People's Explanations of Robot Behavior Subtly Reveal Mental State Inferences

Maartje M.A. de Graaf
Dept. of Information & Computing Sciences
Utrecht University
Utrecht, The Netherlands
m.m.a.degraaf@uu.nl

Bertram F. Malle
Dept. of Cognitive, Linguistic, & Psychological Sciences
Brown University
Providence, USA
bertram_malle@brown.edu

***Abstract*—It has long been assumed that when people observe robots they intuitively ascribe mind and intentionality to them, just as they do to humans. However, much of this evidence relies on experimenter-provided questions or self-reported judgments. We propose a new way of investigating people's mental state ascriptions to robots by carefully studying explanations of robot behavior. Since people's explanations of human behavior are deeply grounded in assumptions of mind and intentional agency, explanations of robot behavior can reveal whether such assumptions similarly apply to robots. We designed stimulus behaviors that were representative of a variety of robots in diverse contexts and ensured that people saw the behaviors as equally intentional, desirable, and surprising across both human and robot agents. We provided 121 participants with verbal descriptions of these behaviors and asked them to explain in their own words why the agent (human or robot) had performed them. To systematically analyze the verbal data, we used a theoretically grounded classification method to identify core explanation types. We found that people use the same conceptual toolbox of behavior explanations for both human and robot agents, robustly indicating inferences of intentionality and mind. But people applied specific explanatory tools at somewhat different rates and in somewhat different ways for robots, revealing specific expectations people hold when explaining robot behaviors.**

Keywords—human-robot-interaction, behavior explanation, mental state inference, folk psychology, theory of mind

I. INTRODUCTION

Many emerging applications of robotic technology involve humans and robots collaborating in healthcare, education, and domestic work. In these contexts, robots display increasingly sophisticated indicators of autonomous action (e.g., approaching, grasping, asking questions) and of mental activity (e.g., eye movements for attention, facial expressions for emotion). Psychological research has shown that people cannot help but infer mind and intentionality from such indicators [1], [2]. People's perceptions and mental models of a machine collaborator are therefore likely to contain deep-seated inferences of mind and agency, which in turn direct human-robot interactions [3]–[7]. In particular, people construct models of other agents to understand and predict their actions and to reduce uncertainty when interacting with them [8].

This research was supported by a Rubicon grant 446-16-007 from the Netherlands Organization for Scientific Research (NWO) and by a grant from the Office of Naval Research (ONR), N00014-16-1-2278. Opinions expressed here are our own and do not necessarily reflect the views of NWO or ONR.

Perhaps the core guide toward understanding and prediction of others' behavior is the human tendency to *explain* those behaviors [9]–[12]. Behavior explanations identify the causes and meaning of behavior, direct people's perceptions and evaluations of others, and regulate their contributions to social interactions. Work in numerous disciplines has shown that human behavior explanations are fundamentally grounded in a conceptual framework of intentional agency and mind, typically referred to as "theory of mind" [13], [14] or "folk psychology" [15], [16]. This framework guides both people's explanations and predictions of behavior but also inferences about specific mental states, such as beliefs, desires, intentions, and emotions [17].

Initial evidence suggests that people may apply this framework of folk psychology to robots, ascribing mental states to them not unlike they do to humans. When observing a robot's behavior, people try to grasp a robot's mind by assigning humanlike traits or characteristics [3], [18] and by using or accepting mentalistic language to make sense of the robot's behavior [19], [20]. Mentalistic inferences are also suggested by the activation of particular brain areas associated with mental model processing [5], [20] and in people's gaze behavior [6]. Mental state ascriptions to artificial agents influence how we differentiate humans from such agents [19], whether we have empathy for robots and treat them well [21], and how we think about robot rights [22]. Thus, people's conceptualization of robots as having a theory of mind may have profound implications for our interactions with these artificial collaborators.

To document and understand people's ascriptions of mind to robots, previous research has mainly relied on direct questions [23]–[27], predictions of certain behaviors [18], [28]–[30], and judgments of robots' mental capacities [31], [32]. Such methods, however, are subject to a number of limitations: They are highly sensitive to the robot's apparent function [33], the nature of the question (forced-choice vs. open-ended; [34]), and the robot's specific physical appearance [35]. Furthermore, the measured mental state ascriptions are often course-grained (e.g., whether the robot generally has emotions or can reason logically [31], [36]), not fine-grained (whether the robot has a specific belief or desire). Finally, most methods suffer from experimenter demand and therefore do not assess people's spontaneous mental state attributions. Attempts to assess such attributions in subtler ways are difficult without a broader theory and can appear ad-hoc (e.g., active vs. passive voice verb use [37]). New approaches using neuroimaging

appear promising, but inconsistent findings have been reported on the similarities between mental state inferences from human and robot behavior (see [20] for a review). Moreover, although neuroimaging techniques overcome some limitations of self-reported perceptions of robots' minds, they still are not sufficiently fine-grained as they document only an overall activation of regions involved in mind inference rather than inferences of specific mental states. Significantly, none of the methods used so far show whether people apply the broader conceptual framework of mind and action (their folk psychology) to robots.

We propose a new way of investigating people's mental state ascriptions to robots by carefully studying people's explanations of robot behavior. Given the grounding of explanations in assumptions of mind and intentional agency, the study of behavior explanations can reveal specific concepts, processes, and even linguistic structures associated with mind inference. Explanations emerge in development just around the time children's theory of mind is maturing [38], and they quickly take on characteristic linguistic forms [39]. Because they are often expressed verbally, they are directly observable with scientific methods [40]. In particular, they arise naturally in answers to simple why questions, so experimentally eliciting them is easy [41], [42]. Most important, they involve fine-grained ascriptions of mental states that can be reliably classified with content analysis [12], [43].

Behavior explanations are of course not meant to replace but rather complement other methods of illuminating people's inferences about robot minds. The present research introduces explanations as a productive and unobtrusive technique to uncover novel insights into people's readiness to infer mental states from robot behavior. This approach allows us to identify both similarities (e.g., using the same conceptual toolbox of folk psychology) and differences (e.g., contrasting rates of specific explanatory tools) in how people make sense of robot behavior. Another significant payoff in studying behavior explanations is that they can inform the development of explainable, intelligible robot behavior. The recent call for transparency and explainability of intelligent machines [44] can succeed only if we know what forms of explanations people actually find helpful and satisfying [45]. By identifying the forms of explanations people naturally offer for robot behavior we can build an initial blueprint for how *robots themselves* could explain their own behavior.

II. Theoretical Framework

A. Behavior Explanations in Robot Studies

Little research exists on how people explain robot behavior and how these explanations compare to corresponding explanations of human behavior [7]. A small number of studies qualitatively explored people's accuracy of mental models of an artificial system [26], [27], [46] or established that mental state inferences occur, without exploring in detail the content and function of these inferences in the overall explanations [25], [47].

