The Intentional Stance Toward Robots: Conceptual and Methodological Considerations

Sam Thellman (sam.thellman@liu.se)
Department of Computer & Information Science, Linköping University
Linköping, Sweden

Tom Ziemke (tom.ziemke@liu.se)
Department of Computer & Information Science, Linköping University
Linköping, Sweden

Abstract

It is well known that people tend to anthropomorphize in interpretations and explanations of the behavior of robots and other interactive artifacts. Scientific discussions of this phenomenon tend to confuse the overlapping notions of folk psychology, theory of mind, and the intentional stance. We provide a clarification of the terminology, outline different research questions, and propose a methodology for making progress in studying the intentional stance toward robots empirically.

Keywords: human-robot interaction; social cognition; intentional stance; theory of mind; folk psychology; false-belief task

Introduction

The use of folk psychology in interpersonal interactions has been described as "practically indispensable" (Dennett, 1989, p. 342), and its predictive power has been proclaimed to be "beyond rational dispute" (Fodor, 1987, p. 6). The emergence of interactive technologies, such as computers and robots, has sparked interest in the role of folk psychology in human interactions with these systems. For example, John McCarthy stated: "It is perhaps never logically required even for humans, but expressing reasonably briefly what is actually known about the state of a machine in a particular situation may require ascribing mental qualities or qualities isomorphic to them" (McCarthy, 1979, p. 2). The usefulness of folk psychology however does not extend to interaction with all artifacts, and it does not necessarily extend to all kinds of interactions with robots. Although the prevalence of folk-psychological interpretation of robot behavior might be considered as being beyond dispute, its predictive power – i.e., the *usefulness* of taking the intentional stance toward robots – remains largely unassessed.

Researchers from diverse fields have explored people's folk-psychological theories about emerging robotic technologies, such as humanoid robots and autonomous vehicles. For example, Krach et al. (2008) and Chaminade et al. (2012) explored the neural activity of persons engaged in interactive games with robots. Waytz et al. (2014) showed that people's ascriptions of mental states to an autonomous vehicle affected their willingness to trust it. Thellman et al. (2017), Petrovych et al. (2018), and de Graaf and Malle (2018, 2019) investigated whether people judge distinctively human behaviors as intentional when exhibited by robots. Terada et al. (2007) asked people directly about whether they adopted the intentional stance toward a robot. Marchesi et al. (2018) developed a questionnaire-based method specifically for assessing whether people adopt the intentional stance toward robots. These studies all provide insight into people's folk-psychological theories about robots. However, none of them assessed how such theories affect people's predictions of behavior to shed light on the usefulness of taking the intentional stance in interactions with robots.

Moreover, research that has so far explicitly addressed the intentional stance toward robots in many cases conflated the intentional stance with overlapping but different notions, such as folk psychology and theory of mind. In particular, the question whether it is useful for people to predict robot behavior by attributing it to mental states (what we in the present paper will call "the intentional stance question") tends to be confounded with whether robots have minds ("the reality question"), whether people think that robots have minds ("the belief question"), and what kinds of mental states people ascribe to robots ("the attribution question"). For example, Chaminade et al. (2012, p. 8) claimed that participants in their experiments did not adopt the intentional stance when interacting with a robot as opposed to a person based on having "[manipulated] participants' belief about the intentional nature of their opponent" (thereby confounding the attribution question with the belief question). Wykowska et al. (2015, p. 768) stated that "it seems indeed very important to know whether the observed entity is an agent with a mind, and thus, whether the entity's behavior provides some social meaningful content" (confounding the attribution question with the reality question). Wiese et al. (2012, p. 2) stated that "adopting the intentional stance is based on a decision as to whether or not an observed agent is capable of having intentions" (confounding the intentional stance question with the belief question).

In view of these confusions, we aim to provide a clarification of the terminology and different research questions related to the folk psychology about robots in general and the intentional stance toward robots in particular. We also discuss in more detail how (not) to approach research questions specifically targeted at the intentional stance toward robots.

Basic Terminology

We here review Griffin and Baron-Cohen's (2002) distinction between *folk psychology*, *theory of mind*, and *the intentional*

stance and relate these overlapping but different notions to the literature surrounding the role of folk psychology in interactions with robots.

