The Perceptual Belief Problem: Why Explainability Is a Tough Challenge in Social Robotics

SAM THELLMAN and TOM ZIEMKE, Linköping University, Sweden

The explainability of robotic systems depends on people's ability to reliably attribute perceptual beliefs to robots, i.e., what robots know (or believe) about objects and events in the world based on their perception. However, the perceptual systems of robots are not necessarily well understood by the majority of people interacting with them. In this article, we explain why this is a significant, difficult, and unique problem in social robotics. The inability to judge what a robot knows (and does not know) about the physical environment it shares with people gives rise to a host of communicative and interactive issues, including difficulties to communicate about objects or adapt to events in the environment. The challenge faced by social robotics researchers or designers who want to facilitate appropriate attributions of perceptual beliefs to robots is to shape human-robot interactions so that people understand what robots know about objects and events in the environment. To meet this challenge, we argue, it is necessary to advance our knowledge of when and why people form incorrect or inadequate mental models of robots' perceptual and cognitive mechanisms. We outline a general approach to studying this empirically and discuss potential solutions to the problem.

CCS Concepts: • **Human-centered computing** → **HCI theory, concepts and models**; *Laboratory experiments*; • **Applied computing** → *Psychology*;

Additional Key Words and Phrases: Human-robot interaction, social robotics, explainability, predictability, understandability, intentionality, intentional stance, common ground, mental state attribution, belief attribution

ACM Reference format:

Sam Thellman and Tom Ziemke. 2021. The Perceptual Belief Problem: Why Explainability Is a Tough Challenge in Social Robotics. *Trans. Hum.-Robot Interact.* 10, 3, Article 29 (July 2021), 15 pages.
<https://doi.org/10.1145/3461781>

1 INTRODUCTION

The line of research that is currently being pursued within the field of human-robot interaction under the banner “explainable robotics”¹ [19, 20] is relatively new, and it is somewhat unclear to what extent the relevant research problems overlap with those of the more general area of explainable artificial intelligence [55]. In this article, we discuss what we consider to be a unique

¹This is sometimes also referred to using the partly overlapping notions of understandability [42, 83], interpretability [40, 54], transparency [49, 76], predictability [26, 82], legibility [26], or explicability [82].

This work was supported by ELLIIT, the Excellence Center at Linköping-Lund in Information Technology. Authors' addresses: S. Thellman and T. Ziemke, Linköping University, Sweden; emails: {sam.thellman, tom.ziemke}@liu.se. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2021 Association for Computing Machinery.

2573-9522/2021/07-ART29 \$15.00

<https://doi.org/10.1145/3461781>

Fig. 1. A robot’s perceptual beliefs are constituted by what the robots knows (or believes) about objects and events in the world based on its perception. The diagram also illustrates the relation between the three problems that this article addresses: the intentionality attribution problem, the belief attribution problem, and the perceptual belief problem.

and significant obstacle in the way of achieving explainability in interactions with robotic systems: the perceptual belief attribution problem or *perceptual belief problem* for short.

In Section 2, we begin the article by asserting the importance of people’s interpretation of robots as agents with intentional states, such as beliefs and desires, to their ability to predict and explain the behavior of robots. We refer to the problem of attributing the appropriate intentional states to robots as the *intentionality attribution problem* in social robotics. We also propose a definition of social robots as intentional systems in the sense of Dennett [24].

The intentionality attribution problem can be broken down into a number of sub-problems, corresponding to different types of intentional states, e.g., beliefs, desires, and intentions. Section 3 focuses on the *belief attribution problem*, i.e., the problem of understanding what robots know or believe (including false beliefs). We review pre-existing work on the belief attribution problem and conclude that little or no work has been concerned with people’s understanding of the perceptual beliefs of robots, i.e., the beliefs that result from a robot’s perception of the world (see Figure 1).²

Section 4 therefore focuses on what we refer to as the *perceptual belief problem*: the problem of understanding what robots know (and do not know) about objects and events in the environment, based on their perception. We explain why this is an important, difficult, and unique problem in social robotics. In particular, the inability to judge what a robot knows about a shared environment gives rise to a host of serious interactive issues, including the inability to communicate about objects and events in the environment and uncertainty about how to act in response to events that transpire in the environment.

In Section 5, we argue that to overcome the perceptual belief problem it is necessary to advance our understanding of when and why people form incorrect or inadequate mental models of the capabilities of robots to perceive and understand the environment. We propose a general approach to studying this, and discuss different types of solutions to the problem.

In Section 6, finally, we conclude our discussion with a broader outlook on the implications of the perceptual belief problem in social interactions with robots.

2 THE INTENTIONALITY ATTRIBUTION PROBLEM

According to Dennett’s *intentional systems theory* [24], the behavior of robots (or any other kind of behaving entity) can be understood, explained, and predicted from the point of view of three different levels of abstraction or “stances.” At the lowest level of abstraction—the *physical*

²It can be argued that all belief is perceptual in the sense expressed in this admittedly rather weak definition of “perceptual belief.” Although the term occurs in the philosophical literature (e.g., [9, 34]), we have found no technical definition that clearly distinguishes perceptual from other belief. This is in line with Quine’s theory that all belief is related to perceptual experience in a relation of “germaneness”: a property of being more or less germane to revision based on sensory experience [62]. We think of perceptual belief as the kind of belief that is particularly germane to sensory experience (i.e., what is perceived).

stance—behavior is viewed in terms of the physical properties of the system. One seldom adopts the physical stance to predict or explain the behavior of robotic systems because of the large number of critical variables in their physical constitution. Although possible in principle, tracing out the effects of the input energies all the way through the robot until some of its effectors are actuated is practically impossible (except in instances of breakdown where the condition preventing normal operation is easily locatable, such as when a robot runs out of power).

The medium level of abstraction is the *design stance*. When taking the design stance, one abstracts away from the physical details of the system and instead view behavior in terms of how the system is designed. Robotic systems are often too complex to understand from the design stance. People generally have little insight into the hardware and software design of robots, and even those who do—e.g., programmers of robotic software—often cannot predict a robot’s next move based on their knowledge as designers of the system (cf. the use of inscrutable machine learning algorithms in robotics [45, 76]). This means that robots are often practically inaccessible from the design stance or the physical stance.