A few studies set out to investigate the form of explanations people give of robot behavior but for various reasons did not systematically report the results. Wortham et al. [27] showed a video of a human-robot interaction and recorded participants' explanations of the robot's behavior. However, the researchers concluded that participants' explanations were ambiguous and they discarded the data. Another study described a detailed coding system of people's explanations of robot behavior but reported only judgments of intentionality [47]. In a third study, people interacted with a robotic ottoman [25] and were asked to explain the ottoman's behavior. Unfortunately, the researchers did not systematically analyze these explanations but offered only sample quotes to support their conclusion that most participants considered the ottoman's behavior intentional and ascribed mental states to it. All these studies raised the novel question of robot behavior explanations but did not yet provide direct evidence for how people explain robot behavior.

The most detailed study of people's explanations of robot behavior was recently presented by Thellman et al. [7]. In the study, participants viewed pictures of a human or a robot (Pepper) perform several behaviors in a kitchen and read an accompanying description of the behavior (e.g., "Ellis mops the floor"). The study protocol did not ask participants to explain in their own words what happened but to rate the plausibility of seven prepared causal explanations (e.g., goal, disposition, outcome) for each behavior. The researchers found these ratings of plausibility to be largely indistinguishable between robot and human actors and speculated that people's folk concepts of intentional behavior may also be similar for human and robot agents.

In the present study we directly examine people's spontaneous explanations of robot behavior and analyze the conceptual assumptions of mind and intentionality revealed by these explanations. To do so we introduce a systematic theoretical framework that captures the fundamental concepts and assumptions people bring to bear on their explanations of behavior, for humans and potentially for robot agents. Within this framework we can methodically analyze people's own explanations, delineate the specific explanatory tools they deploy, and identify intentionality and mental state inferences contained in these explanations.

B. A Folk-Conceptual Theory of Explanation

When explaining behavior, people distinguish sharply between unintentional and intentional behaviors [10], [48]. They explain unintentional behaviors by referring to a variety of causes (e.g., emotions, traits, other people, physical events) that are thought to directly bring about the unintentional behavior [49]. To explain intentional actions, however, people use a sophisticated set of distinctions that follow from their folk concept of intentionality [50] and its constituent mental states: primarily a desire for an outcome and beliefs about how a particular action leads to the outcome. To explain why an agent performed a given intentional action, people thus try to infer what specific desire and specific beliefs prompted the agent to perform the action. These are the *reasons* for which the agent acted. "What was her reason for ordering a second espresso?"—"She wanted to stay alert while preparing her lecture." "What was his reason for suddenly being so nice?"—"He thought he was speaking with an influential politician."

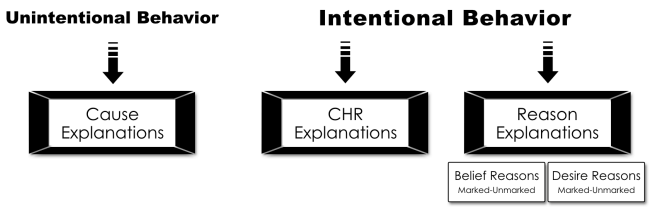

Fig. 1. Folk-conceptual theory of behavior explanations.

Thus, when people conceptualize behaviors as intentional they often explain them by citing the agent's mental states (primarily beliefs and desires), thereby providing *reason explanations*. Sometimes, however, people are at a loss for an explanation involving the agent's reasons, or the reasons are uninteresting; in that case people retreat to "causal history of reason" (CHR) explanations [48] (see Fig. 1). CHR explanations still explain intentional behavior, but they step outside the agent's own reasoning and refer instead to more general background facts—for example, that the agent is a member of a certain culture, has a certain personality, or found themselves nudged by a certain context.

The key contrast, then, is that reasons cite the specific mental states on the agent's mind before performing the action, whereas CHRs cite background factors that may have led up to those reasons—factors literally in the causal history of those reasons. Consider a person who decided not to vote in the last presidential election. We might explain this action by citing a reason ("they didn't think their vote mattered") or by citing a background factor ("their whole family is apolitical"). The family being apolitical is not something the agent actively had on their mind, but it does help explain the decision not to vote, as a "causal history" explanation.

Once people offer a reason explanation, two further choices arise [48]. The first is what type of reason to provide (typically either beliefs, desires, or both). Belief reasons refer to the agent's thinking and knowledge, drawing attention to the agent's rational, deliberative side, whereas desire reasons highlight what the agent wants or needs [51]. Belief reasons, more than desire reasons, also provide idiosyncratic details about the agent's decision-making process, including rejected options, specific plans of action, and likely consequences. As a result, belief reasons are generally less frequent when explaining the behavior of strangers [52].

The second choice surrounding reasons arises for the many cases in which explanations are verbalized: whether to linguistically mark a belief reason or desire reason with a mental state verb (such as "they wanted" for a desire or "they thought" for a belief). This choice is optional but not trivial. Citing or omitting mental state markers can serve significant social functions, such as to distance oneself from the agent's beliefs or to highlight the agent's rationality [51], [52].

This theoretical framework (see Fig. 1) relies on distinctions that derive from an analysis of folk psychology and the components of people's intentionality concept. Moreover, the framework has been empirically supported in numerous studies, showing that it makes correct predictions about how people differently explain their own and others' behavior [52] as well as individual and groups' behaviors [53], and about how people use explanations to manage their own and other people's social standing [51], [52]. These studies relied on a detailed content-coding system [43] that examines people's own verbal explanations and identifies reliably what explanation mode they chose (e.g., cause, reason, CHR), what specific reason they chose (belief, vs. desire), and whether the reason was marked with a mental state verb. Thus, applying this framework to the question of how people explain robot behavior, we have a solid theoretical foundation for differentiating forms of explanation that directly reflect folk psychological concepts. We also have numerous comparison data from the way people explain human behaviors. And we have a methodological tool that makes naturally occurring explanations measurable within this detailed and established conceptual space.

C. Methodological Commitments

In addition to this theoretical foundation, we adopted a number of methodological commitments. First, none of the previous studies on robot behavior explanations ensured that the behaviors people explained were equated, across human and robot agents, for basic properties that are known to influence explanations. In particular, explanations vary dramatically as a function of intentionality, surprisingness, and desirability [41], [54], [55]. Therefore, to determine whether people genuinely *explain* robot and human behaviors differently, we must examine behaviors that have stable intentionality, surprisingness, and desirability, whether performed by a robot or human agent. 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). Second, to avoid prompting participants to make mental state inferences or encouraging them to use certain kinds of explanations, we asked them in a free-response format to explain why a given agent performed a given behavior. Resulting explanations that contain mental state references are then indicative of genuine and spontaneous ascriptions of mind to robots.

D. Working Hypotheses

The general hypothesis is that explanations of robot and human behavior will be similar in the kinds of concepts and linguistic tools people employ but that unique features of robot agents will elicit some reliable differences. Because tightly derived predictions are not possible in the absence of a detailed theory of "robot social perception," we offer working hypotheses on such differences grounded in the folk-conceptual theory of behavior explanation and past findings that employed it. We focus on the three major features of intentional behavior explanation.