Folk psychology about robots

The notion of folk psychology (also known as belief-desire psychology, naïve or intuitive psychology, or commonsense psychology) broadly encompasses all mind-related theories that people have about themselves and others (Griffin & Baron-Cohen, 2002). This includes views about intentional, content-bearing, representational states (beliefs, desires, intentions, hunches, etc.) as well as phenomenal states (e.g., undirected anxieties, feelings and pain), traits, dispositions, and empirical generalizations such as that *people who are tired are generally irritable*, or – as in the context of folk psychology about robots – the cultural platitude that *robots do not have minds* (Fiala et al., 2014).

Research on people's folk-psychological theories about robots in general (as opposed to specific robots) has been pursued in part because of the societal (e.g., political, legal, or ethical) consequences that such theories might have. For example, European citizens' views on and acceptance of emerging robotic technologies, and their use in different areas of society, have been monitored in extensive surveys by the European Commission (2012, 2015). Ethically motivated research has targeted robot abuse, killer robots, robots in elderly care, child-robot interaction, and sex robots (for an overview, see Lin et al., 2014).

Theory of (robot) mind

Theory of mind refers more narrowly to the ability to attribute the behavior of *specific* others or oneself to underlying mental states, in particular intentional states, such as beliefs and desires, that are perceived to have a causal role in behavior (Griffin & Baron-Cohen, 2002).

People's views about the mental attributes of specific robots are frequently probed for the purpose of evaluating human-robot interactions. Examples of such measures are the Godspeed Questionnaire Series (Bartneck et al., 2009) and the Robotic Social Attributes scale (Carpinella et al., 2017). To the best of our knowledge, these measures have so far not been used in conjunction with measures of people's ability to predict the behavior of specific robots in the context of human-robot interaction research.

The intentional stance toward robots

The intentional stance refers to the *use* of intentional constructs (the beliefs, desires, intentions, etc., that are part of people's folk-psychological theories) as an interpretative strategy or framework to predict the behavior of specific others (Griffin & Baron-Cohen, 2002)[1]. The intentional stance is sometimes mistakenly equated with folk psychology. Dennett (1991) describes the intentional stance as "the craft" of folk psychology and distinguishes it from "the theory" itself. The intentional stance concerns what people *do* with folk psychology (i.e., predict and explain behavior using intentional constructs); folk psychology, in Dennett's view, refers to how we talk about what we do.

Although there seems to be a general consensus in the literature concerning the meaning of "intentionality" as denoting the distinguishing characteristic of certain mental phenomena of being "about" or "directed at" something as an object (Brentano, 1874/2012), some authors have treated it as a biological property (e.g., Searle, 1980; Varela, 1997; Ziemke, 2016) whereas others have refrained from doing so (e.g., Dennett, 1989; McCarthy, 1979). It is also important to recognize that intentionality is a separate notion from having certain intentions. Intentionality is a property of a specific set of mental states, namely intentional mental states. This set includes intentions, but also beliefs, desires, hopes, fears, hunches, and so on. Searle (2008, pp. 85–86) noted that the English translation of the German words for intentionality and intention, "Intentionalität" and "Absicht", are confusingly similar, stating that "we have to keep in mind that in English intending is just one form of intentionality among many".

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 that are "usefully and voluminously predictable from the intentional stance" (Dennett, 2009, p. 339). Humans are the most obvious example of intentional systems because human behavior is generally successfully predicted from the intentional stance but not from other modes of interpretation. The label "intentional system" is not restricted to humans, but it also does not extend to all non-human 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 degrees and *knows* that it has fallen during the night, the use of such folk-psychological interpretations does not add predictive value *above and beyond* the corresponding non-psychological interpretation. In the words of John McCarthy (1979, p. 11), "ascribing beliefs to simple thermostats is unnecessary for the study of thermostats, because their operation can be well understood without it". In contrast, the moves of a chess-playing computer are, according to Dennett (1971), practically inaccessible to prediction from any other interpretative mode than the intentional stance.