The third and highest level of abstraction—the one that we focus on in this article—is the *intentional stance*. When taking the intentional stance toward a robot, one treats it as a rational agent whose behavior is governed by underlying intentional states, such as beliefs and desires, without considering its physical constitution and design in much detail. Here is Dennett’s explanation of how the intentional stance works:

First you decide to treat the object whose behavior is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs. A little practical reasoning from the chosen set of beliefs and desires will in most instances yield a decision about what the agent ought to do; that is what you predict the agent will do. [24, p. 17]

In some cases, adopting the intentional stance toward an object is a useful strategy for predicting its behavior; in other cases, it is not. Dennett introduced the notion of an *intentional system* to denote objects or systems whose behavior can be (at least sometimes) explained and predicted *only*³ by being attributed to underlying intentional states, such as beliefs and desires (i.e., based on the intentional stance). Chess-playing computers, for instance, are intentional systems, because not even the designers of chess-playing computer programs always manage to predict the program’s next move in the game based on their knowledge of how the program is designed; the only reliable interpretative strategy available is the intentional one, which is to figure out what the program *thinks* that one is going to do next, and what the program *wants* to accomplish in the next couple of turns given its *beliefs*.⁴ The label “intentional system,” however, does not extend to all objects. Although a person might predict that a thermostat will raise the room temperature in the morning, because it *wants* to keep it at 73° and *knows* that it has fallen during the night, the use of such intentional constructs in the interpretation of the thermostat’s behavior does not add predictive value *above and beyond* the corresponding non-psychological interpretation. A simple thermostat, therefore, is not an intentional system.

³The word “only” was not part of Dennett’s [22] original description of intentional systems. However, in later work [24] Dennett clarified his account of intentional systems, excluding systems (e.g., thermostats) whose behavior can be sufficiently explained using non-psychological interpretation.

⁴Although, as pointed out by Dennett, “one can always refuse to adopt the intentional stance toward the computer, and accept its checkmates” [22, p. 91].

It is crucial to understand that what is and what is not an intentional system is not subjectively “in the eye of the observer.” The point is that a person playing against a chess-playing computer does not *choose* whether to adopt the intentional stance; there is simply no other feasible way (for any person playing against a modern chess-playing computer, as far as is known) to predict the computer’s moves. Furthermore, it is important to note that the fact that an object or system is an intentional system does not necessarily mean that it *really* has beliefs and desires; it only means that one can explain and predict its behavior by ascribing beliefs and desires to it [22].⁵ A third important point is that what kinds of robots (and other systems) are intentional systems is ultimately an empirical question—in the same sense that it turned out *as a matter of fact* that the moves of a chess-playing computer can only be predicted reliably based on the intentional stance.

In human–robot interactions, intentional prediction and explanation is not only common [6, 29, 48, 66, 72, 83] but also works when no other sort of prediction is feasible. People treat robots as intentional agents with beliefs, desires, intentions, and so on, independent of considerations about their physical constitution, functional design, or their status as potentially moral, cognitive, conscious or self-conscious entities. This allows them to interact with robots without necessarily understanding how robots “really” work, i.e., without any in-depth knowledge of the underlying physical and computational mechanisms. However, a person’s understanding of a robotic system from the point of view of the intentional stance is only as good as the person’s grasp of the intentional properties of the system.⁶ Consequently, the predictability and explainability of robotic systems depends on people’s ability to solve the intentionality attribution problem of attributing the appropriate intentional states to robots, i.e., beliefs, desires, and so on. The challenge faced by the social robotics researcher or designer who wants to facilitate this process, to increase the predictability and explainability of robots, is to shape human–robot interactions so that people understand the intentional states of robotic systems.

We argue that this view of intentionality attribution also allows to formulate at least a rough alternative definition of what constitutes the “social” in “social robotics.” Despite the centrality of the term, there is at this point no satisfactory definition of “social” in the context of human–robot interaction. Previously proposed definitions are, in our opinion, usually overly inclusive and focus either on capabilities that robots need to interact socially (cf. [4, 21, 27]) or on people’s judgments of robots as social (cf. [7, 41]). Defining “social” robots in terms of robot capabilities does not sufficiently account for the fact that what constitutes “social” is tied to the perspective of an interpreting observer who must understand the robot’s behavior in order for social interaction to occur. Definitions that rest solely on the perspective of an observer, however, are problematic for at least two reasons. First, people commonly misjudge the social qualities of robots. A person might initially judge a robot to be social only to discover that it is incapable of anything remotely similar to social interaction. Secondly, people have different “social models” (cf. Breazeal [7]) or of opinions on what counts as social, some of which are irreconcilable. Hence, both types of previously proposed definitions independently fail to appropriately restrict the extension or application of the term “social robot,” i.e., they are overly inclusive.

We propose that a better definition of what makes robots “social” is that they are intentional systems (in Dennett’s sense, as discussed above), i.e., systems whose behavior is most appropriately

⁵It might be worth noting, however, that Dennett’s own position—to which there are several alternatives (e.g., [11, 31]), is that “what it is to be a true believer is to be an intentional system” [24, p. 15].

⁶What we mean by grasping/understanding the intentional states/properties of a robot is to have a mental model of the robot that is conducive to predicting and explaining the robot’s behavior based on the intentional stance. We do not take a stance on the issue of if and when people’s ascriptions of intentional states to robotic systems are *true* of such systems [70]. This issue, in our opinion, pertains to cognitive science and philosophy of mind—and therefore falls somewhat outside the scope of social robotics/human–robot interaction research.

interpreted in terms of intentional states. From this it also follows, as a corollary definition, that human–robot interactions should be viewed social only if they involve robots that are intentional systems (in the above sense). This view consolidates the intuitions behind both of the previous types of definitions into a stricter one,⁷ which is based on the notion that social interaction (and the need for agents to communicate) is intimately connected to people’s understanding of others as *intentional agents*. This, in our opinion, better captures what makes interactions with robots *social*. It might be worth noting that this notion of social interaction is of course not limited to humans or to humanlike robots, but also extends to other types of interactive autonomous systems. A human pedestrian’s interaction with an automated car at a crosswalk (cf. Ziemke [83]), for example, can be viewed as a social interaction, because pedestrians in that situation are likely to ask themselves “Has that car seen me?”, “Does it understand I want to cross the road?”, or “Is it planning to stop for me?”, and answering these questions is likely to involve the attribution of intentional states, such as beliefs (e.g., there is a person on the crosswalk), desires (e.g., not to collide with people), and intentions (e.g., to slow down and let the person cross the road).

