Understanding Social Robots: Attribution of Intentional Agency to Artificial and Biological Bodies

Tom Ziemke
Linköping University
Cognition & Interaction Lab
Human-Centered Systems Division
Department of Computer and Information Science
tom.ziemke@liu.se

Abstract Much research in robotic artificial intelligence (AI) and Artificial Life has focused on autonomous agents as an *embodied* and *situated* approach to AI. Such systems are commonly viewed as overcoming many of the philosophical problems associated with traditional computationalist AI and cognitive science, such as the grounding problem (Harnad) or the lack of intentionality (Searle), because they have the physical and sensorimotor grounding that traditional AI was argued to lack. Robot lawn mowers and self-driving cars, for example, more or less reliably avoid obstacles, approach charging stations, and so on—and therefore might be considered to have some form of artificial intentionality or intentional directedness. It should be noted, though, that the fact that robots share physical environments with people does not necessarily mean that they are situated in the same perceptual and social world as humans. For people encountering socially interactive systems, such as social robots or automated vehicles, this poses the nontrivial challenge to interpret them as intentional agents to understand and anticipate their behavior but also to keep in mind that the intentionality of artificial bodies is fundamentally different from their natural counterparts. This requires, on one hand, a "suspension of disbelief" but, on the other hand, also a capacity for the "suspension of belief." This dual nature of (attributed) artificial intentionality has been addressed only rather superficially in embodied AI and social robotics research. It is therefore argued that Bourgine and Varela's notion of Artificial Life as the *practice of autonomous systems* needs to be complemented with a *practice of socially interactive autonomous systems*, guided by a better understanding of the differences between artificial and biological bodies and their implications in the context of social interactions between people and technology.

Keywords
Attribution, embodiment, grounding, human–robot interaction, intentionality, observer's grounding problem, social robots

1 Introduction

Thirty years ago, Bourgine and Varela (1992), in their editorial introduction to the proceedings of the first *European Conference on Artificial Life,* referred to Artificial Life research as "the practice of autonomous systems" as well as "the most recent expression of a relatively long tradition of

 Artificial Life 29: 351–366 (2023) https://doi.org/10.1162/artl_a_00404

thought which searches the core of basic cognitive and intelligent abilities in the very capacity for being alive" (p. xi). Their characterization of autonomy is worth quoting at length:

> The guiding intuition to both cybernetic forerunners and current proponents of artificial life is quite similar: the need to understand the class of processes that endow living creatures with their characteristic autonomy. . . . Autonomy in this context refers to their basic and fundamental capacity to assert their existence and to bring forth a world that is significant and pertinent without be[ing] pre-digested in advance. Thus the autonomy of the living is understood here in regards to its actions and to the way it shapes a world into significance. This conceptual exploration goes hand in hand with the design and construction of autonomous agents and suggests an enormous range of applications at all scales, from cells to societies. (Bourgine & Varela, 1992, p. xi)

In the context of this article, the twofold use of the term *autonomy* in this quotation deserves further attention. Bourgine and Varela's reference to the design and construction of autonomous agents implies a distinction between (a) *biological autonomy* as the *phenomenon* that we seek to understand when using Artificial Life—or computational and robotic models in general—for the purpose of scientific modeling and (b) *artificial autonomy* as a *capacity* with which we seek to endow robotic technologies, based on our scientific understanding of its biological counterpart, or rough abstractions thereof.

Nowadays, 30 years later, there are many examples of robotic real-world technologies with some degree of autonomy in the artificial sense, ranging from robot vacuum cleaners to automated vehicles. Robotic lawn mowers, for example, can relatively reliably mow one's lawn, avoid obstacles while doing so, detect when they are about to run out of energy, navigate to their charging stations for recharging, and so on. Although at some abstract level, these behavioral capacities could be likened to those of living systems, the artificial autonomy with which they operate is, of course, fundamentally different from the biological autonomy of even the simplest living systems. For one thing, robotic lawn mowers, as engineered products with limited adaptivity, can only to a small degree—if any—"bring forth a world that is significant and pertinent without be[ing] pre-digested in advance," which Bourgine and Varela (1992, p. xi) considered characteristic for living systems (cf. earlier discussion and Varela et al., 1991). Your robotic lawn mower might, for example, learn to adapt to the specific layout of your garden and, in that sense, go beyond its preprogramming, but clearly its overall functionality is largely determined by its designers and its owner. After all, it is mowing *your* lawn, and *you* are the one who cares if the robot runs out of energy before reaching its charging station. This seems to imply that interpretations of robot lawn mowers—or similar artificial autonomous agents—as intentional agents might be a modern case of what Searle (1980) called "extend[ing] our own intentionality; our tools are extensions of our purposes, and so we find it natural to make metaphorical attributions of intentionality to them" (p. 419). This would also seem to be in line with Varela's notion of intentionality as fundamentally rooted in biological autonomy (e.g., Varela, 1992, 1997).

On the other hand, much recent research in human–robot interaction (HRI) has shown that people tend to interpret and predict the behavior of robots—or other autonomous artifacts, such as automated vehicles—in folk psychological terms (e.g., Perez-Osorio & Wykowska, 2020; Schellen & Wykowska, 2019; Sciutti et al., 2013; Thellman et al., 2022; Thellman et al., 2017). That means that people tend to *attribute* intentional mental states to robotic systems, such as *beliefs* (e.g., in the lawn mower's case, that the battery level is low), *desires* or *goals* (e.g., to keep the battery level above some threshold), and *intentions* (e.g., to approach the charging station to recharge). It should be noted, though, that this does *not* mean that people *necessarily* really believe that the robots in question actually have such folk psychological, humanlike or animallike mental states (Thellman & Ziemke, 2019). In fact, in one of the earliest empirical studies of such attributions, Heider and Simmel (1944) showed that people even attributed mental states to simple, moving geometric shapes, such as squares and triangles—even though presumably nobody thinks that triangles have

such states. Our recent systematic review of 155 empirical studies (Thellman et al., 2022) has shown that mental state attribution to robots is a common phenomenon and that the known consequences include increased predictability, explainability, and trust. In other words, and somewhat oversimplified, people seem to interpret robots as intentional agents because it makes them easier to interact with.

Needless to say, there is an interesting tension here—which this article seeks to explore further—between the view of intentionality as a biological phenomenon and the fact that people frequently *attribute* intentionality to robotic technologies anyway. Hence the intended contribution of this article to the theme of the special issue of which it is a part—"Biology in AI"—can be described as follows: The article does *not* address the biological understanding or inspiration that in many cases contributes to the development of artificial autonomous agents; instead, it addresses the fact that most of the human cognitive mechanisms used to *interpret* intentional agency in the world around us of course stem from our interactions with a broad spectrum of biological autonomous agents, that is, humans and other animals (cf. Urquiza-Haas & Kotrschal, 2015). We therefore tend to interpret robots, virtual agents, and computer game or cartoon characters—but also, in many cases, inanimate moving objects, such as Heider and Simmel's (1944) geometric shapes—*as if* they were living systems. Or, as Black (2014) put it, we "seem to have an innate propensity to see bodies wherever we look" (p. 16), and therefore "consistently anthropomorphise machines, our attempts to conceptualise unfamiliar new artefacts falling back on the most fundamental and sophisticated frameworks for understanding animation we have—those related to the human body" (p. 38).