Hypothesis 1. For the contrast between *reason explanations* and *causal history of reason (CHR) explanations*, previous findings suggest that reason explanations from an observer perspective are more prevalent for (a) familiar (rather than unfamiliar) agents, (b) agents whose impression one tries to manage, and (c) individual agents (compared to collective agents, whose minds are less clearly delineated [52], [53]). Robots are less familiar than any human agent; people will

generally be less motivated to manage a robot's impression; and robots' minds are less clearly delineated (at least for lay persons). Taken together we expect robot agents to elicit fewer reason explanations (relative to CHRs) than do human agents.

Hypothesis 2. For the contrast between *belief reasons* and *desire reasons*, previous findings suggest that, from the observer perspective, belief reasons are generally more difficult to infer and more prevalent for agents whose impression one tries to manage, especially when highlighting the rationality of an agent [51], [56]. This leads to opposing hypotheses: On the one hand, inferring robots' unfamiliar beliefs may be even more difficult than inferring humans' beliefs, leading to a lower rate of belief reasons relative to desire reasons. On the other hand, assumptions of a robot's rationality would increase the use of belief reasons for robots. In line with the latter hypothesis, the experimental philosophy literature suggests that people are more inclined to attribute beliefs and knowledge to robots than to attribute affect and desires [57]. If true, this tendency would also favor a greater relative use of beliefs over desires when explaining robot behaviors.

Hypothesis 3. For the contrast between *marked* and *unmarked* reason explanations, we need to distinguish between belief markers and desire markers. By using belief markers ("she thinks"; "he believes"), explainers can distance themselves from the agent [56], emphasizing that the agent had a belief that the explainer did not necessarily share. Compare: "They told the joke because it's funny" vs. "They told the joke because they think it's funny." In addition, explainer use belief markers to highlight agents' rationality and reasoning [58]. Thus, to the extent that people see robot minds as distant from their own and/or want to highlight the robot's rationality, people are likely to use more belief markers when explaining robot intentional behaviors. By contrast to belief markers, desire markers ("wanted"; "needed") are conversationally less impactful, because at least in English the grammar of desire reasons reveals the agent's desire state even without mental state verbs (through phrases such as "in order to" or "so that"). However, if people are indeed reluctant to ascribe desires to robots, then they may be particularly reluctant to mark those desires with verbs of "wanting," "needing," and the like.

TABLE I. BEHAVIOR STIMULI WITH MEANS AND STANDARD DEVIATIONS FOR HUMAN AND ROBOT AGENTS

		Intentionality ***M* (*SD*)**		**Surprisingness** ***M* (*SD*)**		**Desirability** ***M* (*SD*)**	
Behavior Class	**Behavior Text**	**Human**	**Robot**	**Human**	**Robot**	**Human**	**Robot**
Surprising Intentional	A [*robot*] nurse is taking care of an ill young boy in a local hospital. [*She / It*] brings him a big present.	4.7 (0.7)	3.6 (1.8)	3.8 (1.6)	4.1 (2.4)	4.1 (1.8)	3.6 (1.3)
	A security [*officer / robot*] is walking on the sidewalk. When [*he / it*] sees a fleeing pick-pocketer, [*he / it*] steps in front of the thief and grabs his arm.	4.7 (0.6)	4.7 (0.7)	3.9 (1.9)	3.1 (1.7)	3.5 (1.4)	3.8 (1.8)
	A [*robot*] assistant is working on [*his / its*] supervisor's computer. [*He / It*] searches through the directory for the supervisor's private files and reads them all.	4.5 (1.0)	2.8 (3.4)	4.9 (1.7)	6.4 (0.8)	-3.9 (1.5)	-3.9 (1.3)
	A personal assistant [*robot*] is sorting through a stack of files. When the managing director asks to get him some lunch, [*she / it*] responds by saying, "Not now, please."	3.1 (2.3)	2.1 (3.1)	4.8 (1.5)	5.3 (2.2)	-1.0 (1.6)	-1.8 (2.5)
Unsurprising Intentional	A [*robot*] host is standing at the entrance of the restaurant. [*He / It*] greets two incoming guests and immediately guides them to a table.	4.8 (0.5)	4.6 (0.7)	1.7 (1.4)	2.2 (1.6)	3.1 (1.8)	3.6 (1.7)
	A [robot] technician is about to replace the hard drive of a customer's computer. [*She / It*] transfers all the files to a backup drive.	4.7 (0.7)	4.6 (0.8)	1.5 (1.2)	2.2 (1.8)	3.5 (1.5)	3.5 (1.8)
	A [*robot*] assistant is helping [*his / its*] supervisor in preparing an important presentation. Just moments before the meeting, [*he / it*] emails the final version of the presentation slides.	3.4 (2.1)	3.1 (2.7)	2.7 (1.8)	2.8 (1.9)	0.8 (2.3)	1.7 (2.7)
	A [*robot*] tutor is grading final exams. [*He / It*] gives a student an A, which makes her pass the semester.	2.2 (2.7)	3.8 (1.7)	2.4 (1.9)	1.9 (1.6)	2.0 (2.0)	2.9 (2.0)
Current Intentional	A [*robot*] soldier is exploring a building after a nuclear explosion. [*He / It*] tells [*his /its*] teammates not to enter the boiling room without protective gear.	4.5 (1.0)	4.9 (0.4)	1.7 (1.2)	1.9 (1.3)	4.2 (1.1)	3.6 (2.8)
	A [*robot*] host is meeting a woman in the lobby. [*He / It*] looks at the woman, performing a beckoning gesture with [*his / its*] right hand.	4.5 (0.9)	2.9 (3.3)	2.7 (1.7)	2.7 (1.9)	0.8 (1.6)	0.0 (2.5)
Unintentional	A [*woman / robot*] is opening the door to enter the apartment building. [*She / It*] knocks out a fleeing burglar who was arrested shortly thereafter.	-3.6 (2.2)	-2.6 (3.3)	5.8 (1.4)	5.1 (2.0)	1.9 (1.9)	1.6 (2.1)
	A [*robot*] technician fails to tightly close a valve at the waste plant. Poisonous gas is released into a neighboring building.	-3.6 (1.6)	-4.1 (1.8)	5.2 (1.7)	4.9 (2.0)	-4.6 (1.2)	-4.5 (1.0)
	A [*robot*] custodian is mopping the floor in the entrance hall of a bank. [*Her / Its*] wet floor makes a bank robber slip, hindering him from exiting the building before the police arrive at the scene.	-3.1 (2.7)	-2.7 (3.5)	5.2 (1.3)	3.7 (2.5)	1.9 (2.1)	2.1 (2.7)
	A [*robot*] cook is baking a cake. [*He / It*] takes the cake out of the oven, moves backwards to set it down, and steps on a co-worker's toes.	-4.1 (2.0)	-4.3 (1.6)	2.7 (1.5)	3.1 (1.9)	-1.0 (0.7)	-1.8 (1.5)

III. METHOD

A. Pretests: A Pool of Matched Stimulus Behaviors

We designed text descriptions of robot and human behaviors for our study, as they allowed us to introduce robots with a variety of functions and roles, left robot appearance unspecified, and used a natural means by which people learn about behavior in general (e.g., in the news, in emails, or conversations). However, as mentioned earlier, in order to study any potential differences between explanations of human and robot behavior, we must ensure that people perceive the explained behaviors in the same way, whether performed by a human or robot agent. It is not enough to simply use the same phrasing in a behavior description and vary the subject noun as "robot" vs. "person." The behaviors must be perceived as highly similar on key properties. As a result, we conducted two initial studies to establish a robust pool of stimulus behaviors that are equated across robot and human agents for three key properties of behavior: intentionality, surprisingness, and social desirability—all of which are critical factors that influence behavior explanations [41], [54], [55].

