Materials. We used the same AMP to measure implicit judgments as we did in Study 1. We also added a series of explicit measures, questions on a seven-point Likert scale, at the end of the study. These questions assessed participants' intentions of interacting with Baxter in the future, and how they viewed Baxter as a teammate.

3.1 Results

Implicit Competence Judgments. As part of our analysis, we first conducted a broad omnibus test of prime (4) x condition (3) at Time 4 on the multivariate trace (Pillai's Trace), $F(6, 208) = 1.66$, $p = .113$. There was no statistically significant evidence of variation in the overall prime effect across the three conditions at Time 4.

When looking within conditions at Time 4, however, we surprisingly found that there was a main effect on implicit measures in Condition C, $F(3, 99) = 3.75$, $p = .013$ (sphericity holds). In that condition, Baxter was more implicitly competent than PR2 ($p = .02$), the cash register ($p = .016$), and no different from the control faces ($p = .969$). A very specific test of that contrast (The contrast of condition C vs. A and B on the difference between Baxter and the PR2 and control faces) was significant, $F(1, 105) = 5.21, p = .024$, as was one that additionally lumps Baxter together with the human faces, $F(1, 105) = 4.37$, $p = .039$. When comparing conditions, Baxter in Condition C ($M = 0.59$) was rated as marginally more significant than Baxter in Condition B ($M = 0.49$) ($p = .076$), but no different from Baxter in Condition A ($M = 0.58$), $p = 0.091$. These results suggest that the positive effects from the warning in Condition C may have only been realized after a delay.

Explicit Judgments. We compared Conditions using a one-way ANOVA with contrast tests for each Likert-scale question. Among our results, the most interesting we found was that the conditions at Time 4 were significantly different with regards to explicit measures of competence $F(2, 104) = 11.66$, $p < .001$. In contrast with implicit impressions, there was no difference in explicit competence levels between Condition B and Condition C, $p = .771$. Baxter in Condition A was rated as significantly more competent than both Condition C, $p < .001$ and Condition B, $p < .001$. This result suggests that people explicitly rated Baxter in Conditions B and C as similarly incompetent. This implies that participants in Condition C had a duality between implicit and explicit impressions of competence.

4 DISCUSSION

Our goal was to provide a more comprehensive understanding of how people form impressions of robot competence. We used the Affect Misattribution Procedure to measure implicit impressions of competence and found a delayed effect of warning. Our result contradicts what previous work on explicit impression formation and updating toward robots find–lowering expectations by using tactics like warnings has been repeatedly shown to mitigate a drop in explicit competence judgments without delay [8, 10]. We also discovered a dissociation between explicit and implicit impressions toward Baxter's competence at Time 4, Condition C. Although participants implicitly believed Baxter to be as competent as Baxter in the success condition (Condition A), they explicitly reported that Baxter was as incompetent as Baxter in the failure condition (Condition B). Our findings highlight the complex impression formation and updating process. We take a first step in providing a more comprehensive understanding of how people form impressions of robot capabilities.

REFERENCES

[1] Mahzarin R Banaji and Anthony G Greenwald. 2013. *Blindspot: Hidden biases of good people.* Random House.

[2] Cindy L Bethel and Robin R Murphy. 2010. Review of human studies methods in HRI and recommendations. *International Journal of Social Robotics* 2, 4 (2010), 347–359.

[3] Jeremy Cone, Thomas C Mann, and Melissa J Ferguson. 2017. Changing Our Implicit Minds: How, When, and Why Implicit Evaluations Can Be Rapidly Revised. *Advances in Experimental Social Psychology* (2017).

[4] Clayton R Critcher and Melissa J Ferguson. 2016. âĂIJWhether I like it or not, itâĂŹs importantâĂİ: Implicit importance of means predicts self-regulatory persistence and success. *Journal of personality and social psychology* 110, 6 (2016), 818.

[5] Russell H Fazio and Michael A Olson. 2014. The MoDe Model. *Dual-process theories of the social mind* (2014), 155.

[6] Bertram Gawronski and Galen V Bodenhausen. 2014. Implicit and explicit evaluation: A brief review of the associative–propositional evaluation model. *Social and Personality Psychology Compass* 8, 8 (2014), 448–462.

[7] Regina Krieglmeyer and Jeffrey W Sherman. 2012. Disentangling stereotype activation and stereotype application in the stereotype misperception task. *Journal of personality and social psychology* 103, 2 (2012), 205.

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

[9] Kristjen B Lundberg and B Keith Payne. 2014. Decisions among the undecided: Implicit attitudes predict future voting behavior of undecided voters. *PloS one* 9, 1 (2014), e85680.

[10] Steffi Paepcke and Leila Takayama. 2010. Judging a bot by its cover: an experiment on expectation setting for personal robots. In *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on.* IEEE, 45–52.

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

[12] Jeffrey W Sherman, Bertram Gawronski, and Yaacov Trope. 2014. *Dual-process theories of the social mind.* Guilford Publications.