Human Expectations of Social Robots

Minae Kwon, Malte F. Jung, and Ross A. Knepper
Computing and Information Science, Cornell University, Ithaca, NY, USA

***Abstract*—A key assumption that drives much of HRI research is that human-robot collaboration can be improved by advancing a robot's capabilities. We argue that this assumption posits a major challenge to developing social robots. Increasing social capabilities in robots can produce an *expectations gap* where humans develop unrealistically high expectations of social robots due to generalization from human mental models. By conducting two studies with 674 participants, we examine how people develop and adjust mental models of robots. We find that both a robot's physical appearance and its behavior influence how we form these models. This suggests it is possible for a robot to unintentionally manipulate a human into building an inaccurate mental model of its overall abilities simply by displaying a few capabilities that humans possess, such as speaking and turn-taking. We conclude that this expectations gap, if not corrected for, could ironically result in less effective collaborations as robot capabilities improve. In this paper, we first describe our research and then discuss related challenges that can arise in real-life settings.**

I. Introduction

Given the difficult nature of integrating robots into tasks that need human collaboration, the advance of anthropomorphic and sociable robots has made significant progress. The effectiveness of human-robot collaboration is limited by the lack of robot skills, both technical and social. By increasing skills in both areas, it is believed that interaction will be deeper, tighter bonds will form, and the collaboration will proceed more smoothly [2].

Often, however, socially intelligent robots give the impression that they are more intelligent than they really are. For example, research has found that perceptions of animacy and intelligence are closely related and simply making a robot more human-like in its appearance and behavior increases perceptions of intelligence [1]. It is therefore likely that in many situations people's perceptions of a robot's intelligence and its actual capabilities are not well aligned. The question of how a robot's embodiment and behavior shapes perceptions of specific capabilities is underexplored. We therefore introduce the term *expectations gap* to describe this under-studied phenomenon that occurs when humans encounter complex engineered systems and form expectations that are misaligned with the system's capabilities. Today's engineers build robots to be good at specific capabilities. In contrast, humans are generally adept at a broad set of capabilities. Humans also have a tendency to assign agency to, or anthropomorphize, human-like objects [5], including robots [7]. When seeing robots that seem sociable or anthropomorphic, it is easy for us to generalize human mental models to robots [2]. We normally trust others to be able to perform a common set of core capabilities, such as speaking or walking. Therefore, when attributing a human mental model to a robot, we hypothesize that humans will initially overestimate the robot's actual breadth of capabilities.

Fig. 1: Anthropomorphism in robots is a double-edged sword, leading to both smooth interaction with humans and unrealistically high expectations due to human mental models that generalize capabilities from humans to social robots.

The harm lies in the fact that incorrectly generalizing capabilities creates misplaced trust due to false expectations, setting people up for disappointment and eventually mistrust [4]. A lack of trust has been shown to impair team performance [8] [3] and an expectations gap could even provoke dangerous situations as robots increasingly support safety-critical tasks in surgery or search and rescue.

Through a prolonged interaction, people do figure out and adapt to a robot's idiosyncrasies. However, there is a whole class of tasks involving brief interactions (such as customer service) in which the interaction is over before the human user has been able to recognize the robot's true capabilities, much less adjust to them. We therefore hold that the expectations gap is a genuine problem in human-robot interaction that must be better understood.

In this paper, we present two studies that contribute preliminary evidence in support of the hypotheses that (1) humans construct distinct theory of mind models of machines and people, (2) people attribute more human mental models to more social robots, and (3) mental models can be changed by the robot's behavior. We then address the implications of these findings on the methodology required to support real-life HRI scenarios deployed long-term outside of the lab.