3 THE BELIEF ATTRIBUTION PROBLEM

The intentionality attribution problem can be broken down according to the two main, primitive⁸ types of intentional states that are conducive to predicting and explaining behavior: beliefs and desires.⁹ We here focus on beliefs. The belief attribution problem (and the corresponding research challenge of shaping human–robot interactions so that people understand what robots know or believe) has been given some treatment in the human-robot interaction (HRI) literature, albeit under the different label of the “common ground problem” in HRI [13, 30, 44]. As convincingly argued by Kiesler [44], one of the ultimate goals of HRI should be to enable humans and robots to communicate in accordance with *the principle of least collaborative effort*, which presupposes mutual or “common ground” knowledge [15]. To paint a picture of the efficient communication and interaction that would be possible if humans and robots shared common ground knowledge, Kiesler considers the example of a robot performing as an airport security guard, providing directions and information to airport visitors. Kiesler asks, *how should these robots explain routes through the airport to locals and strangers, to young and old, to people in a hurry and to people with time to fill?* Such a robot, according to Kiesler, would have to adapt to the needs of different travelers, repair inevitable misunderstandings, speak differently to people with different expertise, and so on—in short, it would have to share common ground knowledge with people to communicate and interact efficiently with them.

Kiesler’s scenario represents a good target for social robotics research. However, turning it into reality requires the endowment of a functional robotic theory of mind, that is, a comprehensive, integrated architecture that allows robots to have rich, usable models of human mental states. This

⁷This view on social behavior echoes the sentiment expressed by Alan Turing on the concept of intelligent behavior: “The extent to which we regard something as behaving in an intelligent manner is determined as much by our own state of mind and training as by the properties of the object under consideration” (Turing [74], p. 127). See Proudfoot [61] for a discussion about Turing’s view of intelligence and thought as “response dependent”.

⁸Beliefs and desires are primitive in the sense that other intentional states (e.g., intentions [51]) are intuitively understood as presupposing their existence.

⁹The problem of understanding (attributing) what robots *want*, in analogy to the belief attribution problem, could be referred to as the desire (or goal) attribution problem. These two problems are to some extent interrelated, because what an agent wants (to do) in a given situation may depend on what the agent believes is the case and vice versa. For example, a robot’s beliefs about the immediate environment (e.g., there is a speck of dust under the table) might cause it to want to interact with it in specific ways (e.g., to clean that speck of dust)—perhaps on account of one of its more general desires (e.g., to keep the house clean). Conversely, a cleaning robot might be expected to be able to form beliefs (e.g., about specs of dust in its vicinity) that are relevant to its desire to clean.

has been singled out as one of the most significant challenges that stem from building robots that interact socially with people [81] and might be considered the holy grail of social robotics research. Despite the important work that has been made in this area (e.g., [25, 36, 64, 80]), the technical challenges involved are immense and the results achieved so far bear only the faintest resemblance to Kiesler’s scenario. It seems, therefore, that until substantial technical progress has been made toward building a functional robotic theory of mind, most of the burden involved in establishing common ground knowledge between humans and robots will have to be carried largely by the human part. This is part of the reason why the belief attribution problem (i.e., focusing on people’s understanding of what robots know) is important.

Kiesler and colleagues took significant steps toward solving the belief attribution problem by elucidating some of the ways in which people can be “helped” to estimate robots’ beliefs [47, 47, 59]. Specifically, they demonstrated in a number of studies that people’s estimates of what a robot knows can be influenced by manipulating social cues related to community co-membership—one of three main factors that allow people to estimate others’ knowledge besides linguistic co-presence and physical co-presence [14]. For instance, Powers and Kiesler [59] demonstrated that increased compliance with an advisory robot can be achieved by indirectly manipulating people’s mental model of the robot through direct manipulation of features in the robot’s physical design. Lee et al. [47] showed that participants in their study made different estimations of a robot’s knowledge about national landmarks when they were told that the robot was made in the U.S. versus China and the robot spoke English versus Chinese. Similarly, participants in an experiment by Powers et al. [60] made different estimations of a robot’s knowledge of dating norms depending on whether the robot was endowed with a female or male voice and appearance. Importantly, these studies identified factors that contribute to the formation and reconstruction of people’s mental models of what robots know and do not know; factors that can be exploited in the design of human–robot interactions, where the accuracy of a user’s mental model is crucial to the predictability and explainability of robotic behavior.

A significant limitation in the way that the belief attribution problem has been approached thus far is that very little attention has been given to how people estimate robots’ *perceptual beliefs*, i.e., what robots know (or believe) about objects and events in the world based on their perception (cf. Figure 1). The ability to estimate what a co-present agent knows about a shared environment is a fundamental interactive skill that is essential in interactions with robots (as physically co-present interactive objects). In the following, we explain why people’s understanding of robots’ perceptual beliefs is subject to a significant limitation that negatively affects the predictability and explainability of robotic systems.

4 THE PERCEPTUAL BELIEF PROBLEM

The physical embodiment of robots allows humans and robots to share the same physical environment. Furthermore, shared knowledge can be facilitated by co-presence and mutual observation. Being co-present with one another, humans and robots can, in principle, see, hear, and interact with what the other is interacting with, allowing them to, among other things, anticipate each other’s actions in relation to objects and events in a shared environment. However, estimating another person’s knowledge based on co-presence presupposes that one is able to reliably infer the other’s beliefs about objects and events in the environment.¹⁰ In cases of human–human interaction, this is, more often than not, unproblematic—people are generally safe to as-

¹⁰Note that we do not suggest that beliefs are necessarily propositional or that belief inference is necessarily active in the sense of being conscious or effortful. More specifically, intuitive or non-inferential processes (e.g., mental simulation, see Gordon [35]) may also play an important role in people’s understanding of the mental states of robots.

Fig. 2. A person’s judgment of what a co-present robot knows about a particular state of affairs (X) in the environment depends on uncertain estimates of the robot’s capability to perceive and understand. In addition to the impediment in predictability and explainability that such uncertainty entails, not knowing what a robot knows about the environment may give rise to uncertainties about one’s own future states and actions when interacting with the robot.

sume that another person who is looking at an object is also *attending* to it, and will draw similar *conclusions* about it [14]. However, in human–robot interactions such assumptions are precarious due to the fact that the mechanisms by which humans and robots perceive and understand the world differ considerably (see Figure 2). While research in embodied AI and cognitive science (e.g., [8, 12, 84]) has acknowledged the fact that any robot necessarily has a different perceptual world—or *Umwelt* [75]—than humans, the implications for HRI have received relatively little attention so far. What we call the perceptual belief attribution problem in social robotics/HRI—or perceptual belief problem for short—is this: *How can people understand what robots know (and do not know) about the shared physical environment?*

One seemingly straightforward way of solving the perceptual belief problem might be to build robots with similar perception and cognition as humans, so that people that interact with a robot can simply assume it to have sufficiently similar beliefs about the environment as a person typically would have in the given situation.¹¹ This solution is, however, implausible in the foreseeable future due to the fact that the perceptual and cognitive mechanisms that govern belief formation in humans are largely unknown and cannot presently be implemented in a machine. Hence, until such a solution is possible (if possible and feasible at all), people encountering and interacting with robots will face the problem of inferring their ability to perceive and understand the environment without relying on the assumption that it is highly similar to one’s own ability.