It should be noted, though, that most of the time, people are of course aware of their attributions being attributions. For cartoon characters like Donald Duck, for example, which are usually not interactive and only appear on a screen or paper—and therefore are easy to consider as "not real"—it might be relatively easy to shake off this intentional interpretation, that is, to recognize that we are attributing intentional agency where maybe there is none. For social robots, on the other hand, which are interactive, physical, and therefore part of the "real world" in some intuitive sense, things might be more complicated. H. Clark and Fischer (2022) have recently referred to this as the *social artifact puzzle*, which they characterize as follows: "It seems self-contradictory, even irrational, for people to hold these two attitudes simultaneously: (a) a willingness to interact with social robots as real people or animals; and (b) a recognition that they are mechanical artifacts" (section 1, para 2).

This article therefore argues that today, when social robots, automated vehicles, and similar interactive autonomous technologies are starting to become a real-world reality for many people, the *practice of autonomous systems*, as Bourgine and Varela (1992) envisioned, with a focus on understanding and modeling biological autonomy, needs to be complemented with what might be called a *practice of socially interactive autonomous systems*, with a focus on the human interpretation of artificial autonomy. Importantly, this needs to be guided by a better understanding of the differences between artificial and biological bodies and their implications in the context of social—or quasi-social—interactions between people and autonomous technologies.

The remainder of this article is structured as follows. Section 2 provides further background by discussing *intentionality* and its relevance to the *grounding problem* and related criticisms of traditional artificial intelligence (AI) as well as embodied AI approaches to overcoming these problems. Section 3 zooms in on current research on social robots and their interpretation in terms of intentional agency, including the *social artifact puzzle*, discussions of *deception* in social robotics, and what I refer to as the *observer's grounding problem*. Section 4, finally, concludes by getting back to the proposal that we need to develop a practice of socially interactive autonomous systems, outlining what the key elements and requirements of such a practice might be.

2 Background: Intentionality and Grounding

Before we move on to social robots in section 3, let me try to quickly recap the (symbol) grounding problem and related criticisms of traditional AI and cognitive science, in particular, their view of

cognition as symbolic computation. There are older criticisms of AI (e.g., Dreyfus, 1972/1979; Taube, 1961), but I will here start with Searle's much-discussed Chinese Room argument from 1980. If you wonder why we need to recap this—after all, aren't these discussions 30–50 years old, and isn't the symbol grounding problem solved anyway? (e.g., Steels, 2008)—then the answer simply is that some of the issues discussed in the context of the human interpretation of social robot behavior are in fact closely related to those "old" arguments. Moreover, Searle's argument might be considered particularly interesting because it addresses the role of *intentionality* in the human interpretation of AI, which is still highly relevant to the human interpretation of social robot behavior. Anyway, I will try to keep this section brief. Before we dive into the discussion of intentionality in AI in section 2.2, though, I will try in section 2.1 to set the scene and structure the discussion by clarifying different notions of intentionality.

2.1 Intentionality

The first thing to note about intentionality is that it is too complex and too controversial a phenomenon to address in much detail here. In fact, one might be tempted to avoid the concept altogether, were it not for the fact that intentionality, independent of philosophical baggage and scientific complexity, simply might be fundamental to the *folk psychology* of how people interpret the world around them and, in particular, how they interpret behavior in terms of intentional agency.[1] As Malle et al. (2004) point out, the concept of intentionality

> brings order to the perception of behavior in that it allows the perceiver to detect structure—intentions and actions—in humans' complex stream of movement . . . [and] supports coordinated social interaction by helping people explain their own and others' behavior in terms of its underlying mental causes. (p. 1)

More specifically, we can roughly distinguish three overlapping aspects or notions of intentionality, corresponding to different ways in which agents can be *directed at* their environment. First, the aspect of *intentionality* discussed most in philosophy of mind is the *aboutness* of internal states (or "representations," for lack of a better term). When your robot lawn mower, for example, turns away from objects it detects, or when it approaches its charging station, this could very well be ascribed to the aboutness of the corresponding internal states. In more philosophical and human-centric terms, Searle (1999) says the following:

> The primary evolutionary role of the mind is to relate us in certain ways to the environment, and especially to other people. My subjective states relate me to the rest of the world, and the general name of that relationship is "intentionality." These subjective states include beliefs and desires, intentions and perceptions. . . . "Intentionality," to repeat, is the general term for all the various forms by which the mind can be directed at, or be about, or of, objects and states of affairs in the world. (p. 85)

Second, central to discussion of intentionality in (social) psychology is the *intentionality of behavior*, which can be illustrated by the fact that in the preceding example, you might say that your robot lawn mower did not approach its charging station by chance but did so *intentionally*, namely, for the purpose of recharging. According to Malle and Knobe (1997), the *folk concept of intentionality* involves several aspects: a desire for a particular outcome, beliefs about an action leading to that outcome, an intention to perform the action, the skill to perform the action, and awareness of fulfilling the intention while performing the action.

1 In Malle et al.'s (2004) words, "if one took a Kantian approach to social cognition, searching for the fundamental concepts without which such cognition is impossible, intentionality would be one of those concepts, on par with space, time and causality in the realm of non-social cognition" (p. 1).

Third, there is what Varela (1992, 1997) referred to as the *biological roots of intentionality*, namely, that *organisms as a whole* are part of and directed toward their environment in the sense that they need to interact with it for the purpose of self-maintenance.

These distinctions also allow us to get a better grip on the nature of the "interesting tension" mentioned in the previous section. From the philosophical perspective of intentionality as aboutness, it is easy to see why one might consider robot lawn mowers—which appear to sense objects, avoid some of them, and approach others—as intentional agents. From the (social) psychology perspective, it might also seem natural to view behaviors like approaching a charging station when energy levels are low as intentional—although there might be disagreement regarding to what extent such systems really fulfill Malle and Knobe's (1997) rather human-centric criteria (e.g., desire or awareness). That current robot lawn mowers lack what Varela (1992, 1997) referred to as the biological roots of intentionality seems relatively clear, but it is less obvious if this does or does not outweigh the other perspectives. Moreover, as discussed in the beginning of the article, Varela seemed perfectly willing to use the term *autonomy* in both the biological sense and an artificial sense, so it might not be far-fetched to think that similar arguments could be made in the case of *intentionality*. Let us therefore move on to the next subsection, which discusses the role of intentionality in discussions of AI.

2.2 Intentionality and AI

In his Chinese Room argument, originally formulated before robots became a focus of attention in much AI research, Searle (1980) imagined himself sitting in a room with the only connections to the outside world being strings of Chinese symbols that he receives as "input" and generates as "output" (e.g., as messages on pieces of paper passed under the door). Searle does not understand any Chinese, so he does not know what the symbols refer to, but he is able to manipulate them with the help of a set of rules (written in some language that Searle does understand), which help him generate output based on the input he receives over time. Searle's argument, intended as a criticism of traditional views of cognition as symbolic computation, was that he, sitting in the room, pretty much does exactly what a computer does when it transforms inputs into outputs according to some program. His point, of course, was that clearly he does not understand any Chinese, in the sense that he does not know to what objects or events in the outside world any of the symbols refer. Nevertheless, people outside the room who speak Chinese might very well be able to attach meaning to those inputs and outputs and therefore might get the impression that whoever or whatever is in the room does understand Chinese—assuming that the program Searle follows is sufficiently good. Searle's conclusion was that the computational view of human cognition is flawed, or must at least be incomplete, because "the operation of such a machine is defined solely in terms of computational processes over formally defined elements" (p. 422), and such "formal properties are not by themselves constitutive of intentionality" (p. 422; in the philosophical sense of aboutness; cf. previous subsection). In Searle's (1990) *Scientific American* article, this was illustrated with a cartoon of a man and a computer facing a symbol on the wall, with the man imagining a horse, whereas the computer only "sees" the symbol.