It is reasonable to conjecture, given the complex behavior and social situatedness (Lindblom & Ziemke, 2003) of emerging robotic technologies, that taking the intentional stance might turn out to be crucial in many cases of human-robot interaction (Hellström & Bensch, 2018; Schellen & Wykowska, 2019; Thill & Ziemke, 2017; Vernon et al.,

[1] As noted by Griffin and Baron-Cohen (2002), the intentional stance theory (also known as intentional systems theory; Dennett, 2009) is both Dennett's take on the role of folk psychology in social interactions and on what intentional states really are. These two components can be considered separately (as in this paper); for example, one might agree with Dennett's claims about the role of the intentional stance in social interaction without subscribing to his views about the reality of ascribed mental states.

2016). However, although there is a growing body of evidence that people take the intentional stance toward robots, the usefulness of doing so remains largely unassessed. Hence, the central question in the context of the intentional stance toward robots is the extent to which the behavior of robots is usefully predicted from the intentional stance. The usefulness of the intentional stance toward robots presumably depends on a number of unknown factors, possibly related to the person interacting with the robot, the interaction context, and the robot in question. Answers to the intentional stance question might thus range from "the intentional stance is a practically dispensable mode of interpretation for predicting robot behavior" (cf. thermostat) to "the intentional stance is practically indispensable for predicting robot behavior" (cf. chess-playing computer), depending on these factors. Research into the usefulness of taking the intentional stance toward robots may also reveal unique social cognitive challenges associated with taking the intentional stance specifically toward robots (e.g., compare inferring what a robot vs. a person can perceive in a given situation in order to predict it's behavior), some of which may be universally present in human-robot interactions.

Four Distinct Research Questions

We have attempted to clarify some of the basic terminology surrounding the intentional stance toward robots. We also identified the central question about the intentional stance toward robots as concerning its usefulness for predicting robot behavior. We now move on to distinguish this question from three overlapping but separate research questions that appear frequently in the literature surrounding the intentional stance toward robots.

The reality question: Do robots have minds?

Questions such as "Do robots have minds?" and "Can machines think?" concern the nature or reality of the mental states of robots and other machines. We here collectively refer to such questions as different formulations of *the reality question*. The reality question is clearly independent from people's beliefs about it, and presumably also from people's disposition to predict and explain robot behavior based on mental state ascriptions (and the potential usefulness of doing so). While it seems plausible that ontological "discoveries" about the minds of robots may have a significant impact on how people relate to and interact with robots, there is no apparent reason to believe that they would affect people's predictions of robot behavior in interactions. What matters for the purpose of predicting behavior, it seems, is how people conceptualize behavior, and not the correspondence of those conceptualizations to reality. For example, Heider (1958, p. 5) noted: "If a person believes that the lines in his palm foretell his future, this belief must be taken into account in explaining certain of his actions". Hence, the reality question is conceptually distinct from questions regarding people's attributions and beliefs about the mental states of robots.

The belief question: Do people think that robots have minds?

People's views on the reality of the mental states of robots are part of folk psychology. As stated in the previous section, it is difficult to foresee how (if at all) such considerations affect people's predictions of robot behavior, regardless if they spring from collective scientific discovery or personal belief. Clearly, a person might attribute the behavior of a robot to mental states without necessarily committing to any ontological position about the reality of those mental states. Indeed, people commonly ascribe mental states to cartoon characters and animated geometric figures (Heider & Simmel, 1944). When, for example, we see Donald Duck angrily chasing chipmunks Chip and Dale because they are stealing his popcorn, we know that Donald, Chip, and Dale do not really have mental states, but we attribute their behavior to mental states nevertheless (Ziemke et al., 2015). As stated by Airenti (2018, p. 10), "anthropomorphism is independent of the beliefs that people may have about the nature and features of the entities that are anthropomorphized". There is to our knowledge no evidence that people's beliefs about the reality of the mental states of robots – or of cartoon characters, thermostats, or fellow humans – affect their disposition or ability to predict behavior. It does not seem to matter, for the purpose of predicting the behavior of an agent, whether the person interpreting the behavior of the agent in question believes that the agent *really* has mental states. *The belief question*, therefore, must be treated as distinct from questions concerning people's ascriptions of mental states to robots as well as the reality question.

The attribution question: What kinds of mental states do people ascribe to robots?