To highlight the significance of the perceptual belief problem, it can be useful to consider how central the assumption of similar perception and cognition is in many types of interpersonal interactions. When, for example, we encounter a road vehicle in traffic, the driver’s beliefs about the ongoing traffic situation often cannot be easily inferred directly based on information in the shared environment. Instead, we must assume the driver to believe the things a driver typically would believe in the given situation. When another vehicle drives straight toward us in our lane on a country road during an overtake, we trust that with high probability—high enough for us not to panic, break, or drive off the road—the other driver is aware of our presence and will return to their own lane as soon as the overtake is complete. Arguably, it is our very strongly held

¹¹This might also require that robots are built with the appropriate external features to communicate humanlike capabilities, such as eye-gaze directionality [1, 53].

assumption that others have similar perception and cognition as ourselves that allow these interactions to take place.

For a less dramatic but more commonplace example of reliance on similar perception and cognition, consider the use of definite reference, that is, linguistic expressions used to refer to a definitely identifiable referent, such as when referring to *this candle*. As argued by Clark and Marshall [14], such linguistic behavior presupposes evidence of *triple co-presence*—of events in which the speaker, addressee, and target object (the candle) are co-present. However, mere co-presence is not enough for definite reference to work. The speaker must be able to count on the addressee being aware of the object. We propose that the reason why people can make such assumptions about one another, with a high degree of predictive success, is that the mechanisms that underlie human perception and cognition are sufficiently similar across individuals in the human population. Because of this similarity, people are able to estimate what others know about the environment based on what they themselves know [57, 68]. However, in human–robot interactions, because speakers often cannot count on robots to be aware of objects and events in the environment, the use of proper nouns and spatial deixis (here, there, this, that, and so on) as communicative devices becomes significantly limited.

Perhaps more serious and fundamental than any communicative issue, the perceptual belief problem makes it difficult for people to predict how robots will behave in response to events that transpire in the environment. This, in turn, has a compounding effect on people’s uncertainty about their own future states and actions. For example, if it is unclear whether a robot is aware of the presence and attributes of an object that is on a collision course with it, it will be difficult to predict whether the robot will act to avoid this collision and therefore likely also difficult to decide whether to intervene in the robot’s execution of its planned actions to prevent an accident that might never happen anyway. In interpersonal interactions, people constantly keep track of others’ perceptual beliefs [46] and act to minimize discrepancies between the beliefs of others and themselves when it is relevant within the context of the given situation (e.g., through explaining their own behavior or asking for explanations of others’ [50]). However, such “interaction-managing actions” are impossible if the beliefs of the robot cannot be tracked in the first place. The perceptual belief problem thus causes difficulty not only in predicting and explaining robot behavior but in managing interactions with robots. This might also negatively impact people’s trust in robotic systems (e.g., Hancock et al. [39]), thereby potentially reducing user acceptance [18, 33].

The perceptual belief problem is largely unique to social robotics. It does not appear in cases of human–robot interaction where people interacting with a robot do not have to consider what the robot knows to predict and explain its behavior (i.e., where the robot does not meet the criteria of being “social” expounded in Section 2.1). For example, the owner of a lawnmower robot typically does not have to reason about what the robot knows to use it and avoid collisions. The perceptual belief problem is not (or much less) present in typical human–computer interactions, such as in interactions with chess-playing computers or screen-confined AI applications, because computers, in contrast to social robots, do not have to deal with the physical environment in a comprehensive way (i.e., computers have relatively few and simple or well-adjusted perceptual beliefs, if any, such as the positions of pieces on a chessboard). The problem also generally does not arise in human–human interactions, because people are able to estimate each other’s perceptual beliefs with high precision, based on the assumption that others generally believe the same things about the environment as themselves [14, 68] (which holds true because of the underlying similarity in perceptual and cognitive mechanism amongst individuals in the human population). The problem might be present to some degree in human–animal interactions. However, the fact that animals have (partly shared) evolutionary and, to some extent, cultural histories that constrain their intra-species perception and cognition makes it generally easier to infer the perceptual beliefs of an animal than

a robot, which could be endowed with virtually any perceptual and cognitive mechanisms technically available. For example, people know relatively well what to expect and not expect of dogs or cats, but at this point that is simply not the case for robots.

In sum, the perceptual belief problem is a largely unique problem in social robotics that adversely affects the predictability and explainability of robotic systems. In particular, the need to treat robots as intentional systems, combined with the inability to judge what a robot knows about a shared environment, gives rise to serious interactive problems, such as the inability to communicate about objects and events in the environment and uncertainty about how to act in response to events that transpire in the environment. It might also be worth noting that belief inference is fundamental to the inference of other types of psychological states, including intentions [51], emotions (e.g., as in “it is upset because it thinks that I am not paying attention”), and desires [16]. Because of this interdependence, inference of other psychological states (possibly including non-intentional ones) may also be affected by the perceptual belief problem. In the following sub-section, we outline a general research approach to studying and overcoming the problem.

5 MEETING THE CHALLENGE

To make progress on the perceptual belief problem in HRI it is necessary to advance our understanding of when and why people form incorrect or inadequate mental models of the perceptual and cognitive capabilities that govern belief formation in robots. We outline a general approach to studying this empirically using our own recent experimental work as an example (Section 5.1) and then discuss different types solutions to the problem (Section 5.2).

5.1 Understanding the Problem

Any approach to studying belief attribution needs to deal with the fact that people’s beliefs about other’s beliefs are not directly observable. That is, they need to be inferred based on some indirect evidence. A common approach in HRI research (e.g., [19, 29, 52, 69]) is to simply ask people about their beliefs about robots, their behavior, and the underlying mechanisms. However, because people’s beliefs are not necessarily accessible to introspection [37, 58, 63] they cannot be reliably and fully assessed based on such self-report data. Indeed, a mounting body of evidence demonstrates that people’s verbal responses to questions about robots are influenced by a wide variety of factors, including the way in which questions are asked, constraints on study participant responses, and cultural platitudes [29]. For example, it has been observed that subjects may describe robots in mental terms, but nevertheless deny that robots have minds when asked directly [32]. Furthermore, questions about robots’ mental states are ambiguous in that they can be interpreted as concerning the reality of the attributed states [70].