Harnad's (1990) formulation of the *symbol grounding problem* was based on Searle's arguments but referred to the problem of intentionality as a lack of "intrinsic meaning" in purely computational systems. This, he argued, could be resolved by what he termed *symbol grounding*, that is, the grounding of internal symbolic representations in sensorimotor interactions with the environment. Others have pointed out that the problem applies not only to symbolic AI but also to subsymbolic neural-network approaches or any other computational formalism and have therefore used terms like *representation grounding* (Chalmers, 1992), *concept grounding* (Dorffner & Prem, 1993), and simply the *grounding problem* (e.g., Ziemke, 1997, 1999). Brooks (1993), one of the pioneers of the embodied AI approach (e.g., Brooks, 1991a, 1991b), used the term *physical grounding* and argued that in robots, unlike with the computer programs of traditional AI, "everything is grounded in primitive sensor motor patterns of activation" (p. 154). Other roboticists, more interested in symbolic representation than Brooks, have also formulated the *perceptual anchoring*

problem (Coradeschi & Saffiotti, 2003), which they describe as the intersection of symbol grounding and pattern recognition, or "the problem of connecting, inside an artificial system, symbols and sensor data that refer to the same physical objects in the external world" (p. 85). The general approach of using simulated or robotic autonomous agents has also been referred to as the "artificial life roots" of AI (Steels, 1993) or the "artificial life route" (Steels & Brooks, 1995) to AI, and this is of course what Bourgine and Varela (1992) referred to as the design and construction of autonomous agents in the quotation discussed at the beginning of this article.

Hence, from the embodied AI perspective, you might think that by putting AI into robots, the grounding problem has been resolved. Proponents of this approach have argued that robots are embodied and situated in roughly the same sense that humans and other animals are and therefore should be able to overcome traditional AI's problems with intentionality or intrinsic meaning. The problem of ungrounded representations, it might be argued, is solved through physical grounding and either not having any representations at all (e.g., Brooks, 1991b; Varela et al., 1991) or acquiring grounded internal representations, as Harnad (1990) proposed, in the course of sensorimotor interaction with the external world. Steels (2008), for example, made the explicit claim that "the symbol grounding has been solved," based on his own large-scale Talking Heads experiments, in which populations of robotic agents, in interaction with people, evolved a shared symbolic vocabulary through a large number of one-to-one—robot–robot as well as robot–human—language games involving labeling objects in a shared physical environment. Steels also explicitly stated that Searle was wrong to consider intentionality a biological property, arguing that the key to intentionality instead lies in adaptation in the course of agent–environment and agent–agent interactions. In Harnad's (1989) terms, this type of embodied AI has gone from a *computational functionalism* to a *robotic functionalism*. Zlatev (2001), for example, formulated the latter position explicitly by stating that there is "no good reason to assume that intentionality is an exclusively biological property (pace e.g., Searle)" and that therefore "a robot with bodily structures, interaction patterns and development similar to those of human beings … could possibly recapitulate ontogenesis, leading to the emergence of intentionality, consciousness and meaning" (p. 155).

It should be noted, though, that the robotic functionalism underlying much current research in robotics and AI is of course far from uncontroversial (e.g., Froese & Ziemke, 2009; Sharkey & Ziemke, 2001; Ziemke, 2022; Zlatev, 2002), and consensus is not likely to be achieved any time soon. For the discussion in this article, however, it suffices to acknowledge that there are different positions on this. One way to move the discussion forward—and toward the discussion of human interpretation of social robots in the next section—is to think about this in terms of *situatedness*. As discussed earlier, robots are commonly referred to as *situated* agents. The fact that robots share *physical* environments with people, however, does not necessarily mean that they are situated in the same *perceptual* and *social world* as humans (Ziemke, 2020). This is obvious to anyone whose robot lawn mower has run over hedgehogs or small toys left on the lawn or who has had their kids' LEGO bricks sucked up by a robot vacuum cleaner. Although these robots are *physically situated* in the human world, many of them are not equipped to detect hedgehogs, LEGO bricks, and so on—or do not attach the same meaning to them as people do. This brings us to the discussion of socially interactive robots, for which a certain degree of *social situatedness* and *shared meaning* is crucial.

3 Discussion: Social Robots as a Challenge

Arguably, a crucial prerequisite for individual and social trust in socially interactive robots—once they have entered the real world, where they must be able to interact with people of different ages as well as educational and cultural backgrounds—is that people can interpret and anticipate the behavior of such systems sufficiently reliably to safely interact with them. Hence the design of interactive autonomous systems needs to be informed by a thorough understanding of the mechanisms underlying human social interaction with such systems (Ziemke, 2020). This section argues

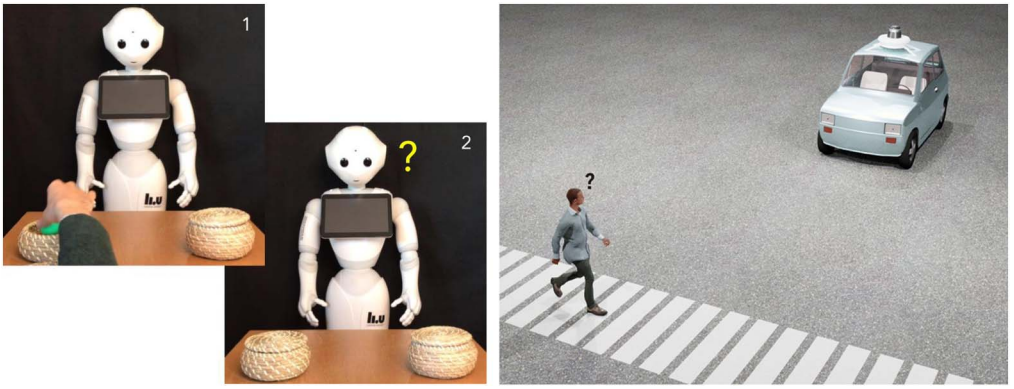

Figure 1. Social robots as intentional agents. (left) An experiment on human mental state attribution to robots (Thellman & Ziemke, 2020), where (1) a humanoid robot observes somebody putting a ball into one of two baskets and (2) a human observer is asked to predict the robot's response to the question of where the ball is. (right) A pedestrian encounters a self-driving car at a crosswalk and wonders if it will let him cross the road (Ziemke, 2020). Reprinted with permission from Ziemke, *Science Robotics*, *5*, eabe2987 (2020). Credit: A. Kitterman/*Science Robotics*.

that this poses significant challenges for the research communities involved as well as for HRI practitioners trying to design social robots and human–robot interactions. Some have even argued that what is needed is a "new ontological category" for "artifacts that appear to think and feel, can be friends, and at least potentially lay some moral claims for kind, fair and just treatment" (Melson et al., 2006, p. 4; cf. H. Clark & Fischer, 2022).