There is now an abundance of evidence that people commonly predict and explain the behavior of robots based on attributing it to underlying intentional states. The assumption that they do is arguably even built into many of the methods that are used to evaluate social human-robot interactions, whereby researchers explicitly ask people to evaluate mental properties of robots. The if-question in "Do people take the intentional stance toward robots?" has thus already been answered in the affirmative. Considerably less is known about what we for the present purposes call *the attribution question*, namely what kinds of mental states people ascribe to robots. The lack of knowledge about the attribution question does not stem from a lack of research effort but, at least in part, from issues in the methodology adopted to tackle the attribution question.

There is so far little agreement about what kinds of mental states people ascribe to robots. Gray, Gray and Wegner (2007) found that people tend to attribute the behavior of robots to mental states related to agency (e.g., memory, planning, and thought) but not subjective experience (e.g., fear, pain, and pleasure). Sytsma and Machery (2010) found, in contrast, that people refrain from attributing subjective states

that have hedonic value for the subject, that is, valenced states (e.g., feeling pain and anger) as opposed to unvalenced states (e.g., smelling a banana or seeing red). Buckwalter and Phelan (2013) further showed that people's tendency to attribute (or not) experiential or valenced states depends on the described function of the robot. Fiala et al. (2014) found that respondents in their experiments – when allowed to choose between different ways of describing the capabilities of a robot (e.g., the robot "identified the location of the box" vs. "knew the location of the box") – preferred not to attribute mental states at all. The authors noted that responses to questions about the mental states of robots are influenced by a wide variety of factors, including the apparent function of the robot, the way in which the question is asked, and cultural platitudes about robots.

In sum, it seems problematic to identify what kinds of mental states people ascribe to robots by asking them directly. Part of the problem, we believe, is that such questions are ambiguously open to interpretation as regarding the reality of the mental states of robots. As pointed out previously, people tend to predict and explain robot behavior with reference to mental states without reflecting on the reality of those states. Thus, when asked directly, a person might deny that a robot has a mind, despite having previously attributed mind to it upon being asked to describe its behavior (Fussell et al., 2008).

The intentional stance question: Is it useful for people to predict robot behavior by attributing it to mental states?

The usefulness of predicting robot behavior by attributing it to mental states is not a pre-given. *The intentional stance question* is therefore distinct from the attribution question. The ability to predict behavior based on the intentional stance is also, as evidenced by studies on mental state attribution from Heider and Simmel (1944) and onwards, independent from the reality of the attributed mental states and from people's beliefs about them.

Although the prevalence of people taking the intentional stance toward robots might be considered as beyond dispute, its predictive power – that is, the usefulness of doing so – remains largely unassessed. Hence, the central question in the context of the intentional stance toward robots is to what extent the behavior of robots is usefully predicted from the intentional stance. Other questions of potential interest concern causes of predictive (mis)judgment from the intentional stance toward robots, how misjudgment can be reduced, and potential effects of taking the intentional stance toward robots on human cognition (e.g., cognitive load).

Measures of the Intentional Stance

If one wants to investigate whether the intentional stance is useful as an interpretative framework for predicting robot behavior, then one must, at the very least, measure people's predictions of behavior and ensure that those predictions stem from specific attributed mental states. Very few previous studies concerned with the intentional stance toward robots employed such measures (one exception is Sciutti et al., 2013). In this section, we review established experimental paradigms in interpersonal psychology that accomplish measuring effects of mental state attribution on behavior prediction, namely explicit and implicit false-belief tasks and anticipatory gaze tasks.

Explicit measures

The standard false-belief task (sometimes referred to as the "Sally–Anne test" or the location-change false-belief test) was outlined by Dennett (1978) in a commentary to Premack and Woodruff's seminal paper "Does the chimpanzee have a theory of mind?". This was later turned into an experimental paradigm in which a human study participant must attribute a false belief to an agent in order to predict its behavior (Wimmer & Perner, 1983). In the experiment, the participant is made aware that an agent observes a certain state-of-affairs x. Then, in the absence of the agent the participant witnesses an unexpected change in the state-of-affairs from x to y. The participant now knows that y is the case and also knows that the agent still (falsely) believes that x is the case. After this, the participant is asked to predict how the agent will behave in some circumstance, given its false belief about the state-of-affairs. If the participant fails to predict the behavior of the agent, this can be directly attributed to a failure of the participant to ascribe a false belief to the agent. Frith and Frith (1999, p. 1692) commented on the strength of the false-belief task: "To predict what a person will do on the basis of a true belief is not a sufficiently stringent test [of the ability to take the intentional stance], since here the belief coincides with reality, and it's hard to tell whether the action is governed by physical reality or mental state. In everyday life, beliefs rather than reality determine what people do, and false beliefs play an important role".

