People's Judgments of Human and Robot Behaviors

A Robust Set of Behaviors and Some Discrepancies

Maartje M.A. de Graaf
Brown University
maartje_de_graaf@brown.edu

Bertram F. Malle
Brown University
bertram_malle@brown.edu

ABSTRACT

The emergence of robots in everyday life raises the question of how people explain the behavior of robots—in particular, whether they explain robot behavior the same way as they explain human behavior. However, before we can examine whether people's explanations differ for human and robot agents, we need to establish whether people judge basic properties of behavior similarly regardless of whether the behavior is performed by a human or a robot. We asked 239 participants to rate 78 behaviors on the properties of intentionality, surprisingness, and desirability. While establishing a pool of robust stimulus behaviors (whose properties are judged similarly for human and robot), we detected several behaviors that elicited markedly discrepant judgments for humans and robots. Such discrepancies may result from norms and stereotypes people apply to humans but not robots, and they may present challenges for human-robot interactions.

KEYWORDS

Behavior Explanations, Folk Psychology, Human-Robot Interaction, Social Cognition

ACM Reference Format:
Maartje M.A. de Graaf and Bertram F. Malle. 2018. People's Judgments of Human and Robot Behaviors: A Robust Set of Behaviors and Some Discrepancies. In *HRI '18 Companion: 2018 ACM/IEEE International Conference on Human-Robot Interaction Companion, March 5-8, 2018, Chicago, IL, USA.* ACM, New York, NY, USA, Article 4, 2 pages. https://doi.org/10.1145/3173386.3177051

1 INTRODUCTION

The rise of robots in everyday life demands an investigation into how people conceptualize robots and their social behaviors—in particular, whether people interpret robot behaviors by way of mental states such as beliefs, desires, and intentions, just as they do for humans [5]. Perhaps the most fundamental use of such mental state inferences lies in people's explanations of other agents' behavior [4]. Examining people's explanations of robot behavior may therefore unveil some of the concepts and cognitive processes that robots elicit in human perceivers, which will in turn help clarify to what extent robotic agents are regarded as social beings.

A small number of studies have provided initial insights into people's explanations of behavior performed by robotic agents (e.g., [9–11]). However, most of these studies did not incorporate the existing bodies of research in philosophy, psychology, and cognitive science on how people generate, select, evaluate, and communicate explanations [7]. Most notably, none of these studies ensured that the behaviors humans and robots performed were equated for some basic properties that are known to influence explanations. In particular, explanations vary dramatically as a function of intentionality, surprisingness, and desirability [1, 6]. Therefore, to determine whether people genuinely *explain* robot and human behaviors differently, we must examine behaviors that are equated, across human and robot, for at least these three properties. Otherwise, any seeming differences in how people explain robot and human behaviors may in reality be due to differences in how people perceive the behaviors (e.g., as more intentional or less surprising) when performed by a robot or human. Both differences are of potential interest, but their theoretical and practical implications differ.

In our investigation we identified a pool of behaviors that people judged as similar on the properties of intentionality, surprisingness, and desirability, regardless of whether they were performed by humans or robots. However, we also detected behaviors that showed markedly discrepant judgments for humans and robots on two or more of the above properties. These behaviors may reveal insights about boundaries of interactions between humans and robotic agents. Both robust and discrepant stimulus behaviors can be found at http://research.clps.brown.edu/SocCogSci/RobotBehaviors.pdf.

2 METHOD

We identified candidate behaviors from the robotics and HRI literatures and from previous studies on human behavior explanations. We aimed for sufficient representation in three classes of behaviors: unsurprising intentional (n = 14), surprising intentional (n = 28), and unintentional (n = 10). Many of these behaviors will be performed only by future robots, so we also identified a fourth class of control behaviors that current robots already perform (n = 26). We recruited 239 participants from Amazon Mechanical Turk and asked them to judge one half of each behavior class (39 out of 78 total) for one agent type (human or robot) on one of the behavior properties (intentionality, surprisingness, or desirability). We examined inter-rater reliability among participants who rated a given agent on a given property across behaviors. We excluded judges (n= 30, 12.6% of all judges) with very low correlations with the rest of the group (r < .30) from further analyses. The remaining judges displayed intra-class correlation coefficients ICC(2,1) in the .50s and .60s for desirability and intentionality, for both agents. More judges had to be excluded for surprisingness and, even then, reliability was in the .30s for robots and .40s for humans.

HRI '18 Companion, March 5-8, 2018, Chicago, IL, USA

ACM ISBN 978-1-4503-5615-2/18/03…$15.00
https://doi.org/10.1145/3173386.3177051

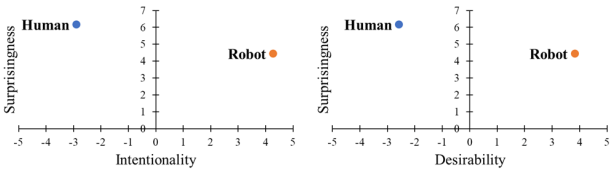

Figure 1: Sample behavior (IN-S-16) judged as intentional and desirable when performed by a robot but as unintentional and undesirable, and even more surprising, when performed by a human

3 RESULTS

We computed the average ratings of intentionality, surprisingness, and desirability for each of the 78 behaviors, separately for robot and human agent. To examine whether the properties differed between agents, we performed three ANOVAs in a 2 (agent: human vs. robot) by 4 (behavior class: control, intentional-surprising, intentional-unsurprising, unintentional) design. We observed the expected main effects for behavior class (e.g., intentional behaviors judged as more intentional than unintentional behaviors) but no interactions with agent. Behavior class explained 62% of the variability in intentionality ($F(3, 147) = 81.0$, $p < .001$), 45% in surprisingness ($F(3, 148) = 40.9$, $p < .001$), and 45% in desirability ($F(3, 148) = 40.3$, $p < .001$). Across the four behavior classes, we identified 28 robust behaviors that were sufficiently similar between the two agent types on all three properties (i.e., no significant agent differences below $p < .001$, nor effect sizes above Cohen's $d > .50$).