Before we dive deeper into discussions of how people interpret social robots, though, let me try to clarify what exactly I mean by the term *social robot* here. In Thellman and Ziemke (2021), we argued in more detail for the position that what makes a robot or any other interactive autonomous system "social" is that its behavior is most appropriately interpreted in terms of intentional states. This is based on the notion that human social interaction is fundamentally connected to people's understanding of others as intentional agents, as pointed out earlier with reference to Malle et al.'s (2004) view of intentionality as a central component of human social cognition. The term *robot* here is used in a broad sense to refer to any physical artifact that interacts with its environment by means of sensors and motors. Figure 1 provides two examples of what is meant by a social robot. The humanoid robot to the left might be an example of what most readers consider a "social robot," whereas some would probably not necessarily apply that term to the self-driving car on the right. The important point here, though, is that in interactions with people, they are both usually easily—and maybe most appropriately—interpreted in folk psychological terms as *intentional agents* with beliefs, goals, and intentions. The humanoid example illustrates a robotic version of a classical false-belief task from developmental child psychology (cf. Thellman & Ziemke, 2020), where the human observer's task is to predict what beliefs some other agent (in this case, the robot) holds regarding the position of an object. The second example is a scenario in which a pedestrian encounters a self-driving car at a crosswalk and needs to assess what the vehicle's perception of the situation might be and how it might act. The latter example is also particularly interesting because it illustrates a real-world scenario in which (a) there is something more at stake than a mere misinterpretation of observed behavior and (b) the pedestrian is under a certain time pressure.

Given what I have said about people's attribution of intentional agency to, for example, Heider and Simmel's (1944) moving geometric shapes or cartoon characters like Donald Duck, it probably does not come as a surprise to anyone that social robots are of course commonly interpreted as intentional agents as well. There is, however, so far relatively little agreement about what kinds of mental states people ascribe to robots (cf. Thellman & Ziemke, 2019): A classical study by Gray et al. (2007), for example, found that people tend to attribute *agency* to robots (e.g., memory or

planning), but not *experience* (e.g., fear or pleasure). Sytsma and Machery (2010) also found that people refrain from attributing subjective states that have *hedonic* value for the subject, that is, *valenced* states (e.g., feeling pain), as opposed to *unvalenced* states (e.g., seeing red). Buckwalter and Phelan (2013) further showed that people's tendency to attribute experiential or valenced states depends on the described *function* of the robot. Fiala et al. (2014) found that, 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"), people preferred *not* to attribute mental states at all. The authors also 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 is difficult to identify what kinds of mental states people attribute to robots by asking them directly. This is at least partly because such questions are open to interpretation regarding the reality of the mental states of robots (cf. Thellman & Ziemke, 2019). Fussell et al. (2008), for example, found that people might deny that a robot has a mind, despite having previously attributed mind (mental states) when describing its behavior.

Owing to the growing research activity in social robotics in recent years, there also has been a growing interest in both conceptual discussions (e.g., Papagni & Koeszegi, 2021; Perez-Osorio & Wykowska, 2020; Schellen & Wykowska, 2019; Thellman et al., 2022; Thellman & Ziemke, 2019) and empirical investigations of why, when, and how people attribute intentional agency and mental states to robots. Most of the theoretical discourse indicates that people commonly interpret the behavior of robots following similar strategies as when interpreting human or animal behavior. Dennett (1971, 1989) referred to this as the *intentional stance*, which he contrasted with design stance and physical stance as alternative approaches to interpreting observed behavior. However, according to the *cognitive default hypothesis* in studies of human interpretation of animal behavior (Caporael & Heyes, 1997; Guthrie, 1997; Urquiza-Haas & Kotrschal, 2015), anthropomorphism in fact emerges as an automatic response to any sufficiently humanlike behavior or feature, especially when a swift response is required and/or other modes of interpretation are not readily available. Urquiza-Haas and Kotrschal have developed a (neuro)psychological model of how automatic/bottom-up and reflective/top-down mechanisms of physical/embodied and social cognition interact in the human anthropomorphic interpretation of animal behavior. Their hypothesis is that the balance of automatic and reflective cognitive processes depends on phylogenetic distance and on shared morphological and behavioral features.

Regarding empirical investigations of mental state attribution to robots, my coauthors and I recently published a systematic review of 155 studies (Thellman et al., 2022). Most empirical research so far has been concerned with determinants and consequences, that is, the questions of *when* people attribute mental states to robots, and *why*. *Determinants* include *human factors*, such as age and motivation, as well as *robot factors*, such as appearance and behavior. *Consequences* include increased predictability, explainability, and trust but also increases in cognitive drain and moral concern. Relatively little, however, is known about the *how*, that is, the mechanisms underlying such mental state attributions. A multidimensional account similar to Urquiza-Haas and Kotrschal's (2015) model might also be a promising starting point for investigations of the neural and psychological mechanisms involved in social human–robot interactions, especially given that one of the findings of our review (Thellman et al., 2022) is that there is a "computer < robot < human" pattern in the tendency to attribute mental states[2] that appears to be moderated by the presence of socially interactive behavior. However, an account of the human interpretation of autonomous technologies is likely to require additional dimensions and complexity, given that (a) autonomous technologies are a "moving target" (i.e., unlike cats or cows, robots today are not what they were 10 years ago) and (b) most people presumably do not yet have firmly established categories.

2 That means that people are more likely to attribute mental states to humans than to robots and more likely to attribute them to robots than to computers, which is in line with Urquiza-Haas and Kotrschal's (2015) arguments regarding the role of shared morphological and behavioral features.

As briefly mentioned earlier, H. Clark and Fischer (2022) have recently characterized the state of the art regarding human interpretation of social robots as puzzling:

> Social robots are a puzzle. On the one hand, people interact with them as if they were humans or pets. They talk with them, show them things, and engage with them in joint activities. At the same time, people know that social robots are mechanical artifacts. They have metal and plastic parts, sensors for vision and hearing, and speech that sounds artificial. It seems self-contradictory, even irrational, for people to hold these two attitudes simultaneously: (a) a willingness to interact with social robots as real people or animals; and (b) a recognition that they are mechanical artifacts. The puzzle is not only theoretical but practical. When a robot stops moving, people must decide "Did the social agent fall asleep, or did the artifact's battery die?" And when its finger breaks off, "Am I sad because the social agent is in pain, or because the artifact needs repairing?" Call this the *social artifact puzzle*. (section 1, para 2, emphasis original)