For Study 1 we identified candidate behaviors from the robotics and HRI literatures and from previous studies on human behavior explanations. We focused on likely-intentional behaviors, both surprising and unsurprising ones, and a few likely-unintentional ones (which are almost always surprising). Several of these behaviors are out of reach for extant robots, so we added a few behaviors that such robots are capable of performing—a class of behaviors we labeled "current." We also attempted to populate each behavior class with socially desirable and undesirable exemplars. We recruited 239 participants from Amazon Mechanical Turk to judge one half of each behavior class (unsurprising intentional, n = 14; surprising intentional, n = 28; current intentional, n = 26; and unintentional, n = 10), completing judgments for one agent type (human or robot) on one of the behavior properties (intentionality, surprisingness, or desirability). More details about the method and analysis can be found in [59]. In light of the results, we selected a pool of 28 matched stimulus behaviors whose properties were judged similarly for human and robot agents. However, the resulting behaviors did not evenly cover all behavior classes (e.g., there were few matched surprising intentional behaviors), so we conducted a second study. We began by rephrasing several stimulus behaviors from Study 1 that were closely matched on some but not all properties. In addition, we created six new unintentional and desirable behaviors, which had proven difficult to match in the first study. The resulting candidate behaviors again represented the four classes of behaviors: unsurprising intentional (n = 7), surprising intentional (n = 11), current intentional (n = 5), and unintentional (n = 11). The procedure and measurements were identical to those in Study 1, with the sole exception that participants in Study 2 (n = 126 from Amazon Mechanical Turk) judged all 35 behaviors for one agent type (human or robot) on one of the three behavior properties (intentionality, surprisingness, or desirability).

In light of the results from both studies, we established a pool of 25 stimulus behaviors that were equated between robot and human agents for all three key properties of behavior: intentionality, surprisingness, and desirability. This pool consisted of four classes: unsurprising-intentional (n = 6), surprising-intentional n = 7), current-intentional (n = 2), and unintentional (n = 10) [59].

B. Study Design and Procedure

For the present study, we selected 14 stimulus behaviors falling into the following behavior classes: 4 surprising intentional, 4 unsurprising intentional, 2 current intentional, and 4 unintentional; for each class, half of the behaviors were desirable, half undesirable. To achieve this balanced set we selected 12 items from the pool of stimulus behaviors described above and added two surprising intentional items (one desirable, one undesirable) from the extended pool that were matched on all properties, albeit not identical in degree. The added desirable one showed a statistically significant difference in surprisingness, but the behavior was clearly judged as very surprising for both human and robot agent. Likewise, the added undesirable one showed a statistically significant difference in intentionality, but the behavior was clearly judged as intentional for both human and robot agent. All 14 selected behavior stimuli are shown in Table 1.

We chose a between-subjects manipulation of the agent factor to avoid making the purpose of our study salient, which a mix of human and robot agents in the same stimulus material would have done. We thus created two versions of an online questionnaire, differing only in the agent descriptors, human or robot, and participants were randomly assigned to either version. After participants gave their consent, we explained the study task: to consider various behaviors and to provide an explanation in their own words for why each agent performed the respective behavior (i.e., "Why did [*he / she / it*] do that?"). In the robot condition, we added that we were interested in the participant's opinions about robot behaviors in the near future. Participants completed one practice item, the 14 experimental items in random order, and ended by providing basic demographics and indicating their knowledge of or experience with robots or the robotics field (using 1-6 point rating scales).

C. Classification and Analysis of Explanations

All explanation classifications followed the systematic guidelines of the F.Ex. coding system [43], as used in previous studies [51]–[53], [60]. To determine an explanation code, coders first decide on whether the explained behavior is likely intentional (such as "greet," "save"); if it is intentional, coders determine whether a reason is mentioned (something the agent likely had on their mind when deciding to act—primarily beliefs and desires) or a factor that was in the causal background of the agent's decision (CHR) but not on their mind (e.g., personality, environment, role).

For example, the explained behavior "greets two guests" would be judged intentional (firmly supported by the pretest data). An explanation for it such as "The robot wanted to be polite" would be a (desire) reason; "its job was to be host" would be a CHR. The main focus of this paper is on such explanations of intentional behaviors. However, we also classified and analyzed all cause explanations of unintentional behavior, which are available in the supplementary materials.)

To prepare the free-response data for analysis, two coders were trained and independently identified codable clauses in all verbal responses (87.1% agreement and Gwet's AC_1 = .85 [61][1]). Coders then classified all explanations into reasons vs. CHRs (90.2% agreement, κ = .79), reason type (91.8%, κ = .84), and mental state markers (93.2%, κ = .79). Beyond the standard F.Ex codes we also created a code for explanations of robot behavior in which participants referred to the robot being programmed (e.g., "It was programmed to read files"; "it did its programmed tasks"). Coding was done via automated text search and human checking (agreement was above 95%). Such program references could in principle be reasons (e.g., "it was programmed to want to be polite") or CHRs ("it is programmed as a host"). In the present data, only a single program reference was a reason explanation ("The robot determined through its programming that the student was deserving of the A"); all other program references for intentional behaviors were causal history (CHR) explanations.

The dependent variables were raw counts of explanation parameters (e.g., number of reason explanations per intentional behavior; number of belief reasons per behavior explained by reasons). These explanation parameters were averaged across all intentional behaviors to gain maximal representativeness of potential human-robot differences (but we briefly report the level of consistency within behavior classes). In addition to the average counts, we report percent scores when illustrating the relative dominance of one explanation type over another (e.g., percent reasons out of reasons + CHRs). We examined all hypotheses using mixed between-within analyses of variance (ANOVA), focusing on the statistical interactions of the agent factor with the relevant within-subject explanation parameter (e.g., reason vs. CHR). We report Cohen's *d* effect sizes[2].

D. Participants

We recruited 121 participants (58 males, 62 females, 1 other) from Amazon Mechanical Turk (human agent condition n = 61, robot agent condition n = 60). The participants' age ranged from 21 to 84 (M = 38.5, SD = 12.3), their educational level ranged from high school degree to doctoral and professional degrees, and they had little knowledge about (M = 2.5, SD = 1.2) and experience with (M = 2.1, SD = 1.2) robots. There were no significant differences in those measures between participants from the two agent conditions.

IV. Results

Analyses were based on a total of 121 participants (61 explained human behaviors, 60 explained the corresponding robot behaviors). Average length of explanations was comparable for robot behaviors (M = 16.7) and human behaviors (M = 14.3), suggesting that participants put similar cognitive effort into their explanations. Overall, participants provided 1.34 explanations per behavior (95% CI: 1.24, 1.44). That number was slightly higher for human agents (M = 1.43, SD = 0.64) than robot agents (M = 1.25, SD = 0.43), $F(1, 119) = 3.44$ p = .066, d = 0.34.