II. Study 1: Measuring Expectations

In a two (Context: industrial vs domestic) by three (Level of anthropomorphism of agent: industrial robot vs. humanoid robot vs. human) between subjects study with N=600 participants from Amazon Mechanical Turk (AMT), we examined

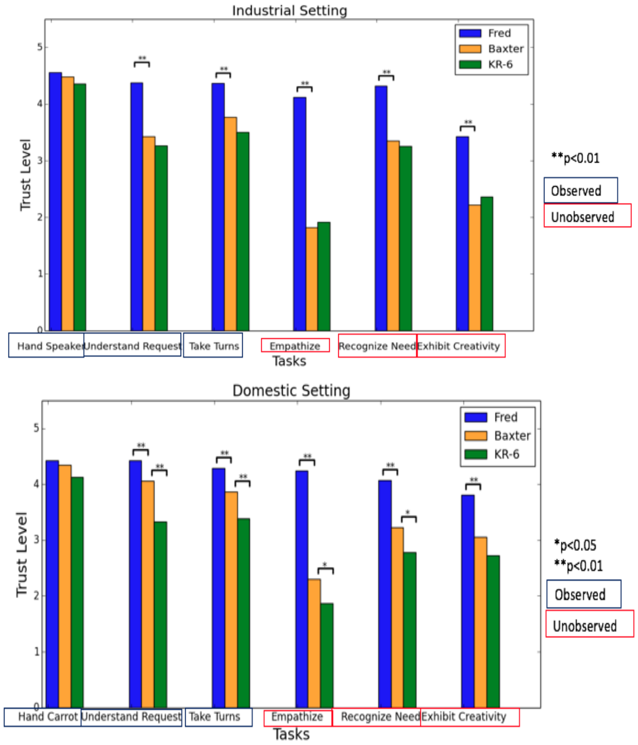

Fig. 2: Users' trust of robots performing social tasks in industrial (top) and domestic (bottom) settings. Robots were better discriminated by social tasks in the kitchen setting.

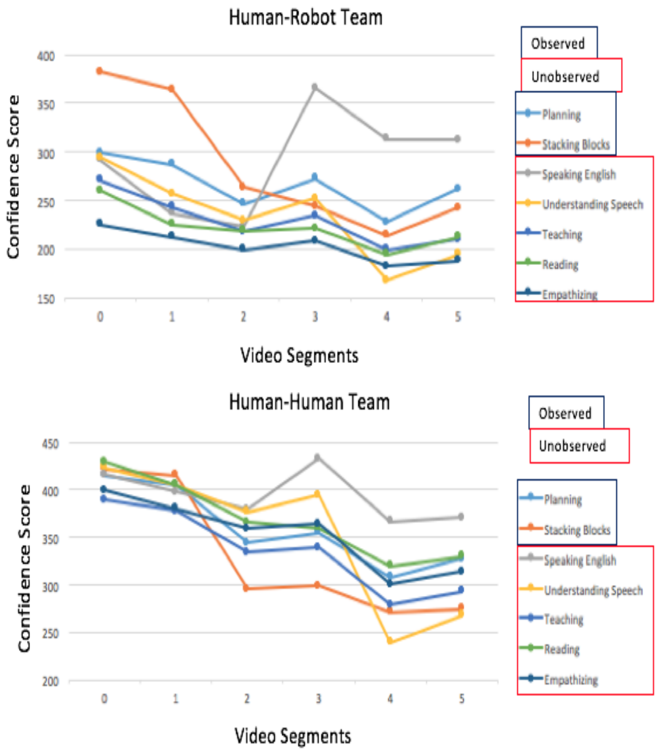

Fig. 3: Mean scores of participants' confidence levels in the featured teammate's ability to complete the seven tasks. Confidence scores were recorded for each video segment[1-5], including the initial still shot[0].

the impact of varying levels of anthropomorphism on people's trust that an agent is capable of performing specific tasks.

Method. We created six surveys that each presented participants with a vignette describing a human worker collaborating on a task with one of our three featured agents in either an industrial setting or a domestic setting. The industrial setting pictured a team in a factory working to install a speaker into a car door and the domestic setting pictured a team cooking dinner in a household. Levels of anthropomorphism were manipulated by displaying a picture of either an industrial robot named "KR-6,", a humanoid robot named "Baxter," or a human named "Fred" at the start of the survey. As dependent variables, we asked participants to rate how much they would trust the featured teammate to accomplish six related tasks using a 5-point Likert scale (1-"Completely Distrust", 5-"Completely Trust"). These tasks all involved social interaction such as handing speakers to a teammate or taking turns. Of the six tasks, three were "observed tasks" that were included in the task description in the survey and three were "unobserved" but related tasks.