However, 17 of the 78 behaviors had significant ($p < .001$) and substantial ($d > .50$) human-robot discrepancies on at least two properties. Three of these behaviors even showed substantial discrepancies for all three properties. The first such behavior was the following (IN-S-16): *"A security [officer | robot] is walking on the sidewalk. When [she | it] sees a fleeing pick-pocket, [she | it] steps in front of him and grabs the man's arm."* When performed by a robot (compared to a human), this behavior was evaluated as clearly intentional (rather than moderately unintentional), middling in surprisingness (rather than clearly surprising), and clearly desirable (rather than moderately undesirable). We visualize the discrepancy of this result in Figure 1 as an example for all the discrepant behaviors. The second behavior was the following (IN-S-23): *"An [robot] assistant is managing the financial information of [his | its] supervisor. [He | It] releases information of [his | its] supervisor's current income to an advertiser."* When performed by a robot (compared to a human), this behavior was evaluated as moderately unintentional (rather than moderately intentional), clearly surprising (rather than moderately surprising), and clearly undesirable (rather than moderately undesirable). The third behavior was the following (IN-S-25): *"A [robot] nurse is treating a woman diagnosed with colon cancer. [He | It] advises a specific medical treatment shown to be successful in men."* When performed by a robot (compared to a human), this behavior was evaluated as middling in intentionality (rather than clearly intentional), middling in surprisingness (rather than clearly unsurprising), and moderately undesirable (rather than moderately desirable).

4 DISCUSSION AND CONCLUSION

For the purpose of investigating how people explain the behavior of robots, we have developed a robust pool of stimulus behaviors that are equated between robot and human agents for three key properties of behavior: intentionality, surprisingness, and desirability. However, we also discovered some behaviors whose properties differed starkly as a function of whether a robot or human performed the behavior. One source of these contrasting perceptions may be different norms that people impose on humans and robots for those behaviors. Previous research has shown that social norms influence anticipated human-robot interactions [3] and expectations of appropriate application domains for robots in society [2]. This impact of norms may help explain why certain robot behaviors were rejected (i.e., rated as less desirable and more surprising) in select domains (such as health care in case of IN-S-25). Another source of the contrasting perceptions may be specific stereotypes people have about human agents, which they (currently) do not apply to robotic agents. For example, people judged the female security officer's intervention in IN-S-16 as unintentional, surprising, and even undesirable. Apparently, her firm behavior violated gender stereotypes, and such violations often lead to devaluation and rejection [8].

Many factors impinge on people's interpretations of behavior, including norms, stereotypes, and inferred intentionality. We cannot expect all of those factors to be equal between humans and robots, but it is remarkable that we have been able to identify a considerable number of behaviors that are. For those behaviors, we can now study in detail how people's explanations differ when the behaviors are performed by humans or robots. For the other behaviors, we have a new task before us: to investigate why some behaviors are seen as so different when performed by human or robot. Shedding light on those differences may reveal challenges for some future human interactions with robots.

REFERENCES

[1] Gifford W. Bradley. 1978. Self-serving biases in the attribution process: A reexamination of the fact or fiction question. *Journal of Personality and Social Psychology* 36, 1 (1978), 56–71.
[2] Maartje M.A. de Graaf and Somaya Ben Allouch. 2016. Anticipating our future robot society: The evaluation of future robot applications from a user's perspective. In *RO-MAN 2016*. IEEE, 755–762.
[3] Maartje M.A. de Graaf, Somaya Ben Allouch, and Jan A.G.M. van Dijk. 2017. Why Would I Use This in My Home?: A Model of Domestic Social Robot Acceptance. *Human-Computer Interaction* accepted (2017).
[4] Bertram F Malle. 2004. How the mind explains behavior. *Folk Explanation, Meaning and Social Interaction. Massachusetts: MIT-Press* (2004).
[5] Bertram F. Malle and Sara D. Hodges (Eds.). 2005. *Other minds: How humans bridge the divide between self and others.* Guilford Press, New York, NY.
[6] Bertram F. Malle and Joshua Knobe. 1997. Which behaviors do people explain? A basic actor-observer asymmetry. *Journal of Personality and Social Psychology* 72, 2 (1997), 288–304.
[7] Tim Miller. 2017. Explanation in artificial intelligence: Insights from the social sciences. *arXiv preprint: 1706.07269* (2017).
[8] Charles Stangor, Linda A Sullivan, and T.E. Ford. 1991. Affective and cognitive determinants of prejudice. *Social Cognition* 9, 4 (1991), 359–380.
[9] Sam Thellman, Annika Silvervarg, and Tom Ziemke. 2017. Folk-psychological interpretation of human vs. humanoid robot behavior: exploring the intentional stance toward robots. *Frontiers in psychology* 8 (2017), 1962.
[10] Ning Wang, David V Pynadath, and Susan G Hill. 2016. Trust calibration within a human-robot team: Comparing automatically generated explanations. In *HRI 2016*. IEEE Press, 109–116.
[11] Robert H. Wortham, Andreas Theodorou, and J. J. Bryson. 2016. What does the robot think? Transparency as a fundamental design requirement for intelligent systems. In *IJCAI 2016 Workshop on Ethics for AI*. AAAI, New York, NY, USA.