Hypothesis 1 stated that robot agents would elicit fewer reason explanations (relative to CHRs) than would human agents. As Figure 2 shows, this hypothesis was supported, as explainers offered relatively fewer reasons for robots (M = 0.67, SD = 0.37) than for humans (M = 1.01, SD = 0.35) but relatively more CHRs for robots (M = 0.70, SD = 0.39) than for humans (M = 0.50, SD = 0.41), interaction $F(1,119) = 28.26$, p < .001, d = 0.71. This human-robot asymmetry in the use of reasons vs. CHRs held consistently within each behavior class (desirable-undesirable, surprising-unsurprising, and for current behaviors). However, more than half of participants' CHR explanations for robot actions referred to the robot's programming. So we tested whether the reason-CHR asymmetry was driven by such program-referring explanations. Indeed, when we conducted the analysis again, considering only program-*un*related CHRs, the above interaction effect dropped sharply, $F(1,119) = 3.42$, p = .067, d = 0.19. (The effects within behavior classes were also consistently reduced.). Thus, people explain robot actions overall with fewer reasons and more CHRs, but the strong presence of CHR explanations relative to reason explanations (51% for robot actions vs. 33% for human actions) is spurred by over half of the CHRs referring to the robot's programming.

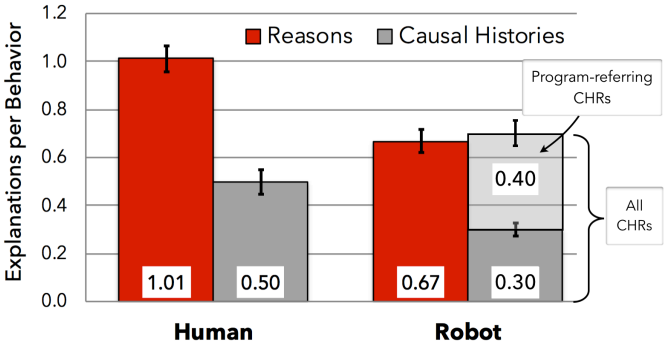

Fig. 2. Explainers of human intentional actions, relative to robot actions, used more reason explanations and fewer causal history of reason (CHR) explanations. More than half of the considerable number of CHR explanations given for robot actions referred to the robot's programming.

Hypothesis 2 concerned the contrast between *belief* reasons and *desire* reasons.[3] As Figure 3 shows, explainers of human action offered more desire reasons (M = 0.77, SD = 0.25) than belief reasons (M = 0.49, SD = 0.29), $F(1,116) = 18.71$, p < .001, d = 1.04; by contrast, explainers of robot actions offered as many desire reasons (M = 0.57, SD = 0.33) as belief reasons (M = 0.58, SD = 0.29), F = 0. This pattern yields a human-robot asymmetry for reason types, *interaction* $F(1,116) = 9.27$, p = .003, d = 0.39, and it is consistent across behavior classes (desirable-undesirable, surprising-unsurprising, and for current behaviors). This asymmetry supports the hypothesis that people are reluctant to attribute affect and desire to robots but feel comfortable attributing beliefs to them [57].

[1] This index has been derived mathematically and supported by Monte Carlo simulations as an improvement over Cohen's κ measure when margin rates are extreme (as is the case for identifying codable clauses).

[2] For interaction effects, proper effect size calculation requires dividing the full interaction contrast by two times the pooled standard deviation [62], $d = ((M_{11} – M_{12}) – (M_{21} – M_{22}))/2\sigma$.

[3] Within reason explanations, people may offer beliefs, desires, or valuings [12], [43]. However, as is often the case, the rate of valuings was negligible (< 2%), so we conducted all reason type analyses on beliefs and desires only.

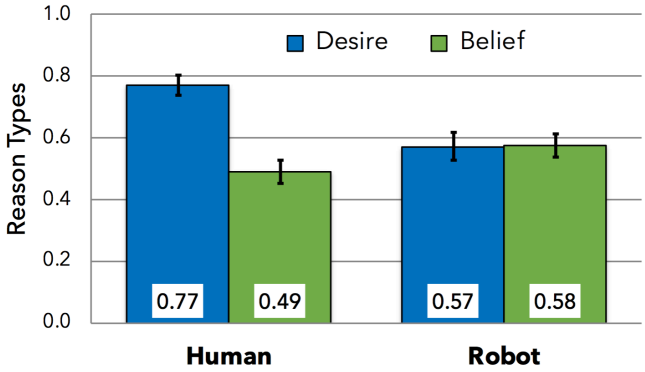

Fig. 3. Explainers of human actions provided more desire reasons than belief reasons, whereas explainers of robot actions provided an equal number of the two reason types.

Hypothesis 3 involved the contrast between marked and unmarked reason explanations as a result of people either distancing themselves from the agent and/or highlighting the agent's rationality. For both types of reasons (i.e., beliefs and desires), we assessed whether the explainer explicitly marked the offered reason with a mental state verb. Generally, people tended to offer reason explanations without markers, but the use of such markers showed two contrasting human-robot asymmetries (see Fig. 4). On the one hand, people offered marked belief reasons more often for robots (31.6% of the time) than for humans (10.4%), $F(1,112) = 22.2$, $p < .001$, $d = 0.88$. On the other hand, they offered marked desire reasons more often for humans (33.6%) than for robots (18.3%), $F(1,107) = 9.2$, $p < .001$, $d = 0.58$. Moreover, the three-way interaction between reason type, marker, and agent type was strong, $F(1,105) = 29.3$, $p < .001$, $d = 0.85$. This pattern suggests that, when explaining robot behaviors with reasons, people are far more comfortable ascribing explicitly marked beliefs to robots than to ascribe explicitly marked desires to them (those that use the words "want," "need," etc.).

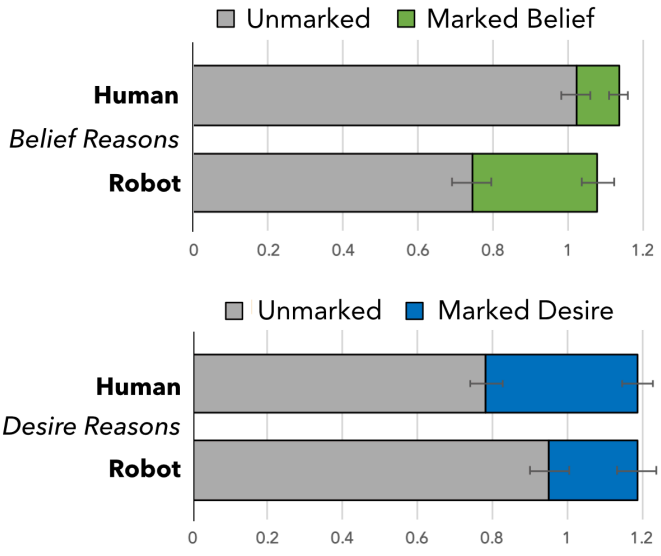

Fig. 4. When offering belief reasons (top), people use more mental state markers for robot agents than for human agents; when offering desire reasons (bottom), people use *fewer* mental state markers for robot agents.