As has been argued in more detail elsewhere (Ziemke & Thellman, 2023), the claim that there is something "self-contradictory, even irrational" about how people interpret social robots might be overstated. Let us take the crosswalk scenario illustrated in Figure 1 as an example. Taking the *intentional stance*, the pedestrian might wonder, "Has that car seen me?," "Does it understand that I want to cross the road?," or "Is it planning to stop for me?" (cf. Ziemke, 2020). Alternatively, the pedestrian could of course take what Dennett refers to as the *design stance* and try to predict the car's behavior based on the *general* assumption that such vehicles are designed to detect people and not harm them. That might seem more appropriate to some readers—and safer to pedestrians—but this would still require the pedestrian to make additional, more *situation-specific* assumptions about whether the car has actually detected him in that particular situation (Thellman & Ziemke, 2021; Ziemke, 2020). This brings us back to what I said earlier about review findings regarding the consequences of mental state attribution to robots (Thellman et al., 2022): In a nutshell, such attributions seem to lead to increased predictability, explainability, and trust, which means that treating such artifacts as intentional, social agents might simply make them easier to interact with. In that sense, H. Clark and Fischer's (2022) "social artifact puzzle" is less puzzling than it might seem at first (Ziemke & Thellman, 2023). Moreover, we could speculate that in safety-critical situations like the crosswalk scenario, it might in fact be *more rational* to view the driverless car as *both* an intentional agent and a mechanical artifact, instead of picking only one of them.[3]

Related to the question of what is rational or irrational, in discussions of human interpretations of robot behavior, there is sometimes a tendency to view folk psychological interpretations of robots as intentional agents as problematic (e.g., Fuchs, 2022; Sharkey & Sharkey, 2021). This is closely related to similar discussions of anthropomorphism, which generally can be characterized as "the human tendency to attribute human traits to non-human entities" (Damiano & Dumouchel, 2018, p. 2), including both animals and artifacts (cf. earlier discussion). As Damiano and Dumouchel point out, anthropomorphism traditionally "has been viewed as a bias, a category mistake, an obstacle to the advancement of knowledge, and as a psychological disposition typical of those who are immature and unenlightened, i.e., young children and 'primitive people' " (p. 2; see, e.g., Caporael, 1986; Fisher, 1996; S. D. Mitchell, 2005).

Related negative views can be found in discussions of "deception" in social robotics. Sharkey and Sharkey, for example, have argued that "efforts to develop features that promote the illusion of mental life in robots could be viewed as forms of deception" (Sharkey & Sharkey, 2011, p. 34) because "current robots have neither minds nor experiences" (Sharkey & Sharkey, 2021, p. 309). Their argument that the "appearance and behaviour of a robot can lead to an overestimation of its functionality or to an illusion of sentience or cognition that can promote misplaced trust and

3 This might, admittedly, not be in line with the "swift response" aspect discussed earlier.

inappropriate uses such as care and companionship of the vulnerable" (Sharkey & Sharkey, 2021, p. 309) is in fact much in line with my own suggestion that social robot design needs to "better guide users of interactive autonomous systems by encouraging appropriate attributions of intentionality and mental states and discouraging inappropriate ones, thereby reducing unrealistic expectations on such systems" (Ziemke, 2020, p. 2).

Following Coeckelbergh's (2018) discussion of deception by information technologies in terms of magic and performance, Sharkey and Sharkey (2021) elaborate their position on deception in social robotics as follows:

> As in the case of a magic show, the users/audience may not have been fooled into thinking that a robot is sentient, or has emotions, and may know that it is a trick. This is likely to be the case when a social robot is displayed to an audience with a reasonable knowledge of robotics. The audience members could enjoy the performance, at the same time as looking for clues or asking questions about how the performance was accomplished. This is less likely to be the case with naïve audiences, or vulnerable groups of people such as the very young, or older people with cognitive limitations. (p. 311)

The concern about naive audiences and vulnerable user groups is of course clearly warranted, especially where people are involuntarily exposed to robots they might not be prepared to deal with. The crosswalk scenario illustrated in Figure 1, for example, would be rather different if we were to consider very young or very old pedestrians and the potential cognitive limitations that are characteristic for different age groups. The notion of "deception" in social robotics is not unproblematic, though, owing to the difficulties discussed earlier in assessing what mental states people attribute to robots for the purpose of making them interpretable and interactable versus what states they think those robots really have (cf. Thellman & Ziemke, 2019). In a similar vein, Damiano and Dumouchel (2018) point out that

> anthropomorphic projections do not [necessarily] rest on the prior belief that an object or animal has human like mental states. It rests on the recognition that one is dealing with an entity that acts . . . and that the relation has changed, from, say, a relation of use to a form of interaction. That is: to a relation that requires the coordination of the actions of two "agents" for any one of them to be able to achieve his, her or its goal. (p. 6)

This leads us to what might be called the *observer's grounding problem*. In the original symbol grounding problem (Harnad, 1990), as illustrated in the discussion of Searle's (1980, 1990) Chinese Room argument, the problem was that the AI system/the computer/Searle-inside-the-room lacked original intentionality or intrinsic meaning because of an inability to make the connection between the symbols manipulated in the room and the objects or events in the real world (outside the room) that observers outside the room might consider the symbols to refer to. In robotic autonomous agents that are physically grounded and situated in the real world (i.e., in the same environment/room as the human observer), the problem is "solved" (cf. Steels, 2008) in the sense that the robot's internal mechanisms are grounded in sensorimotor interaction with the environment. Whatever internal mechanisms, representations, and so on your robot lawn mower may use, for example, to avoid obstacles, keep track of its battery level, navigate to the charging station, and so on, it clearly works, at least most of the time.

As I have pointed out, though, the fact that robots share physical environments with people does not necessarily mean that they are situated in the same perceptual and social world as humans. This was earlier exemplified with hedgehogs and small toys left on the lawn, which robot lawn mowers simply might not be equipped to perceive and/or to which they might not attach the same meaning as you or your kids do. This is relatively obvious but becomes both more problematic and more significant in cases when one actually needs to interact with a robot, instead of just observing

it. If we, for example, again consider the crosswalk scenario illustrated in Figure 1, the cognitive burden on the pedestrian as the human observer interpreting and anticipating the car's behavior is rather high. He needs to improvise—in "real time," and possibly without having encountered that particular car before—an appropriate understanding of the situation that allows, in Damiano and Dumouchel's (2018) terms, both agents to achieve their goals.

What I call the *observer's grounding problem*, then, is the fact that human observers, for the reasons discussed, are of course likely to interpret the situation through the lens of their own folk psychological interpretations and anthropomorphic projections; the observers use their own grounding, based on their own biological autonomy and intentional agency, plus their experiences of interactions with other living systems, to interpret the behavior of a robotic system, whose physical and sensorimotor grounding is likely to be very different. Moreover, although the robot and its human observers are now, in terms of the Chinese Room metaphor, "in the same room," observers are still in an important sense in the same situation as the observers outside the room in Searle's original argument—the human observers have no direct access to what goes inside the robot's "head" (the software controlling the car, in the crosswalk scenario) but have to use their own groundings (perception, experience, etc.) to interpret how the robot might perceive the situation.[4]

For people encountering socially interactive systems, this poses the nontrivial challenge to interpret them as intentional agents to understand and predict their behavior, but also keep in mind that the intentionality of artificial, robotic bodies is fundamentally different from that of their natural counterparts. This requires, on one hand, a "suspension of disbelief" (cf. Duffy & Zawieska, 2012) but, on the other hand, also a capacity for the "suspension of belief." The "suspension of disbelief" occurs when, in our crosswalk example, the pedestrian adopts some folk psychological/anthropomorphic interpretation like "the car has seen me, understands that I want to cross the road, and intends to slow down and let me do that"—as if the car had a human driver. The "suspension of belief" occurs, or might occur, when the observer reminds himself that his interpretation of the robot's (in this case, the car's) interpretation of the situation is not necessarily accurate—or when it turns out that the interaction does not actually unfold in the way the observer had expected.