Results. To analyze our data, we conducted Single-Factor Analysis of Variance (ANOVA) and Tukey's Post Hoc tests. ANOVA tests comparing scores for Fred, Baxter, and KR-6 in the industrial setting revealed a significant difference between groups, $p<0.01$, for all tasks except for "Hand speaker," $F(2,288)=1.52$, $p>0.05$. Similarly, ANOVA tests for the domestic setting show significant differences between the three groups for all tasks except for "Hand vegetable," $F(2,297)=2.98$, $p>0.05$. From the Tukey's Post Hoc tests, we found significant differences between Baxter and KR-6 in the "Understand request," "Take turns," "Empathize," and "Recognize need" tasks, $p<0.01$, in the domestic setting. However, there were no significant differences for any trait between Baxter and KR-6 in the industrial setting, as shown in Fig. 2. The results indicate that people seem to generalize capabilities for a humanoid robot more than an industrial robot when in a domestic setting. This suggests the importance of designing robot behavior in a way that will be able to mitigate high expectations when interacting with humans in less industrial settings.

III. Study 2: Defying Expectations

After gaining support for the idea that people generate different expectations based on appearance-based preconceived mental models, we wanted to see how behavior can alter these preconceived expectations. We conducted a between subjects survey-study (N = 74) on AMT with type of team partner (robot vs. human) as our independent variable.

Method. For this study we created two video clips of a human-robot team (with Baxter as a humanoid robot partner) and a human-human team each completing a simple block-building task (Fig. 1). The task involved stacking blocks in an alternating color sequence. In both videos, each partner was responsible for one color of blocks. In order to defy preconceived expectations, we programmed Baxter to be incapable of stacking blocks. The human team-mate needed to help it stack blocks, suggesting that the robot's set of capabilities was narrow. The human-human team followed the same script as the human-robot team, including exhibiting the same limitations. The videos showed the interactions in chronological segments with each segment introducing a new limitation or skill. Dependent on the experimental condition, AMT workers were presented with either the series of segments of the human-human, or the human-robot video. To measure people's preconceived expectations based on appearance, we included a still shot of the featured teammate at the beginning of the survey. For the still shot and each consecutive video segment, we asked participants to rate, on a 5-point Likert scale, how well they thought the featured teammate would be able to perform a list of observed and unobserved tasks with 1 being ”Not at all capable” and 5 being ”Extremely capable.”

Results. We took the mean confidence scores for each task and plotted them for each video segment (Fig. 3). For both teams, people's expectations of task completion fluctuated based on each newly demonstrated skill or limitation. In order to measure variance across the six segments for each task, we conducted Single-Factor ANOVAs on the confidence scores. Although both results were significant, the robot-human team displayed greater variance for the observed tasks, ”Speaking English,” $F(5{,}594)=16.7$, $p<0.01$ and ”Stacking blocks” $F(5{,}594)=32.82$, $p<0.01$, compared to the human-human team who had $F(5{,}594)=5.19$, $p<0.01$ and $F(5{,}594)=30.58$, $p<0.01$ respectively. For the rest of the tasks, the human-human team displayed greater variance. This finding suggests that people are more willing to modify their expectations based on a robot's perceived capabilities compared to a human. Furthermore, by the end of the survey, people's expectations of the human dropped for all tasks while expectations for Baxter dropped for all but one of the tasks, “speaking English.” This is presumably because speaking was a skill Baxter exhibited that people did not initially expect. Overall, participants seemed to modify their expectations based on behavioral evidence for both robot and human.

IV. Discussion

Our preliminary findings suggest that (1) people tend to generalize social capabilities more for anthropomorphic robots in more social settings, and (2) we can override preconceived, appearance-based notions of capabilities using behavior. The first study implies that robots designed to work in social settings are more likely to breed an expectations gap, which presents a challenge when designing robots for social settings. The second study suggests that changes in behavior can mitigate these high expectations people have of social robots, thus suggesting the need for new guidelines in interaction design.

V. Challenges in Real-Life Settings

Developing robots that can predict what humans will expect of them is a huge challenge in laboratory settings and even more so in real-life. People differ along many dimensions such as personality or past exposure to robots – all of which inform the expectations people form of a robot's capabilities. Furthermore, as we have shown in our studies, the context in which a robot is employed might play an important role in the perceptions people form about its capabilities.