False-belief tasks have primarily been used to test for the possession of a theory of mind (e.g., Baron-Cohen, Leslie & Frith, 1985). However, they can also be used to explore the relative difficulty of reasoning about others' beliefs (Bloom & German, 2000). We argue that the false-belief task is a suitable paradigm for assessing the usefulness of the intentional stance toward robots because it enables measuring the extent to which a person's mental state ascriptions to a specific robot are conducive to predicting its behavior. Hence, false-belief tasks would be used in the context of human-robot interaction studies not to test for a person's *possession* of a theory of a specific robot's mind but for the successful or unsuccessful *use* of such theories in interactions with robots.

Concerns have been raised previously in the theory of mind literature about whether the explicit formulation of false-belief questions might impute folk-psychological theory to the task participant or affect his or her disposition to ascribe mental states. In some false-belief experiments, participants were asked questions, such as "Where does the agent *believe/think* that the object is now?", which explicitly suggest that the agent possesses beliefs or thoughts. Other experi-

ments used questions such as "Where will the agent look for the object?" which implicitly suggest the possession of beliefs or thoughts. However, an extensive meta-study of theory of mind research on children showed that the type of question (e.g., explicit vs. implicit statements of belief) provided to participants did not significantly affect participants' success in the false-belief task (Wellman, Cross & Watson, 2001). This finding can be taken as supporting the view expressed by Dennett that "whether one calls what one ascribes to the computer beliefs or belief-analogues or information complexes or Intentional whatnots makes no difference to the nature of the calculation one makes on the basis of the ascription" (Dennett, 1971, p. 91). Regardless of this meta-analytic finding, researchers concerned with the risk of imputing folk-psychological theories about robots to study participants, in the context of false-belief tasks, can employ implicit intentional stance measures. In the following section we review two such measures: implicit false-belief tasks and goal-directed anticipatory gaze tasks.

Implicit measures

Implicit false-belief tasks employ non-verbal measures to assess people's behavior predictions (for an overview, see Schneider & Slaughter, 2015). Using implicit measures, the intentional stance can be investigated by recording anticipatory gaze behavior (Clements & Perner, 1994) or reaction times (Kovács, Téglás & Endress, 2010), even without instructions to predict behavior or providing questions about the mental states of agents (Kovács, Téglás & Endress, 2010; Schneider et al., 2012). Implicit measures also provide an opportunity to investigate the potential effort involved in tracking the beliefs of robots whose sensory perspectives significantly differ from the human case.

Goal-directed anticipatory gaze tasks represent another way to measure the intentional stance toward robots. Using an anticipatory gaze paradigm, Sciutti et al. (2013) showed that people shift their gaze toward perceived "goals states" of robot actions prior to the execution of the actions themselves. One limitation of this paradigm is that it is not always possible to infer which gaze behaviors are anticipatory gazes (and therefore reflect goal ascriptions) and which are not. As such, goal-directed anticipatory gaze measures might not be as strong a measure of the intentional stance as false-belief tasks. Nevertheless, studying goal ascription through anticipatory gaze measures might be suitable as a complement to studying belief ascription using false-belief tasks.

Conclusion

We have attempted to clarify the difference between three overlapping concepts that are used (in many cases confusedly) in the literature surrounding the intentional stance toward robots: folk psychology, theory of mind, and the intentional stance. The central question in research on the intentional stance toward robots was identified as the extent to which the intentional stance is a useful (and potentially even indispensable) interpretative strategy or framework for predicting behavior in interactions with robots. We argued that this question is distinct from questions regarding the reality of the mental states of robots, people's beliefs about the mental states of robots, and what kinds of mental states people ascribe to robots. We also established a "methodological criterion" for investigating the usefulness of the intentional stance toward robots: the measurement of people's predictions of robot behavior and reliable inference that those predictions stem from specific attributed mental states. Last, but not least, we identified explicit and implicit false-belief tasks and anticipatory gaze tasks as fulfilling these criteria, thereby constituting a promising experimental paradigm for future empirical investigations of the intentional stance toward robots.