This dual nature of (attributed) artificial intentionality has been addressed only rather superficially in embodied AI and social robotics research (cf. Froese & Ziemke, 2009). As discussed in more detail elsewhere (Thellman & Ziemke, 2021; Ziemke & Sharkey, 2001), research in embodied AI and cognitive science (e.g., Brooks, 1991b; A. Clark, 1997; Suchman, 1987; Ziemke, 1997; Ziemke & Sharkey, 2001) acknowledged early on that any robot necessarily has a different *perceptual world*—or *Umwelt* (von Uexküll, 1982)—than humans, but the implications for HRI have received relatively little attention so far. Thellman and Ziemke (2021) have referred to this as the *perceptual belief attribution problem* in human–robot interaction: How can people understand what robots know (and do not know) about the shared physical environment? The *observer's grounding problem*, as formulated in this article, then, is the other side of the same coin, you might say, because it addresses the fact that people (a) tend to interpret interactive situations in terms of their own folk psychological/anthropomorphic perception and understanding and (b) might not have any other choice anyway, unless they happen to have the technical expertise or interaction experience to have a sufficiently in-depth understanding of how a particular robot *really* sees the world.

4 Conclusion: Toward a Practice of Socially Interactive Autonomous Systems

Let me summarize the arguments in this article. We started off by discussing the distinction between *biological autonomy* and *artificial autonomy* implied in Bourgine and Varela's (1992) notion of Artificial

4 It might be worth noting that the term *observer's grounding problem* is of course ambiguous—and intentionally so; no pun intended—because it can be parsed in two ways and thus refers to both (a) the observer having a grounding problem (it's for him that things are at stake, e.g., the crosswalk scenario in Figure 1) and (2) the observer's grounding being the problem (rather than the AI system's grounding, as in the original symbol grounding problem).

Life research as constituting or moving toward a *practice of autonomous systems*. We also addressed Varela's (1992, 1997) notion of *intentionality*—as rooted in biological autonomy—which, at least at that time, traditional AI systems were considered to lack. We then moved on to address the *grounding problem* and related criticisms of traditional AI, as well as embodied AI approaches to overcoming these problems. Today's embodied AI and robots might be said to have some form of *artificial intentionality*, because clearly they are capable of successfully interacting with objects in the real world, most of the time. The philosophical question as to what degree such artificial intentionality constitutes *original intentionality*—or just another case of *extended human intentionality*—will probably have to be left to a few more decades of discussion of the Chinese Room argument. We then zoomed in on current research on social robots and people's folk psychological interpretations in terms of intentional agency, addressing the *social artifact puzzle*, discussions of *deception* in social robotics, and what I referred to as the *observer's grounding problem*.

We found that the distinction—or, should we say, the *relation*—between biological and artificial intentional agency plays a central role in many of these discussions. On one hand, some would argue that, as a matter of fact, "current robots have neither minds nor experiences" (Sharkey & Sharkey, 2021, p. 309) and therefore also should not be attributed mental or emotional states, at least not where naive audiences and vulnerable user groups are involved. On the other hand, empirical studies indicate (Thellman et al., 2022) that people in many cases—but notably not in all cases!—*choose* to make folk psychological attributions of intentional agency and that such attributions seem to increase predictability, explainability, and trust but also seem to increase cognitive drain and moral concern. Damiano and Dumouchel (2018) have also argued that instead of taking a dichotomous approach to the ethics of social robotics that considers them "a 'cheating' technology, as they generate in users the illusion of reciprocal social and affective relations" (p. 1), we might want to simply acknowledge anthropomorphism, folk psychology, and so on as distinctly human fundamental mechanisms of social interaction (cf. Duffy, 2003; Papagni & Koeszegi, 2021)—which would also allow us to develop what they call "a critical experimentally based ethical approach to social robotics" (Damiano & Dumouchel, 2018, p. 1).

These discussions illustrate that today, when social robots are starting to become a part of people's everyday lives, we need to complement Bourgine and Varela's (1992) notion of a *practice of autonomous systems* by also developing a *practice of socially interactive autonomous systems*. One of the elements of such a practice will need to be a concern for what Sharkey and Sharkey (2021) refer to as naive audiences and vulnerable users, that is, people who might have difficulties dealing with the cognitive complexities involved in social human–robot interactions. One way of addressing this is what Sharkey and Sharkey (2021) refer to as "allocation of responsibility for harmful deception" (p. 309), a way of holding different stakeholders accountable for people's interpretation of what robots can and cannot do. But it should also be noted that much more can be done in terms of fundamental AI literacy (e.g., Ng et al., 2021) to make "naive audiences" less naive. The cognitive science and AI communities, large parts of which still hold on to traditional views of cognition as computation (cf. Ziemke, 2022) and "mind as machine" (e.g., Boden, 2006), certainly could and should contribute more to clarifying the differences between people and existing technologies. This would need to include a clearer acknowledgment that, at least at this point, (human) cognition and (machine) computation are still very different things.

Widespread misconceptions of AI and robotic technologies are apparent in, for example, recent discussions of self-driving cars, which have been characterized by overly optimistic predictions (cf. M. Mitchell, 2021). Part of the problem is that such systems lack the embodied experience that allows human drivers to understand the perspectives of other drivers, cyclists, and pedestrians. Despite decades of discussion of embodied cognition and social interaction in the cognitive sciences (e.g., Varela et al., 1991; Ziemke, 2022), many people still fail to appreciate that such embodied understanding would be difficult to acquire or replicate in self-driving cars—which, after all, differ radically from people in their "embodiment." This is illustrated by recent experiments of ours in which people tended to assume that self-driving cars (a) have humanlike perceptual capabilities (Thellman, Pettersson, et al., 2023) and (b) have the cognitive capacity of *object*

permanence (which usually takes children four to eight months of sensorimotor interaction to develop) (Thellman, Holmgren, et al., 2023). For similar reasons, after accidents involving automated vehicles, people often wonder why such systems lack the common sense required to understand human behavior (e.g., jaywalking; Marshall & Davies, 2019). These are examples of expectations of AI and robotic technologies that are overly human centered or simply too high (cf. Ziemke, 2020). *Common sense*, for example, has been discussed as a problem in AI research for decades (e.g., Dreyfus, 1972/1979; Taube, 1961) and is not likely to be resolved any time soon. Many researchers in AI and cognitive science are, of course, well aware of this but have not managed to communicate these fundamental limitations to the general public sufficiently well. More specifically, overly high expectations would be easier to avoid if researchers in fields like cognitive science, AI, and social robotics could manage to develop and communicate a clearer understanding of what humans, as evolved living systems, share with other species, in particular, the domestic animals with which we have coevolved over extended periods of biological and cultural evolution (cf. Urquiza-Haas & Kotrschal, 2015)—and what we consequently do not share with computers, robots, self-driving cars, and so on, whose cognitive and behavioral capacities lack those shared biological roots.