As an alternative to using verbal measures to study belief attribution, our own approach to investigating how people deal with the perceptual belief problem in human–robot interactions uses non-verbal false-belief tasks [71]. False-belief tasks [3, 23, 79] are predicated on a basic insight: Some beliefs have highly specific and predictable behavioral consequences in particular situations. If, for example, in the case of the classical Sally–Anne test [3], experimental subjects think that Sally holds a false belief about the location of an object (e.g., having observed the object being displaced by Anne in Sally’s absence), then they are likely to predict that Sally will look for the object in the wrong place. If, however, they think Sally has a true belief (e.g., having observed that Sally is aware of the displacement), then they are likely to predict that Sally will look for the object in the right place. That means, the subjects’ predictions of Sally’s behavior implicitly tell us something about their beliefs about Sally’s beliefs.

In recent empirical work [71] our goal is to study people’s assumptions about the perceptual and cognitive capabilities of robots by observing their attributions of perceptual beliefs, including

when and why such attributions succeed or fail. This can be accomplished by modifying the standard false-belief task in such a way that subjects have to predict the behavior a robot based on what they think it can or cannot perceive in the environment. In a recent study [71], we used a non-verbal anticipatory-looking paradigm, which helps to circumvent the methodological issues discussed above. The experiment involved three types of false-belief task scenarios, in which the behavior of observed humans is easy to predict, because subjects understand what other people can and cannot see or hear, but much harder to predict in the case of a robot, whose visual and auditory capacities subjects know very little about. The experimental results (see Thellman & Ziemke [71] for details) indicate three types of circumstances in which people's assumptions about the perceptual capabilities of robots are inadequate. More specifically, our results indicate that one of the reasons why people form incorrect mental models of the perceptual and cognitive capabilities of robots is that they—unsurprisingly—assume these capabilities to be more or less humanlike until they are presented with evidence to the contrary, i.e., they default to an “anthropocentric bias” (cf. Thellman & Ziemke [71]). This underscores the importance as well as the difficulty of shaping human–robot interactions so that people can more easily infer the non-human capabilities of robots.

It should be noted that the proposed false-belief task methodology used in our own experimental work can of course be modified to study other types of human–robot interaction scenarios that demand perceptual belief attribution. Clearly, people's assumptions about what robots can and cannot do, in terms of perceptual and cognitive capabilities that govern belief formation, are likely to vary in different real-world scenarios in which they might encounter robots. It is crucial that these variations—and the human capability to deal with them—are identified and studied. Presumably, there are other ways to approach the perceptual belief problem that are complementary to—and possibly better than—our own approach. Our particular approach has the advantage that it does not rely on verbal self-report evidence in the assessment of perceptual belief attribution. But clearly a fuller understanding of how people deal with (perceptual) belief attribution to robots will require a richer toolbox of alternative and complementary approaches.

5.2 Possible Solutions to the Problem

Regarding possible practical solutions to the perceptual belief problem in human–robot interaction, we can roughly distinguish the following two general types of solutions, which will also be addressed in more detail in the following two subsections:

- Endogenous (or internal) solutions are based on repeated exposure to the problematic situation or interaction, in the course of which people learn to correctly interpret the robot's behavior.
- Exogenous (or external) solutions rely on providing additional explicit signals of the robot's (perceptual) beliefs that help guide people through the interaction (e.g., a robot that indicates that it has detected an approaching person by establishing eye contact and/or asking “Can I help you?”).

5.2.1 Endogenous Solutions. People might be able to adapt their understanding of the perceptual beliefs and mechanisms of a robot on their own based on simply observing it interacting with the environment [71]. Examining whether this is the case is critical. The ways to manipulate people's estimations of robots' beliefs that have so far been identified (e.g., [47, 59, 60]) mainly rely on informing participants or designing the behavior or physical appearance of robots prior to interactions. Realistically, however, not all uncertainties that people have about robots (or any agent) can be resolved before interactions take place [28]. This means that some of the uncertainties that people have when engaging in interactions with robots must be resolved during *ongoing* interac-

tions. We believe that estimating what a co-present robot knows about objects and events in a shared environment is a particularly important case in point: presumably, people learn about the perceptual and cognitive capabilities of robots when they observe them interact with the environment much in the same way that people naturally manage and repair misunderstandings based on observation in interpersonal interaction and conversation [5, 65]. If, depending on the situation at hand, a person can quickly learn about a robot's capabilities from observation, then that person's initial lack of understanding might be inconsequential to his or her ability to interact with that robot. However, in cases where people are unable to improve their understanding based on observation, it becomes relevant (perhaps even necessary) to "help" people to form adequate mental models of robots' perceptual and cognitive capabilities so that they can interact with them. Below, we briefly outline four different types of "exogenous" solutions whereby perceptual belief inference is facilitated by providing external sources of information.

5.2.2 Exogenous Solutions. A white cane, which is used as a navigation device by many people who are blind or visually impaired, helps onlookers to identify its user as visually impaired. That means, the cane, or observers' interpretation of the cane, functions as a partial explanation of the observed blind person's behavior. Perhaps robots can be made to signal their perceptual and cognitive capabilities to onlookers in similar ways that help people infer their perceptual beliefs (cf. [17, 38, 73]). While such "white cane solutions" might be useful to communicate the presence or absence of a particular capability (such as in the case of the blind person with the cane), they might be of limited value for communicating nuances in robots' awareness, such as that a robot can see or hear things in certain conditions or with a certain (e.g., non-human) acuity, or that its understanding of objects and events only covers a particular domain of knowledge. For this purpose, overt attentional cueing (e.g., through robotic eye-gaze behavior [56]) might be a more suitable method. However, the potential of these methods as solutions to the perceptual belief problem has not been systematically evaluated.

Second, it might be useful to communicate the capabilities and limitations of a robotic systems verbally to users before the interaction takes place. Verbal instructions/explanations can be used to describe and highlight specific features of a robot's perceptual and cognitive capabilities, thereby making it easier for people to infer what perceptual beliefs the robot ought to have in a situation given those capabilities. However, it might be difficult or impossible to specify in advance how a robot will behave in a highly dynamic and unpredictable environment even though one has a good understanding of its perceptual and cognitive capabilities. Instructions might also be of limited use in time-critical situations where there is little or no time for people to digest them.

A third category of solutions might leverage augmented, mixed, or virtual reality technology to communicate what a robot sees, hears, attends to or understands about a present state of affairs. Williams et al. [78], for example, identified a number of different approaches to generating deictic robot gestures in mixed reality. For example, a robot could be made to point toward or highlight present and future states of objects and events in a person's field of view inside a head-mounted display or by using a projector [2, 10, 67, 77]. These different approaches have distinct advantages and disadvantages. However, a general limitation is the lack of the technical solutions needed to allow robots to intelligently decide for themselves when and how to generate such gestures [78].