Since most HRI studies that have explored the influence of a robot's embodiment and behavior on perceptions of capabilities or intelligence have been done in more or less fixed laboratory contexts (e.g. [1], [9]) the question of how a robot can elicit accurate perceptions of its capabilities irrespective of the setting it is employed in is unexplored. In real life settings, robots will be deployed not only in a wide variety of contexts, making it important for robots to adapt to individual differences but also to adapt dynamically to changes in context. Building a robot that can anticipate over-generalizations of its skills in different domains will be difficult. For example, a household robot that helps with cooking and cleaning will need to know a broad range of domain-specific skills in those areas in order to predict over-generalizations.

We expect that there will also be significant methodological challenges when evaluating the expectations gap in real-life settings and over long periods of time. Many robots capable of sophisticated social interaction, like Baxter, do not yet have the social sophistication to be used outside of laboratory or industrial settings. Thus, when testing our model outside the lab, we will need to gather data from robots currently used in everyday social settings, such as Paro or the Roomba, and apply these lessons to other robots. Differences among these robots may complicate generalization.

VI. Conclusion and Future Work

In this paper, we have presented preliminary evidence to support the hypothesis of an expectations gap. Specifically, we showed experiments supporting the notions that (1) humans construct distinct theory of mind models of machines and people, (2) people attribute more human mental models to more social robots, and (3) mental models can be changed by the robot's behavior. In addition to documenting this phenomenon in crowd-sourced studies, we hope to demonstrate its effect in more personal human-robot interaction settings in the laboratory.

In the remaining paragraphs, we lay out important problems that must be addressed in order to solve the problem of the expectations gap. Their solutions will require a combination of mathematical modeling and machine learning.

An important related question is how perceptions of capabilities transfer within a mental model of a single agent based on individual observations. For example, if we hear a robot speaking English, then we expect it to understand English as well, even though from an engineering standpoint these two implementations are unrelated. We need a quantitative, semantic metric on capabilities in order to estimate a human's perceived likelihood of certain capabilities based on generalizations of similar, observed traits. Our focus for ongoing work is on how such a metric can be designed and evaluated. We can then revisit the questions raised in this paper about how and when human mental models generalize.

In the longer term, we plan to build an algorithm to predict when the human will incorrectly estimate a robot's capabilities. Robots could then reduce the expectations gap by issuing corrective behavior that sets realistic expectations. Even having these models in place, the inference problem will be a great challenge. Humans rarely state perceptions or assumptions about capabilities directly, since they typically take human capabilities for granted. Instead, the robot will need to employ a learned model of human behavioral cues to infer capabilities. Our existing crowd-sourcing tools will be valuable in collecting the training data for this model. Finally, we may construct a second learned model mapping robot behaviors onto changes in human perception. At this point, the *inverse semantics* [6] technique will implement an effective controller to adapt the robot's behavior to fine tune human perceptions of the robot's capabilities.

References

[1] C. Bartneck, T. Kanda, O. Mubin, and A. 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] K. Dautenhahn. Design spaces and niche spaces of believable social robots. In *Robot and Human Interactive Communication, 2002. Proceedings. 11th IEEE International Workshop on*, pages 192–197. IEEE, 2002.

[3] K. T. Dirks. The effects of interpersonal trust on work group performance. *Journal of applied psychology*, 84(3):445, 1999.

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

[5] F. Heider and M. Simmel. An experimental study of apparent behavior. *The American Journal of Psychology*, pages 243–259, 1944.

[6] R. A. Knepper, S. Tellex, A. Li, N. Roy, and D. Rus. Recovering from failure by asking for help. 39(3):347–362, 2015.

[7] S. Lemaignan, J. Fink, and P. Dillenbourg. The dynamics of anthropomorphism in robotics. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 226–227. ACM, 2014.

[8] R. C. Mayer, J. H. Davis, and F. D. Schoorman. An integrative model of organizational trust. *Academy of management review*, 20(3):709–734, 1995.

[9] L. Takayama, D. Dooley, and W. Ju. Expressing thought: improving robot readability with animation principles. In *Proceedings of the 6th international conference on Human-robot interaction*, pages 69–76. ACM, 2011.