It should be noted, though, that making a clear distinction between biological and artificial autonomy, agency, and intentionality as they currently are does not necessarily imply a dichotomous approach that insists on fundamental differences between biological and artificial autonomous systems that cannot possibly be overcome. Humans and social robots can, of course, also coadapt, codevelop, and coevolve (cf. Damiano & Dumouchel, 2018), in other, nonbiological ways. How this will affect the future of human–robot social interactions remains to be seen—humans are, after all, a biological, cultural, and technological species. The practice of socially interactive autonomous systems, therefore, needs to be informed and guided by a thorough understanding of the differences between artificial and biological bodies and their implications in the context of social interactions between people and autonomous technologies.

Acknowledgments

This work was supported by ELLIIT, the Excellence Center at Linköping-Lund in Information Technology (https://elliit.se/), and a Swedish Research Council (VR) grant (2022-04602) on "Social Cognition in Human–Robot Interaction." The author thanks Sam Thellman, Stevan Harnad, Luisa Damiano, Pasquale Stano, and three anonymous reviewers for useful feedback on earlier versions of this article.

References

Black, D. (2014). *Embodiment and mechanisation: Reciprocal understanding of body and machine from the Renaissance to the present*. Ashgate.

Boden, M. (2006). *Mind as machine: A history of cognitive science* (2 vols.). Oxford University Press.

Bourgine, P., & Varela, F. J. (1992). Towards a practice of autonomous systems. In F. J. Varela & P. Bourgine (Eds.), *Toward a practice of autonomous systems* (pp. xi–xvii). MIT Press.

Brooks, R. A. (1991a). Intelligence without reason. In *Proceedings of the twelfth International Joint Conference on Artificial Intelligence (IJCAI-91)* (pp. 569–595). Morgan Kauffmann.

Brooks, R. A. (1991b). Intelligence without representation. *Artificial Intelligence*, *47*(1–3), 139–159. https://doi.org/10.1016/0004-3702(91)90053-M

Brooks, R. A. (1993). The engineering of physical grounding. In *Proceedings of the fifteenth annual conference of the Cognitive Science Society* (pp. 153–154). Erlbaum.

Buckwalter, W., & Phelan, M. (2013). Function and feeling machines: A defense of the philosophical conception of subjective experience. *Philosophical Studies*, *166*(2), 349–361. https://doi.org/10.1007/s11098-012-0039-9

Caporael, L. R. (1986). Anthropomorphism and mechanomorphism: Two faces of the human machine. *Computers in Human Behavior*, *2*(3), 215–234. https://doi.org/10.1016/0747-5632(86)90004-X

Caporael, L. R., & Heyes, C. M. (1997). Why anthropomorphize? Folk psychology and other stories. In R. W. Mitchell, N. S. Thompson, & H. L. Miles (Eds.), *Anthropomorphism, anecdotes, and animals* (pp. 59–73). University of New York Press.

Chalmers, D. J. (1992). Subsymbolic computation and the Chinese room. In J. Dinsmore (Ed.), *The symbolic and connectionist paradigms: Closing the gap* (pp. 25–48). Erlbaum.

Clark, A. (1997). *Being there: Putting brain, body, and world together again*. MIT Press. https://doi.org/10.7551/mitpress/1552.001.0001

Clark, H. H., & Fischer, K. (2022). Social robots as depictions of social agents. *Behavioral and Brain Sciences*, *28*, 1–33. https://doi.org/10.1017/S0140525X22000668, PubMed: 35343422

Coeckelbergh, M. (2018). How to describe and evaluate "deception" phenomena: Recasting the metaphysics, ethics, and politics of ICTs in terms of magic and performance and taking a relational and narrative turn. *Ethics and Information Technology*, *20*, 71–85. https://doi.org/10.1007/s10676-017-9441-5

Coradeschi, S., & Saffiotti, A. (2003). An introduction to the anchoring problem. *Robotics and Autonomous Systems*, *43*(2–3), 85–96. https://doi.org/10.1016/S0921-8890(03)00021-6

Damiano, L., & Dumouchel, P. (2018). Anthropomorphism in human–robot co-evolution. *Frontiers in Psychology*, *9*, 468. https://doi.org/10.3389/fpsyg.2018.00468, PubMed: 29632507

Dennett, D. C. (1971). Intentional systems. *Journal of Philosophy*, *68*(4), 87–106. https://doi.org/10.2307/2025382

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

Dorffner, G., & Prem, E. (1993). Connectionism, symbol grounding, and autonomous agents. In *Proceedings of the fifteenth annual meeting of the Cognitive Science Society* (pp. 144–148). Erlbaum.

Dreyfus, H. (1979). *What computers can't do* (Rev. ed.). Harper and Row. (Original work published 1972).

Duffy, B. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, *42*(3–4), 177–190. https://doi.org/10.1016/S0921-8890(02)00374-3

Duffy, B., & Zawieska, K. (2012). Suspension of disbelief in social robotics. In *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication* (pp. 484–489). IEEE. https://doi.org/10.1109/ROMAN.2012.6343798

Fiala, B. Arico, A., & Nichols, S. (2014). You, robot. In E. Machery & E. O'Neill (Eds.), *Current controversies in experimental philosophy*. Routledge. https://doi.org/10.4324/9780203122884-5

Fisher, J. A. (1996). The myth of anthropomorphism. In M. Bekoff & D. Jamieson (Eds.), *Readings in animal cognition* (pp. 3–16). MIT Press.

Froese, T., & Ziemke, T. (2009). Enactive artificial intelligence: Investigating the systemic organization of life and mind. *Artificial Intelligence*, *173*(3–4), 466–500. https://doi.org/10.1016/j.artint.2008.12.001

Fuchs, T. (2022). Understanding Sophia? On human interaction with artificial agents. *Phenomenology and Cognitive Sciences*. https://doi.org/10.1007/s11097-022-09848-0

Fussell, S. R., Kiesler, S., Setlock, L. D., & Yew, V. (2008). How people anthropomorphize robots. In *2008 ACM/IEEE international conference on human-robot interaction* (pp. 145–152). IEEE. https://doi.org/10.1145/1349822.1349842

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619. https://doi.org/10.1126/science.1134475, PubMed: 17272713

Guthrie, S. E. (1997). Anthropomorphism: A definition and a theory. In R. W. Mitchell, N. S. Thompson, & H. L. Miles (Eds.), *Anthropomorphism, anecdotes, and animals* (pp. 59–73). University of New York Press.

Harnad, S. (1989). Minds, machines and Searle. *Journal of Experimental and Theoretical Artificial Intelligence*, *1*(1), 5–25. https://doi.org/10.1080/09528138908953691

Harnad, S. (1990). The symbol grounding problem. *Physica D*, *42*, 335–346. https://doi.org/10.1016/0167-2789(90)90087-6

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology*, *57*(2), 243–259. https://doi.org/10.2307/1416950

Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology*, *33*(2), 101–121. https://doi.org/10.1006/jesp.1996.1314

Malle, B. F., Moses, L. J., & Baldwin, D. A. (2004). Introduction: The significance of intentionality. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 1–24). MIT Press.