V. Discussion

A growing body of research indicates that many of the processes that guide people's interactions with humans also guide their interactions with robots. In particular, initial studies suggest that people ascribe mental states to robots not unlike they do to humans [20]. However, much of this evidence relies on experimenter-provided questions, self-reported judgments, or general indicators of mind inference. The present research sought to contribute new evidence on people's mind perception of robots through the study of behavior explanations. People naturally provided such explanations when asked simple why questions, and the concepts and linguistic structures they used in explaining robot behavior revealed several patterns of inferences about robot minds and intentional agency.

Overall, people used the same toolbox of explanations for both robots' and humans' intentional actions, namely reason explanations that refer to the agent's own beliefs and desires supporting the given action, and causal history explanations that provide the broader causal background to those reasons. In addition, however, people applied the specific explanatory tools at somewhat different rates and in somewhat different ways for robots. We briefly discuss these patterns in turn.

First, as is standard in ordinary behavior explanations [60], people provided about twice as many reasons as causal history of reason explanations (CHRs) for human agents. In giving reason explanations, people simulate the agent's mental states (primarily beliefs and desires) that led to the decision to act. Accordingly, as half of people's explanations of robots' actions were reasons, people seemed to exhibit spontaneous mental state inferences for robot agents. Further inspection also shows that only 3/60 explainers of robot actions used no reason explanations at all. At the same time, people offered relatively more CHR explanations when explaining robot actions than when explaining human actions. The unique contributor to this asymmetry was a class of CHRs that referred to the robot's program or its programmers. It is noteworthy that such program references never occurred in reason explanations of robot behavior, so the reasons people provided for these behaviors represent uncontaminated evidence for inferences of robot mental states. But in spite of people's general inclination to apply folk psychology to robot behavior, they did not ignore the fact that a robot is a mechanical device and sometimes (in about 30 percent of all intentional behavior explanations) offered simple mechanistic explanations (e.g., that the robot was programmed to greet a guest or designed to protect people).

Second, examining the type of reasons people offered, we observed that desire reasons were more frequent than belief reasons in explanations of human behavior, as is typical for people's explanations of strangers' behaviors [52]. In explanations of robot behaviors we saw a relative increase to use belief reasons and a decrease to use desire reasons. As suggested by previous research [57], people are comfortable considering robots as having beliefs and knowledge, as being rational; but they are less comfortable portraying robots as having needs and wants, as being affective. Nonetheless, the differences here were small, and people still cited desire

reasons in just over half of the cases of providing reason explanations for robot behaviors.

Third, the use of mental state markers was overall quite similar—just as for human agents, people explained robot agents' behaviors often without an explicit mental state verb. However, people used more belief markers and fewer desire markers when explaining robot actions compared to human actions. This diverging pattern of using markers reinforces the above interpretation of people's greater comfort with robots' *thinking* (permitting relatively more belief markers than humans used) and less comfort with robots' *desires* (prompting relatively fewer desire markers than humans used).

Increased use of belief markers, in particular, also functions as a way of "distancing" oneself from the agent [51]—indicating that one does not necessarily share the agent's beliefs that explain their actions. It is possible, then, that participants distanced themselves to some degree from the "minds" of robots. In the future it might be instructive to assess people's general attitudes toward robots to see whether greater rejection of machines goes along with greater belief marker use for distancing purposes. Future studies might also compare people's distancing tendencies from robots with their distancing from outgroups; after all, robots are in significant ways a novel kind of outgroup.

All the documented patterns—both the similar general explanation tendencies and the specific human-robot asymmetries—held up consistently across behavior classes (desirable-undesirable, surprising-unsurprising), and for current behaviors as well as near-future behaviors. This consistency strengthens the conclusion that we have identified genuine explanation tendencies for robot behaviors, not just contingent associations for specific actions. Furthermore, it strengthens the conclusion that we have identified systematic patterns of mental state inferences: (a) a considerable tendency to ascribe mental states to robots in the form of reasons (but still less overall than the rate of ascribing mental states to humans); (b) a preference specifically to ascribe belief reasons and a reluctance to ascribe desire reasons; and (c) an amplification of this combination of preference and reluctance when using explicit mental state verbs to ascribe beliefs and desires. Taken together, we have both amply documented spontaneous mental state inferences to robots and also discovered boundary conditions for such inferences.

A. Limitations and Future Directions

A limitation of the current research is the focus on text descriptions of the target behaviors. We attempted to narrow the ambiguities of interpretation by referring to concrete roles and locations for each behavior, and we systematically equated each behavior for perceptions of intentionality, desirability, and surprise across human and robot agents. The consistency of results across behavior classes boosts our confidence that the findings are generalizable. Nonetheless, we cannot be certain that enriched contexts (including live interactions) will lead to the same patterns of results.

Another limitation is that the manipulation of agent type was implemented by means of minimal descriptions (e.g., a soldier → a robot soldier; a cook → a robot cook). Even though many studies [63]–[65] suggest that people, at least in Western societies, hold similar representations of robots, we exerted no experimental control over those representations. Once more, however, the consistency of results across behaviors, roles, and locations provide some assurance of generalizability. Even if people imagined entirely different robots for each behavior, these images had little impact on the results. It thus appears promising that the small manipulation (of placing a "robot" into the same role and location as a human) had such systematic and interpretable results. However, future variations of wording, information, and images or videos of agent behaviors would clearly strengthen the generalizability of our results. The challenges of achieving consistent live human-robot interactions are well-known, and experimenters have to worry that participants will be distracted by the robot's imperfections and even failures, and that they will actually not explain the same behaviors for human and robot targets. An intermediate step might be virtual- or augmented-reality experiments, which can guarantee the robot's reliability and still allow for inclusion of a range of behaviors, roles, and robot appearances.

B. Conclusion

To our knowledge, this is the first study that has rigorously analyzed people's free-response explanations of robot (compared to human) behavior. Given the grounding of explanations in assumptions of mind and intentionality, the study of behavior explanations can reveal how the specific concepts, processes, and even linguistic structures associated with mind inference are applied to robots. Indeed, our results show that people's spontaneous, free-response explanations of robot behavior rely on similar conceptual tools as they do for human behavior. Even though people also offer unique program-referring explanations of robot behaviors, these explanations occurred only as CHR explanations, not as reason explanations (e.g., people didn't write *It was programmed to think X*). This suggests that people's reason explanations are a relatively clean reflection of folk psychology and mental state inference in robot behavior explanations. Beyond these general similarities, people also used some specific explanation tools differently for robots, and previous work suggests that these patterns reveal a general preference to think of robots as rational rather than motivational-affective beings.

The explanations participants provided for robot behaviors could serve as an initial blueprint for how robots themselves should explain their own behavior. In order to make AI and robots explainable, transparent, and trustworthy [66] designers need to understand how people make sense of such agents, and therefore what kinds of explanations would be illuminating [67]. In earlier work, we presented how the conceptual framework and psychological mechanisms of human behavior explanation could inform the development of explainable robot behavior [45]. Our present results suggest that, at least at current levels of familiarity with technology, people readily apply the familiar concepts and linguistic tools of folk psychology when making sense of robot behavior. This implies that people will be comfortable when robots explain their own behavior using the framework of reasons and the language of beliefs and desires that are so dominant in human interactions.