The ability to infer the intentional states (beliefs, desires, etc.) of robots is presumably in many cases crucial to the successful prediction of robot behavior and, consequently, to well-functioning and socially acceptable human-robot interaction (Hellström & Bensch, 2018; Schellen & Wykowska, 2019; Thill & Ziemke, 2017; Vernon et al., 2016). However, continuously tracking changes in the intentional states of robots as interactions unfold represents a potentially difficult and demanding challenge to humans: robots have different "perspectives" on or sensorimotor couplings with the world than humans. Consider the task of simultaneously navigating interactions with three different types of robots in a crowded environment (e.g., a busy street): the first robot can detect objects behind humanly opaque structures such as walls, vehicles, or humans; the second robot cannot see through glass; and the third robot is sensory-equivalent to most humans. How do humans fare in an interaction scenario like this? We propose that taking the intentional stance toward robots must in some cases be more difficult (in terms of predictive accuracy) and demanding (e.g., in terms of cognitive load; Sweller, 1988) than taking the intentional stance toward humans, and view this as a hypothesis worthwhile exploring in the context of human-robot interaction research. In particular, we speculate that people employ a reasoning heuristic which can be described as "anthropocentric anchoring and adjustment", consistent with the accounts in Epley et al. (2004) and Nickerson (1999) but where people adopt the perspective of specific robots by serially adjusting from their own (human) perspective.

Another question relevant to the intentional stance toward robots is the extent to which its usefulness or predictive power can be improved by providing information about the capabilities and limitations of robots prior to interactions. People base their estimations of the knowledge of robots partly on their assumptions about people (Kiesler, 2005). People's knowledge estimations of robots have been shown to be affected by the physical attributes of robots (Powers & Kiesler, 2006) and information about the robot given beforehand, such as robot gender (Powers et al., 2005) or country of origin (Lee et al., 2005). However, it has to our knowledge not yet been investigated whether providing information or manipulating

social cues can improve the accuracy with which people predict the behavior of a robot. Would prior knowledge about the sensory capabilities of the three types of robots in the example above help people interact with them? This is another question worthwhile exploring in studying the intentional stance toward robots.

We believe that cognitive science has important contributions to make in the continued exploration of the role of folk psychology in human interaction with robots, especially in the development of appropriate methodological approaches to investigating the intentional stance toward robots. As suggested in this paper and elsewhere, the intentional stance can be a confusing concept (Griffin & Baron-Cohen, 2002) and a difficult phenomenon to measure, perhaps especially in the context of interactions with robots (Schellen & Wykowska, 2019). In the folk psychology about robots, robots might not have *real* minds but have *attributed* minds nevertheless, and as for the science of mind, the jury is still out regarding the extent to which mind possession and mind attribution go hand-in-hand in the case of robots (cf. Dennett, 1989; Fodor, 1987; Searle, 1980). We therefore hope that the conceptual clarifications and methodological proposals presented here will pave the way for fruitful research on the intentional stance toward robots.

Acknowledgments

The authors would like to thank Fredrik Stjernberg, Robert Johansson, and members of the Cognition & Interaction Lab at Linköping University for valuable input on the ideas presented in this paper.

References

Airenti, G. (2018). The development of anthropomorphism in interaction: Intersubjectivity, imagination and theory of mind. *Frontiers in Psychology*, *9*, 2136.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a theory of mind? *Cognition*, *21*(1), 37–46.

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Int. J. Soc. Robot.*, *1*(1), 71–81.

Bloom, P., & German, T. P. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition*, *77*(1), B25–B31.

Brentano, F. (2014). *Psychology from an empirical standpoint*. Routledge. (Original work published 1874).

Buckwalter, W., & Phelan, M. (2013). Function and feeling machines: a defense of the philosophical conception of subjective experience. *Philos. Stud.*, *166*(2), 349–361.

Carpinella, C. M., Wyman, A. B., Perez, M. A., & Stroessner, S. J. (2017). The robotic social attributes scale (rosas): Development and validation. In *Proc. 2017 ACM/IEEE Int. Conf. on Human-Robot Interaction* (pp. 254–262).