Finally, a fourth possibility is to design robots that verbally communicate their current state of belief about their surroundings when appropriate (e.g., by utilizing self-explanations of behavior (e.g., Kaptein et al. [43])). This presumably requires that the robot is able to infer discrepancies between its own state of belief and people's mental models of its beliefs, which presupposes the endowment of a functional robotic theory of mind [64] (see discussion in Section 3 about the present limitations of such technology).

The different types of solutions to the perceptual belief problem outlined in this section are not intended as an exhaustive list; we merely attempt to sketch out a few possible trajectories for future research on the problem.

6 CONCLUSION

People's ability to predict and explain robot behavior based on the intentional stance is significantly limited by uncertainty about the capabilities of robots to perceive and understand the environment [71, 83]. The inability to judge what a robot knows (and does not know) about a shared environment makes it difficult or impossible to communicate with it about things in the environment (e.g., using proper nouns or deictic words such as "here," "there," "this," or "that") and is likely to give rise to uncertainty about one's own future states and events when interacting with the robot (e.g., deciding whether to intervene when it is unclear whether a robot is aware that something negative is about to happen). We refer to this problem as the *perceptual belief problem* in social robotics/HRI, which is part of the broader *belief attribution problem*, which in turn is part of the *intentionality attribution problem*.

The challenge faced by social robotics researchers or designers who want to facilitate appropriate attributions of perceptual beliefs to robots is to shape human–robot interactions so that people understand what robots know about objects and events in the environment. To meet this challenge, we argued, it is necessary to advance our knowledge of when and why people form incorrect or inadequate mental models of robots' perceptual and cognitive mechanisms. We proposed a general approach to studying this based on the idea that flaws in people's understanding can be identified by exposing them to situations in which they can only predict the behavior of a robot if they have an adequate understanding of some of its perceptual or cognitive capabilities (i.e., a modified version of the standard false-belief task mainly used in theory of mind research). When such flaws are identified, it becomes possible to both explore endogenous solutions (people learn about the robot's capabilities based on observing it interact with the environment) and evaluate exogenous solutions (e.g., "white cane solutions," verbal instructions, or self-explanation) to the perceptual belief problem.

The perceptual belief problem arises more or less uniquely in social interactions with robots, because robots, at least for the foreseeable future, have particularly poorly understood perceptual systems. This is partly a consequence of the fact that the perceptual systems of robots, in contrast to humans or animals, are not constrained by evolution. The main factor limiting the inferential problem space that one is confronted with when trying to figure out a robot's perception and cognition is therefore likely, instead of biological and cultural evolution, the current state of the art in robotic technology. The perceptual and cognitive mechanisms of robots will remain considerably different from those of humans (at least) until substantial progress has been made in the cognitive sciences on the elucidation of human perception and cognition. This means that people interacting with robots in the foreseeable future will have to adapt to agents whose perception and cognition are a considerably diverse and constantly moving target. It also means that, as long as robots lack the appropriate socio-cognitive faculties, humans have to do most of the work of establishing common ground and coordinating social interaction. While people might in some cases be able to work out the perceptual and cognitive mechanisms of a robot by themselves, it is likely that designers of human–robot interactions will have to think more consciously about solutions to the perceptual belief problem as technological developments allow robots to interact with their physical and social environment in increasingly complex ways.

REFERENCES

- [1] Henny Admoni and Brian Scassellati. 2017. Social eye gaze in human-robot interaction: A review. *J. Hum.-Robot Interact.* 6, 1 (2017), 25–63.
- [2] Rasmus S. Andersen, Ole Madsen, Thomas B. Moeslund, and Heni Ben Amor. 2016. Projecting robot intentions into human environments. In *Proceedings of the 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'16)*. IEEE, 294–301.
- [3] Simon Baron-Cohen, Alan M. Leslie, and Uta Frith. 1985. Does the autistic child have a “theory of mind”? *Cognition* 21, 1 (1985), 37–46.
- [4] Christoph Bartneck and Jodi Forlizzi. 2004. A design-centred framework for social human-robot interaction. In *Proceedings of the 13th IEEE International Workshop on Robot and Human Interactive Communication*. IEEE, 591–594.
- [5] Carla Bazzanella and Rossana Damiano. 1999. The interactional handling of misunderstanding in everyday conversations. *J. Pragmat.* 31, 6 (1999), 817–836.
- [6] Francesco Bossi, Cesco Willemse, Jacopo Cavazza, Serena Marchesi, Vittorio Murino, and Agnieszka Wykowska. 2020. The human brain reveals resting state activity patterns that are predictive of biases in attitudes towards robots. *Sci. Robot.* 5, eabb6652 (2020).
- [7] Cynthia Breazeal. 2003. Toward sociable robots. *Robot. Autonom. Syst.* 42, 3-4 (2003), 167–175.
- [8] Rodney A. Brooks. 1991. Intelligence without representation. *Artif. Intell.* 47, 1-3 (1991), 139–159.
- [9] Tyler Burge. 2003. Perceptual entitlement. *Philos. Phenom. Res.* 67, 3 (2003), 503–548.
- [10] Ravi Teja Chadalavada, Henrik Andreasson, Robert Krug, and Achim Lilienthal. 2015. That’s on my mind! robot to human intention communication through on-board projection on shared floor space. In *Proceedings of the 2015 European Conference on Mobile Robots (ECMR'15)*. IEEE, 1–6.
- [11] Paul M. Churchland. 1981. Eliminative materialism and propositional attitudes. *J. Philos.* 78, 2 (1981), 67–90.
- [12] Andy Clark. 1998. *Being There: Putting Brain, Body, and World Together Again*. MIT Press.
- [13] Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.
- [14] Herbert H. Clark and Catherine Marshall. 1981. Definite knowledge and mutual knowledge. In *Elements of Discourse Understanding*, Aravind Joshi, Bonnie Webber, and Ivan Sag (Eds.). Cambridge University Press, 10–63.
- [15] Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition* 22, 1 (1986), 1–39.
- [16] Gergely Csibra and György Gergely. 2007. ‘Obsessed with goals’: Functions and mechanisms of teleological interpretation of actions in humans. *Acta Psychol.* 124, 1 (2007), 60–78.
- [17] Koen De Clercq, Andre Dietrich, Juan Pablo Núñez Velasco, Joost de Winter, and Riender Happee. 2019. External human-machine interfaces on automated vehicles: Effects on pedestrian crossing decisions. *Hum. Fact.* 61, 8 (2019), 1353–1370.
- [18] Maartje de Graaf and Somaya Ben Allouch. 2013. Exploring influencing variables for the acceptance of social robots. *Robot. Autonom. Syst.* 61, 12 (2013), 1476–1486.
- [19] Maartje de Graaf and Bertram Malle. 2019. People’s explanations of robot behavior subtly reveal mental state inferences. In *Proceedings of the 2019 ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, 239–248.
- [20] Maartje de Graaf, Bertram F Malle, Anca Dragan, and Tom Ziemke. 2018. Explainable robotic systems. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 387–388.
- [21] Sergi del Moral, Diego Pardo, and Cecilio Angulo. 2009. Social robot paradigms: An overview. In *Proceedings of the International Work-Conference on Artificial Neural Networks*. Springer, 773–780.
- [22] Daniel Dennett. 1971. Intentional systems. *J. Philos.* 68, 4 (1971), 87–106.
- [23] Daniel Dennett. 1978. Beliefs about beliefs. *Behav. Brain Sci.* 1, 4 (1978), 568–570.
- [24] Daniel Dennett. 1989. *The Intentional Stance*. MIT Press.
- [25] Sandra Devin and Rachid Alami. 2016. An implemented theory of mind to improve human-robot shared plans execution. In *Proceedings of the 2016 ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, 319–326.
- [26] Anca D. Dragan, Kenton C. T. Lee, and Siddhartha S. Srinivasa. 2013. Legibility and predictability of robot motion. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction*. IEEE Press, 301–308.
- [27] Brian Duffy. 2003. Anthropomorphism and the social robot. *Robotics and Autonomous Systems* 42, 3–4 (2003), 177–190.
- [28] Oriel FeldmanHall and Amitai Shenhav. 2019. Resolving uncertainty in a social world. *Nat. Hum. Behav.* 3, 5 (2019), 426–435.
- [29] Brian Fiala, Adam Arico, and Shaun Nichols. 2014. You, robot. In *Current Controversies in Experimental Philosophy*, Edouard Machery and Elizabeth O’Neill (Eds.). Routledge.
- [30] Kerstin Fischer. 2011. How people talk with robots: Designing dialog to reduce user uncertainty. *AI Mag.* 32, 4 (2011), 31–38.
- [31] Jerry A. Fodor. 1975. *The Language of Thought*. Vol. 5. Harvard University Press.
- [32] Susan R. Fussell, Sara Kiesler, Leslie D. Setlock, and Victoria Yew. 2008. How people anthropomorphize robots. In *Proceedings of the 2008 ACM/IEEE International Conference on Human-Robot Interaction*. 145–152.