References

[1] A. L. Woodward, "Infants' grasp of others' intentions.," *Curr. Dir. Psychol. Sci.*, vol. 18, no. 1, pp. 53–57, Feb. 2009.

[2] S. C. Johnson, "The recognition of mentalistic agents in infancy," *Trends Cogn. Sci.*, vol. 4, no. 1, pp. 22–28, Jan. 2000.

[3] S. Kiesler and J. Goetz, "Mental models of robotic assistants," in *CHI '02 Extended Abstracts on Human Factors in Computing Systems*, New York, NY: ACM, 2002, pp. 576–577.

[4] T. Kim and P. Hinds, "Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction," in *Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN06)*, Hatfield, UK, 2006, pp. 80–85.

[5] F. Hegel, S. Krach, T. Kircher, B. Wrede, and G. Sagerer, "Theory of Mind (ToM) on robots: A Functional neuroimaging study," in *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction*, New York, NY, 2008, pp. 335–342.

[6] A. Sciutti, A. Bisio, F. Nori, G. Metta, L. Fadiga, and G. Sandini, "Robots can be perceived as goal-oriented agents," *Interact. Stud.*, vol. 14, no. 3, pp. 329–350, Jan. 2013.

[7] S. Thellman, A. Silvervarg, and T. Ziemke, "Folk-psychological interpretation of human vs. humanoid robot behavior: Exploring the intentional stance toward robots," *Front. Psychol.*, vol. 8, 2017.

[8] N. Epley, A. Waytz, and J. T. Cacioppo, "On seeing human: A three-factor theory of anthropomorphism," *Psychol. Rev.*, vol. 114, no. 4, pp. 864–886, Oct. 2007.

[9] A. Gopnik, "Explanation as orgasm," *Minds Mach.*, vol. 8, no. 1, pp. 101–118, 1998.

[10] F. Heider, *The psychology of interpersonal relations*. New York: Wiley, 1958.

[11] D. J. Hilton, "Causal explanation: From social perception to knowledge-based causal attribution," in *Social psychology: Handbook of basic principles*, 2nd ed., A. W. Kruglanski and E. T. Higgins, Eds. New York, NY: Guilford Press, 2007, pp. 232–253.

[12] B. F. Malle, *How the mind explains behavior: Folk explanations, meaning, and social interaction*. Cambridge, MA: MIT Press, 2004.

[13] S. Baron-Cohen, "Without a theory of mind one cannot participate in a conversation," *Cognition*, vol. 29, pp. 83–84, 1988.

[14] D. Premack and G. Woodruff, "Does the chimpanzee have a theory of mind?," *Behav. Brain Sci.*, vol. 1, no. 4, pp. 515–526, Dec. 1978.

[15] J. D. Greenwood, Ed., *The future of folk psychology: Intentionality and cognitive science*. Cambridge: Cambridge University Press, 1991.

[16] T. Horgan and J. Woodward, "Folk psychology is here to stay," *Philos. Rev.*, vol. 94, pp. 197–226, 1985.

[17] R. G. D'Andrade, "A folk model of the mind," in *Cultural models in language and thought*, D. Holland and N. Quinn, Eds. New York, NY: Cambridge University Press., 1987, pp. 112–148.

[18] F. Eyssel, D. Kuchenbrandt, and S. Bobinger, "Effects of anticipated human-robot interaction and predictability of robot behavior on perceptions of anthropomorphism," in *Proceedings of the 6th International Conference on Human-robot Interaction*, New York, NY, 2011, pp. 61–68.

[19] S. Kiesler, S. I. Lee, and A. D. I. Kramer, "Relationship effects in psychological explanations of nonhuman behavior," *Anthrozoös*, vol. 19, no. 4, pp. 335–352, 2006.

[20] A. Wykowska, T. Chaminade, and G. Cheng, "Embodied artificial agents for understanding human social cognition," *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 371, no. 1693, May 2016.

[21] H. Ku, J. J. Choi, S. Lee, S. Jang, and W. Do, "Designing Shelly, a robot capable of assessing and restraining children's robot abusing behaviors," in *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, New York, NY, USA, 2018, pp. 161–162.

[22] D. J. Gunkel, "The other question: Can and should robots have rights?," *Ethics Inf. Technol.*, vol. 20, no. 2, pp. 87–99, Jun. 2018.

[23] F. Eyssel and D. Kuchenbrandt, "Manipulating anthropomorphic inferences about NAO: The role of situational and dispositional aspects of effectance motivation," presented at the International Symposium on Robot and Human Interactive Communication, 2011.

[24] B. Mutlu, F. Yamaoka, T. Kanda, H. Ishiguro, and N. Hagita, "Nonverbal leakage in robots: Communication of intentions through seemingly unintentional behavior," in *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*, New York, NY, USA, 2009, pp. 69–76.

[25] D. Sirkin, B. Mok, S. Yang, and W. Ju, "Mechanical ottoman: Engaging and taking leave," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, New York, NY, 2015, pp. 275–275.

[26] J. Tullio, A. K. Dey, J. Chalecki, and J. Fogarty, "How it works: A field study of non-technical users interacting with an intelligent system," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, 2007, pp. 31–40.

[27] R. H. Wortham, A. Theodorou, and J. J. Bryson, "What does the robot think? Transparency as a fundamental design requirement for intelligent systems," in *Proceedings of the 2016 IJCAI Workshop on Ethics for Artificial Intelligence, New York*, 2016.

[28] A. D. Dragan, S. Bauman, J. Forlizzi, and S. S. Srinivasa, "Effects of robot motion on human-robot collaboration," in *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction*, New York, NY, 2015, pp. 51–58.

[29] D. T. Levin, M. M. Saylor, and S. D. Lynn, "Distinguishing first-line defaults and second-line conceptualization in reasoning about humans, robots, and computers," *Int. J. Hum.-Comput. Stud.*, vol. 70, no. 8, pp. 527–534, Aug. 2012.

[30] L. Takayama, D. Dooley, and W. Ju, "Expressing thought: Improving robot readability with animation principles," in *Proceedings of the 6th ACM/IEEE International Conference on Human-robot Interaction*, New York, NY, 2011, pp. 69–76.

[31] H. M. Gray, K. Gray, and D. M. Wegner, "Dimensions of mind perception," *Science*, vol. 315, no. 5812, pp. 619–619, Feb. 2007.

[32] H. Takahashi, M. Ban, and M. Asada, "Semantic differential scale method can reveal multi-dimensional aspects of mind perception," *Front. Psychol.*, vol. 7, p. 1717, 2016.

[33] X. Wang and E. G. Krumhuber, "Mind perception of robots varies with their economic versus social function," *Front. Psychol.*, vol. 9, Jul. 2018.

[34] B. Fiala, A. Arico, and S. Nichols, "You, robot," in *Current controversies in experimental philosophy*, E. Machery and E. O'Neill, Eds. Routledge, 2014, pp. 31–47.