Chaminade, T., Rosset, D., Da Fonseca, D., Nazarian, B., Lutscher, E., Cheng, G., & Deruelle, C. (2012). How do we think machines think? An fMRI study of alleged competition with an artificial intelligence. *Frontiers in Human Neuroscience*, *6*, 103.

Clements, W. A., & Perner, J. (1994). Implicit understanding of belief. *Cognitive Development*, *9*(4), 377–395.

de Graaf, M., & Malle, B. F. (2018). People's judgments of human and robot behaviors: A robust set of behaviors and some discrepancies. In *Comp. 2018 ACM/IEEE Int. Conf. on Human-Robot Interaction* (pp. 97–98).

de Graaf, M. M., & Malle, B. F. (2019). People's explanations of robot behavior subtly reveal mental state inferences. In *Proc. 2019 ACM/IEEE Int. Conf. on Human-Robot Interaction* (pp. 239–248).

Dennett, D. C. (1971). Intentional systems. *The Journal of Philosophy*, *68*(4), 87–106.

Dennett, D. C. (1978). Beliefs about beliefs [P&W, SR&B]. *Behavioral and Brain Sciences*, *1*(4), 568–570.

Dennett, D. C. (1989). *The intentional stance*. MIT press.

Dennett, D. C. (1991). Two contrasts: folk craft versus folk science, and belief versus opinion. In J. D. Greenwood (Ed.), *The future of folk psychology: Intentionality and cognitive science* (pp. 135–148). Cambridge University Press.

Dennett, D. C. (2009). Intentional systems theory. *The Oxford handbook of philosophy of mind*, 339–350.

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *J. Pers. Soc. Psychol.*, *87*(3), 327.

European Commission. (2012). *Special eurobarometer 382: Public attitudes towards robots.*

European Commission. (2015). *Special eurobarometer 427: Autonomous systems.*

Fiala, B., Arico, A., & Nichols, S. (2014). You, robot. In E. Machery & E. O'Neill (Eds.), *Current controversies in experimental philosophy.* Routledge.

Fodor, J. A. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind* (Vol. 2). MIT press.

Frith, C. D., & Frith, U. (1999). Interacting minds–a biological basis. *Science*, *286*(5445), 1692–1695.

Fussell, S. R., Kiesler, S., Setlock, L. D., & Yew, V. (2008). How people anthropomorphize robots. In *2008 ACM/IEEE Int. Conf. on Human-Robot Interaction* (pp. 145–152).

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619–619.

Griffin, R., & Baron-Cohen, S. (2002). The intentional stance: Developmental and neurocognitive perspectives. In A. Brook & D. Ross (Eds.), *Daniel Dennett* (pp. 83–116). Cambridge University Press.

Heider, F. (1958). *The psychology of interpersonal relations*. New York: Whiley.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *Am. J. Psychol.*, *57*(2), 243–259.

Hellström, T., & Bensch, S. (2018). Understandable robots-what, why, and how. *Paladyn*, *9*(1), 110–123.

Kiesler, S. (2005). Fostering common ground in human-

robot interaction. In *2005 IEEE Int. Workshop on robot and human interactive communication* (pp. 729–734).

Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science*, *330*(6012), 1830–1834.

Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., & Kircher, T. (2008). Can machines think? Interaction and perspective taking with robots investigated via fMRI. *PloS one*, *3*(7), e2597.

Lee, S.-l., Lau, I. Y.-m., Kiesler, S., & Chiu, C.-Y. (2005). Human mental models of humanoid robots. In *Proc. 2005 IEEE Int. Conf. Robot. Autom.* (pp. 2767–2772).

Lin, P., Abney, K., & Bekey, G. A. (2014). *Robot ethics: the ethical and social implications of robotics*. The MIT Press.

Lindblom, J., & Ziemke, T. (2003). Social situatedness of natural and artificial intelligence: Vygotsky and beyond. *Adaptive Behavior*, *11*(2), 79–96.

Marchesi, S., Ghiglino, D., Ciardo, F., Perez-Osorio, J., Baykara, E., & Wykowska, A. (2019). Do we adopt the intentional stance toward humanoid robots? *Frontiers in Psychology*, *10*.