- [33] Ilaria Gaudiello, Elisabetta Zibetti, Sébastien Lefort, Mohamed Chetouani, and Serena Ivaldi. 2016. Trust as indicator of robot functional and social acceptance. An experimental study on user conformation to iCub answers. *Comput. Hum. Behav.* 61 (2016), 633–655.
- [34] Alvin I. Goldman. 1976. Discrimination and perceptual knowledge. *J. Philos.* 73, 20 (1976), 771–791.
- [35] Robert M. Gordon. 1992. The simulation theory: Objections and misconceptions. *Mind Lang.* 7, 1–2 (1992), 11–34.
- [36] Jesse Gray and Cynthia Breazeal. 2014. Manipulating mental states through physical action. *Int. J. Soc. Robot.* 6, 3 (2014), 315–327.
- [37] Anthony G. Greenwald and Mahzarin R. Banaji. 1995. Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychol. Rev.* 102, 1 (1995), 4.
- [38] Azra Habibovic, Victor Malmsten Lundgren, Jonas Andersson, Maria Klingegård, Tobias Lagström, Anna Sirkka, Johan Fagerlönner, Edgren Claes, Rikard Fredriksson, Stas Krupenia, Dennis Saluäär, and Pontus Larsson. 2018. Communicating intent of automated vehicles to pedestrians. *Front. Psychol.* 9 (2018), 1336.
- [39] Peter A. Hancock, Deborah R. Billings, Kristin E. Schaefer, Jessie Y. C. Chen, Ewart J. De Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Hum. Fact.* 53, 5 (2011), 517–527.
- [40] Bradley Hayes and Julie A. Shah. 2017. Improving robot controller transparency through autonomous policy explanation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, 303–312.
- [41] Frank Hegel, Claudia Muhl, Britta Wrede, Martina Hielscher-Fastabend, and Gerhard Sagerer. 2009. Understanding social robots. In *Proceedings of the 2nd International Conference on Advances in Computer-Human Interactions*. IEEE, 169–174.
- [42] Thomas Hellström and Suna Bensch. 2018. Understandable robots-what, why, and how. *Paladyn* 9, 1 (2018), 110–123.
- [43] Frank Kaptein, Joost Broekens, Koen Hindriks, and Mark Neerinx. 2017. Personalised self-explanation by robots: The role of goals versus beliefs in robot-action explanation for children and adults. In *Proceedings of the 2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'18)*. IEEE, 676–682.
- [44] Sara Kiesler. 2005. Fostering common ground in human-robot interaction. In *Proceedings of the 2005 IEEE International Workshop on Robot and Human Interactive Communication*. IEEE, 729–734.
- [45] Jens Kober, J. Andrew Bagnell, and Jan Peters. 2013. Reinforcement learning in robotics: A survey. *Int. J. Robot. Res.* 32, 11 (2013), 1238–1274.
- [46] Ágnes Kovács, Ernő Téglás, and Denis Endress. 2010. The social sense: Susceptibility to others' beliefs in human infants and adults. *Science* 330, 6012 (2010), 1830–1834.
- [47] Sau-lai Lee, Ivy Yee-man Lau, Sara Kiesler, and Chi-Yue Chiu. 2005. Human mental models of humanoid robots. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*. 2767–2772.
- [48] Daniel T. Levin, Stephen S. Killingsworth, Megan M. Saylor, Stephen M. Gordon, and Kazuhiko Kawamura. 2013. Tests of concepts about different kinds of minds: Predictions about the behavior of computers, robots, and people. *Hum.-Comput. Interact.* 28, 2 (2013), 161–191.
- [49] Joseph B. Lyons. 2013. Being transparent about transparency: A model for human-robot interaction. In *Proceedings of the 2013 AAAI Spring Symposium Series*. 48–53.
- [50] Bertram Malle. 2006. *How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction*. MIT Press.
- [51] Bertram Malle and Joshua Knobe. 1997. The folk concept of intentionality. *J. Exp. Soc. Psychol.* 33, 2 (1997), 101–121.
- [52] Serena Marchesi, Davide Ghiglino, Francesca Ciardo, Jairo Perez-Osorio, Ebru Baykara, and Agnieszka Wykowska. 2019. Do we adopt the intentional stance toward humanoid robots? *Front. Psychol.* 10 (2019).
- [53] Andrew N. Meltzoff, Rechele Brooks, Aaron P. Shon, and Rajesh P. N. Rao. 2010. “Social” robots are psychological agents for infants: A test of gaze following. *Neur. Netw.* 23, 8–9 (2010), 966–972.
- [54] Joseph E. Mercado, Michael A. Rupp, Jessie Y. C. Chen, Michael J. Barnes, Daniel Barber, and Katelyn Procci. 2016. Intelligent agent transparency in human-agent teaming for Multi-UxV management. *Hum. Fact.* 58, 3 (2016), 401–415.
- [55] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* 267 (2019), 1–38.
- [56] Bilge Mutlu, Fumitaka Yamaoka, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2009. Nonverbal leakage in robots: Communication of intentions through seemingly unintentional behavior. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*. 69–76.
- [57] Raymond Nickerson. 1999. How we know—and sometimes misjudge—what others know: Imputing one's own knowledge to others. *Psychol. Bull.* 125, 6 (1999), 737–759.
- [58] Richard E. Nisbett and Timothy D. Wilson. 1977. Telling more than we can know: Verbal reports on mental processes. *Psychol. Rev.* 84, 3 (1977), 231.
- [59] Aaron Powers and Sara Kiesler. 2006. The advisor robot: Tracing people's mental model from a robot's physical attributes. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction*. ACM, 218–225.
- [60] Aaron Powers, Adam Kramer, Shirlene Lim, Jean Kuo, Sau-lai Lee, and Sara Kiesler. 2005. Eliciting information from people with a gendered humanoid robot. In *Proceedings of the 2005 IEEE International Workshop on Robot and Human Interactive Communication*. IEEE, 158–163.