[35] M. C. Martini, C. A. Gonzalez, and E. Wiese, "Seeing minds in others – Can agents with robotic appearance have human-like preferences?," *PLOS ONE*, vol. 11, no. 1, p. e0146310, Jan. 2016.

[36] K. Weisman, C. S. Dweck, and E. M. Markman, "Rethinking people's conceptions of mental life," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 114, no. 43, pp. 11374–11379, Oct. 2017.

[37] E. Short, J. Hart, M. Vu, and B. Scassellati, "No fair!! An interaction with a cheating robot," in *Proceedings of the 5th ACM/IEEE International Conference on Human-robot Interaction*, Piscataway, NJ, 2010, pp. 219–226.

[38] H. M. Wellman, A. K. Hickling, and C. A. Schult, "Young children's psychological, physical, and biological explanations," *New Dir. Child Dev.*, vol. 75, pp. 7–25, 1997.

[39] K. Bartsch and H. M. Wellman, *Children talk about the mind*. Oxford University Press, 1995.

[40] C. Antaki, "Explanations, communication and social cognition," in *Analyzing everyday explanation: A casebook of methods*, C. Antaki, Ed. London: Sage Publication, 1988, pp. 1–14.

[41] B. F. Malle and J. Knobe, "Which behaviors do people explain? A basic actor-observer asymmetry.," *J. Pers. Soc. Psychol.*, vol. 72, no. 2, pp. 288–304, 1997.

[42] J. McClure, D. J. Hilton, J. Cowan, L. Ishida, and M. Wilson, "When people explain difficult actions, is the causal question how or why?," *J. Lang. Soc. Psychol.*, vol. 20, no. 3, pp. 339–357, 2001.

[43] B. F. Malle, "F.Ex: Coding scheme for people's folk explanations of behavior. (Originally published 1998.)," Retrieved September 2018 from http://research.clps.brown.edu/SocCogSci/CodingSchemes.html, 2014.

[44] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 3063289, Oct. 2017.

[45] M. M. A. de Graaf and B. F. Malle, "How people explain action (and autonomous intelligent systems should too)," in *AAAI Fall Symposium*

on Artificial Intelligence for Human-Robot Interaction., Arlington, VA, USA, pp. 1–8.

[46] N. Wang, D. V. Pynadath, and S. G. Hill, "Trust calibration within a human-robot team: comparing automatically generated explanations," in *Proceeedings of the 11th ACM/IEEE International Conference on Human Robot Interaction*, Piscataway, NJ, 2016, pp. 109–116.

[47] E. Wang, C. Lignos, A. Vatsal, and B. Scassellati, "Effects of head movement on perceptions of humanoid robot behavior," in *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-robot Interaction*, New York, NY, 2006, pp. 180–185.

[48] B. F. Malle, "How people explain behavior: A new theoretical framework," *Personal. Soc. Psychol. Rev.*, vol. 3, no. 1, pp. 23–48, Feb. 1999.

[49] A. R. Buss, "Causes and reasons in attribution theory: A conceptual critique," *J. Pers. Soc. Psychol.*, vol. 36, pp. 1311–1321, 1978.

[50] B. F. Malle and J. Knobe, "The folk concept of intentionality," *J. Exp. Soc. Psychol.*, vol. 33, no. 2, pp. 101–121, 1997.

[51] B. F. Malle, J. Knobe, M. J. O'Laughlin, G. E. Pearce, and S. E. Nelson, "Conceptual structure and social functions of behavior explanations: Beyond person-situation attributions.," *J. Pers. Soc. Psychol.*, vol. 79, no. 3, pp. 309–326, 2000.

[52] B. F. Malle, J. Knobe, and S. E. Nelson, "Actor-observer asymmetries in explanations of behavior: New answers to an old question.," *J. Pers. Soc. Psychol.*, vol. 93, no. 4, pp. 491–514, 2007.

[53] M. J. O'Laughlin and B. F. Malle, "How people explain actions performed by groups and individuals.," *J. Pers. Soc. Psychol.*, vol. 82, no. 1, pp. 33–48, 2002.

[54] G. W. Bradley, "Self-serving biases in the attribution process: A reexamination of the fact or fiction question," *J. Pers. Soc. Psychol.*, vol. 36, no. 1, pp. 56–71, Jan. 1978.

[55] B. Weiner, "'Spontaneous' causal thinking," *Psychol. Bull.*, vol. 97, pp. 74–84, 1985.

[56] B. F. Malle, "Time to give up the dogmas of attribution: A new theory of behavior explanation," in *Advances of Experimental Social Psychology*, vol. 44, M. P. Zanna and J. M. Olson, Eds. San Diego, CA: Academic Press, 2011, pp. 297–352.

[57] B. Huebner, "Commonsense concepts of phenomenal consciousness: Does anyone care about functional zombies?," *Phenomenol. Cogn. Sci.*, vol. 9, pp. 133–155, 2010.

[58] B. F. Malle, J. Knobe, M. J. O'laughlin, G. E. Pearce, and S. E. Nelson, "Conceptual Structure and Social Functions of Behavior Explanations: Beyond Person–situation Attributions," *J. Pers. Soc. Psychol.*, vol. 79, no. 3, pp. 309–326, Sep. 2000.

[59] M. M. A. de Graaf and B. F. Malle, "People's Judgments of Human and Robot Behaviors: A Robust Set of Behaviors and Some Discrepancies," in *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, New York, NY, USA, 2018, pp. 97–98.

[60] J. Korman and B. F. Malle, "Grasping for traits or reasons? How people grapple with puzzling social behaviors," *Pers. Soc. Psychol. Bull.*, vol. 42, no. 11, pp. 1451–1465, Nov. 2016.

[61] K. L. Gwet, "Computing inter-rater reliability and its variance in the presence of high agreement," *Br. J. Math. Stat. Psychol.*, vol. 61, no. 1, pp. 29–48, May 2008.

[62] J. Westfall, "Cohen's d for 2x2 anova interaction," 28-Oct-2015. [Online]. Available: https://stats.stackexchange.com/questions/179098/cohens-d-for-2x2-anova-interaction. [Accessed: 19-Oct-2018].

[63] M. M. A. de Graaf and S. Ben Allouch, "Anticipating our future robot society: The evaluation of future robot applications from a user's perspective," in *25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2016, pp. 755–762.

[64] A. Weiss, J. Igelsböck, D. Wurhofer, and M. Tscheligi, "Looking Forward to a 'Robotic Society'?," *Int. J. Soc. Robot.*, vol. 3, no. 2, pp. 111–123, Apr. 2011.

[65] C. Ray, F. Mondada, and R. Siegwart, "What do people expect from robots?," in *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2008, pp. 3816–3821.

[66] F. Alaieri and A. Vellino, "Ethical decision making in robots: Autonomy, trust and responsibility," in *Social Robotics*, vol. 9979, A. Agah, J.-J. Cabibihan, A. M. Howard, M. A. Salichs, and H. He, Eds. Cham: Springer International Publishing, 2016, pp. 159–168.

[67] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artifical Intell.*, vol. 267, pp. 1–38, Feb. 2019.