McCarthy, J. (1979). Ascribing mental qualities to machines. In M. Ringle (Ed.), *Philosophical perspectives in artificial intelligence.* Humanities Press.

Nickerson, R. S. (1999). How we know – and sometimes misjudge – what others know: Imputing one's own knowledge to others. *Psychol. Bull.*, *125*(6), 737–759.

Petrovych, V., Thellman, S., & Ziemke, T. (2018). Human interpretation of goal-directed autonomous car behavior. In *Proc. 40th Annual Cognitive Science Society Meeting* (pp. 2235–2240). Madison, WI.

Powers, A., & Kiesler, S. (2006). The advisor robot: Tracing people's mental model from a robot's physical attributes. In *Proc. 1st ACM SIGCHI/SIGART Conf. on Human-Robot Interaction* (pp. 218–225).

Powers, A., Kramer, A. D., Lim, S., Kuo, J., Lee, S.-l., & Kiesler, S. (2005). Eliciting information from people with a gendered humanoid robot. In *2005 IEEE Int. Workshop on Robot and Human Interactive Communication* (pp. 158–163).

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515–526.

Schellen, E., & Wykowska, A. (2019). Intentional mindset toward robots – open questions and methodological challenges. *Frontiers in Robotics and AI*, *5*, 139.

Schneider, D., Bayliss, A. P., Becker, S. I., & Dux, P. E. (2012). Eye movements reveal sustained implicit processing of others' mental states. *J. Exp. Psychol. Gen.*, *141*(3), 433.

Schneider, D., Slaughter, V. P., & Dux, P. E. (2015). What do we know about implicit false-belief tracking? *Psychonomic Bulletin & Review*, *22*(1), 1–12.

Sciutti, A., Bisio, A., Nori, F., Metta, G., Fadiga, L., & Sandini, G. (2013). Robots can be perceived as goal-oriented agents. *Interaction Studies*, *14*(3), 329–350.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424.

Searle, J. R. (2008). *Mind, language and society: Philosophy in the real world*. Basic books.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285.

Sytsma, J., & Machery, E. (2010). Two conceptions of subjective experience. *Philos. Stud.*, *151*(2), 299–327.

Terada, K., Shamoto, T., Mei, H., & Ito, A. (2007). Reactive movements of non-humanoid robots cause intention attribution in humans. In *2007 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems* (pp. 3715–3720).

Thellman, S., Silvervarg, A., & Ziemke, T. (2017). Folk-psychological interpretation of human vs. humanoid robot behavior: Exploring the intentional stance toward robots. *Frontiers in Psychology*, *8*, 1962.

Thill, S., & Ziemke, T. (2017). The role of intentions in human-robot interaction. In *Proc. 2017 ACM/IEEE Int. Conf. on Human-Robot Interaction* (pp. 427–428).

Varela, F. J. (1997). Patterns of life: Intertwining identity and cognition. *Brain and cognition*, *34*(1), 72–87.

Vernon, D., Thill, S., & Ziemke, T. (2016). The role of intention in cognitive robotics. In A. Esposito & L. C. Jain (Eds.), *Toward Robotic Socially Believable Behaving Systems – Volume I* (pp. 15–27). Springer.

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *J. Exp. Soc. Psychol.*, *52*, 113–117.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, *72*(3), 655–684.

Wiese, E., Wykowska, A., Zwickel, J., & Müller, H. J. (2012). I see what you mean: How attentional selection is shaped by ascribing intentions to others. *PloS one*, *7*(9), e45391.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103–128.

Wykowska, A., Kajopoulos, J., Obando-Leitón, M., Chauhan, S. S., Cabibihan, J.-J., & Cheng, G. (2015). Humans are well tuned to detecting agents among non-agents: examining the sensitivity of human perception to behavioral characteristics of intentional systems. *Int. J. Soc. Robot.*, *7*(5), 767–781.

Ziemke, T. (2016). The body of knowledge: on the role of the living body in grounding embodied cognition. *Biosystems*, *148*, 4–11.

Ziemke, T., Thill, S., & Vernon, D. (2015). Embodiment is a double-edged sword in human-robot interaction: Ascribed vs. intrinsic intentionality. In *Proc. 10th ACM/IEEE Human Robot Interaction Conference* (pp. 1–2).