- [61] Diane Proudfoot. 2013. Rethinking turing’s test. *J. Philos.* 110, 7 (2013), 391–411.
- [62] Willard V. Quine. 1951. Main trends in recent philosophy: Two dogmas of empiricism. *Philos. Rev.* 60, 1 (1951), 20–43.
- [63] Arthur S. Reber. 1989. Implicit learning and tacit knowledge. *J. Exp. Psychol.: Gen.* 118, 3 (1989), 219.
- [64] Brian Scassellati. 2002. Theory of mind for a humanoid robot. *Auton. Robots* 12, 1 (2002), 13–24.
- [65] Emanuel A. Schegloff. 1987. Some sources of misunderstanding in talk-in-interaction. *Linguistics* 25, 1 (1987), 201–218.
- [66] Elef Schellen and Agnieszka Wykowska. 2019. Intentional mindset toward robots – open questions and methodological challenges. *Front. Robot. AI* 5 (2019), 139.
- [67] Elena Sibirtseva, Dimosthenis Kontogiorgos, Olov Nykvist, Hakan Karaoguz, Iolanda Leite, Joakim Gustafson, and Danica Kragic. 2018. A comparison of visualisation methods for disambiguating verbal requests in human-robot interaction. In *Proceedings of the 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN’18)*. IEEE, 43–50.
- [68] Diana I Tamir and Jason P Mitchell. 2013. Anchoring and adjustment during social inferences. *J. Exp. Psychol.: Gen.* 142, 1 (2013), 151.
- [69] Sam Thellman, Annika Silvervarg, and Tom Ziemke. 2017. Folk-psychological interpretation of human vs. humanoid robot behavior: Exploring the intentional stance toward robots. *Front. Psychol.* 8 (2017), 1962.
- [70] Sam Thellman and Tom Ziemke. 2019. The intentional stance toward robots: Conceptual and methodological considerations. In *Proceedings of the 41st Annual Conference of the Cognitive Science Society*. 1097–1103.
- [71] Sam Thellman and Tom Ziemke. 2020. Do You See what I See? Tracking the Perceptual Beliefs of Robots. *iScience* 23, 10 (2020), 101625. <https://doi.org/10.1016/j.isci.2020.101625>
- [72] Serge Thill and Tom Ziemke. 2017. The role of intentions in human-robot interaction. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 427–428.
- [73] Sofia Thunberg and Tom Ziemke. 2021. User-centred design of humanoid robots’ communication. *J. Behav. Robot.* 12, 1 (2021), 58–73.
- [74] Alan M. Turing. 1948. *Intelligent Machinery*. Technical Report. NPL. Mathematics Division.
- [75] Jakob von Uexküll. 1982. The theory of meaning. *Semiotica* 42, 1 (1982), 25–79.
- [76] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. 2017. Transparent, explainable, and accountable AI for robotics. *Sci. Robot.* 2, 6 (2017), eaan6080.
- [77] Michael Walker, Hooman Hedayati, Jennifer Lee, and Daniel Szafrir. 2018. Communicating robot motion intent with augmented reality. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 316–324.
- [78] Tom Williams, Daniel Szafrir, Tathagata Chakraborti, and Heni Ben Amor. 2018. Virtual, augmented, and mixed reality for human-robot interaction. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 403–404.
- [79] Heinz Wimmer and Josef Perner. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children’s understanding of deception. *Cognition* 13, 1 (1983), 103–128.
- [80] Alan F. T. Winfield. 2018. Experiments in artificial theory of mind: From safety to story-telling. *Front. Robot. AI* 5 (2018), 75.
- [81] Guang-Zhong Yang, Jim Bellingham, Pierre Dupont, Peer Fischer, Luciano Floridi, Robert Full, Neil Jacobstein, Vijay Kumar, Marcia McNutt, Robert Merrifield, et al. 2018. The grand challenges of Science Robotics. *Sci. Robot.* 3, 14 (2018), 7650.
- [82] Yu Zhang, Sarath Sreedharan, Anagha Kulkarni, Tathagata Chakraborti, Hankz Hankui Zhuo, and Subbarao Kambhampati. 2017. Plan explicability and predictability for robot task planning. In *Proceedings of the IEEE International Conference on Robotics and Automation*. IEEE, 1313–1320.
- [83] Tom Ziemke. 2020. Understanding robots. *Sci. Robot.* 5, 46 (2020), eabe2987.
- [84] Tom Ziemke and Noel E. Sharkey. 2001. A stroll through the worlds of robots and animals: Applying Jakob von Uexküll’s theory of meaning to adaptive robots and artificial life. *Semiotica* 134, 1/4 (2001), 701–746.

Received December 2019; revised October 2020; accepted March 2021