Marshall, A., & Davies, A. (2019). Uber's self-driving car didn't know pedestrians could jaywalk. *Wired*, May 11. http://www.wired.com/story/ubers-self-driving-car-didnt-know-pedestrians-could-jaywalk/

Melson, G. F., Kahn, P. H. Jr., Beck, A., & Friedman, B. (2006, July 17). *Toward understanding children's and adults' encounters with social robots* [Paper presentation]. AAAI Workshop on Human Implications of Human-Robot Interaction, Boston, MA, United States.

Mitchell, M. (2021). *Why AI is harder than we think*. https://doi.org/10.48550/arXiv.2104.12871

Mitchell, S. D. (2005). Anthropomorphism and cross-species modeling. In L. Daston & G. Mitman (Eds.), *Thinking with animals* (pp. 100–118). Columbia University Press.

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers, and Education: Artificial Intelligence*, *2*, 100041. https://doi.org/10.1016/j.caeai.2021.100041

Papagni, G., & Koeszegi, S. A. (2021). Pragmatic approach to the intentional stance—semantic, empirical and ethical considerations for the design of artificial agents. *Minds and Machines*, *31*, 505–534. https://doi.org/10.1007/s11023-021-09567-6

Perez-Osorio, J., & Wykowska, A. (2020). Adopting the intentional stance toward natural and artificial agents. *Philosophical Psychology*, *33*(3), 369–395. https://doi.org/10.1080/09515089.2019.1688778

Schellen, E., & Wykowska, A. (2019). Intentional mindset toward robots—Open questions and methodological challenges. *Frontiers in Robotics and AI*, *5*, 139. https://doi.org/10.3389/frobt.2018.00139, PubMed: 33501017

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. https://doi.org/10.1075/is.14.3.02sci

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–457. https://doi.org/10.1017/S0140525X00005756

Searle, J. R. (1990). Is the brain's mind a computer program? *Scientific American*, January, 26–31. https://doi.org/10.1038/scientificamerican0190-26, PubMed: 2294583

Searle, J. R. (1999). *Mind, language and society: Philosophy in the real world*. Basic Books.

Sharkey, A., & Sharkey, N. (2011). Children, the elderly, and interactive robots. *IEEE Robotics and Automation Magazine*, *18*(1), 32–38. https://doi.org/10.1109/MRA.2010.940151

Sharkey, A., & Sharkey, N. (2021). We need to talk about deception in social robotics! *Ethics and Information Technology*, *23*, 309–316. https://doi.org/10.1007/s10676-020-09573-9

Sharkey, N., & Ziemke, T. (2001). Mechanistic versus phenomenal embodiment: Can robot embodiment lead to strong AI? *Cognitive Systems Research*, *2*(4), 251–262. https://doi.org/10.1016/S1389-0417(01)00036-5

Steels, L. (1993). The artificial life roots of artificial intelligence. *Artificial Life*, *1*(1_2), 75–110. https://doi.org/10.1162/artl.1993.1.1_2.75

Steels, L. (2008). The symbol grounding problem has been solved, so what's next? In M. de Vega, A. Glenberg, & A. Graesser (Eds.), *Symbols and embodiment: Debates on meaning and cognition* (pp. 223–244). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199217274.003.0012

Steels, L., & Brooks, R. A. (1995). *The Artificial Life route to artificial intelligence: Building embodied, situated agents*. Routledge.

Suchman, L. A. (1987). *Plans and situated actions: The problem of human–machine communication*. Cambridge University Press.

Sytsma, J., & Machery, E. (2010). Two conceptions of subjective experience. *Philosophical Studies*, *151*(2), 299–327. https://doi.org/10.1007/s11098-009-9439-x

Taube, M. (1961). *Computers and common sense: The myth of thinking machines*. Columbia University Press. https://doi.org/10.7312/taub90714

Thellman, S., de Graaf, M., & Ziemke, T. (2022). Mental state attribution to robots: A systematic review of conceptions, methods, and findings. *ACM Transactions on Human–Robot Interaction*, *11*(4), 41. https://doi.org/10.1145/3526112

Thellman, S., Holmgren, A., Pettersson, M., & Ziemke, T. (2023). Out of sight, out of mind? Investigating people's assumptions about object permanence in self-driving cars. In *HRI '23: Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction, March 2023* (pp. 602–606). Association for Computing Machinery Digital Library. https://doi.org/10.1145/3568294.3580156

Thellman, S., Pettersson, M., Holmgren, A., & Ziemke, T. (2023). In the eyes of the beheld: Do people think that self-driving cars see what human drivers see? In *HRI '23: Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction, March 2023* (pp. 612–616). Association for Computing Machinery Digital Library. https://doi.org/10.1145/3568294.3580158

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. https://doi.org/10.3389/fpsyg.2017.01962, PubMed: 29184519

Thellman, S., & Ziemke, T. (2019). The intentional stance toward robots: Conceptual and methodological considerations. In *The 41st annual conference of the Cognitive Science Society, July 24–26, Montreal, Canada* (pp. 1097–1103). Cognitive Science Society.

Thellman, S., & Ziemke, T. (2020). Do you see what I see? Tracking the perceptual beliefs of robots. *iScience*, *23*(10), 101625. https://doi.org/10.1016/j.isci.2020.101625, PubMed: 33089112

Thellman, S., & Ziemke, T. (2021). The perceptual belief problem: Why explainability is a tough challenge in social robotics. *ACM Transactions on Human–Robot Interaction*, *10*(3), 29. https://doi.org/10.1145/3461781

Urquiza-Haas, E. G., & Kotrschal, K. (2015). The mind behind anthropomorphic thinking: Attribution of mental states to other species. *Animal Behaviour*, *109*, 167–176. https://doi.org/10.1016/j.anbehav.2015.08.011

Varela, F. J. (1992). Autopoiesis and a biology of intentionality. In B. McMullin & N. Murphy (Eds.), *Proceedings of Autopoiesis and Perception: A Workshop with ESPRIT BRA 3352* (pp. 4–14). Dublin City University.

Varela, F. J. (1997). Patterns of life: Intertwining identity and cognition. *Brain and Cognition*, *34*(1), 72–87. https://doi.org/10.1006/brcg.1997.0907, PubMed: 9209756

Varela, F. J., & Bourgine, P. (Eds.) (1992). *Toward a practice of autonomous systems*. MIT Press.

Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. MIT Press. https://doi.org/10.7551/mitpress/6730.001.0001

von Uexküll, J. (1982). The theory of meaning. *Semiotica*, *42*(1), 25–79. https://doi.org/10.1515/semi.1982.42.1.25

Ziemke, T. (1997). Rethinking grounding. In A. Riegler & M. Peschl (Eds.), *Does representation need reality? Proceedings of the international conference "New Trends in Cognitive Science" (NTCS 97)—Perspectives from cognitive science, neuroscience, epistemology, and artificial life* (pp. 87–94; Technical Report No. 9701). Austrian Society for Cognitive Science.

Ziemke, T. (1999). Rethinking grounding. In A. Riegler, M. Peschl, & A. von Stein (Eds.), *Understanding representation in the cognitive sciences* (pp. 177–190). Plenum Press. https://doi.org/10.1007/978-0-585-29605-0_20

Ziemke, T. (2020). Understanding robots. *Science Robotics*, *5*(46), eabe2987. https://doi.org/10.1126/scirobotics.abe2987, PubMed: 32999050

Ziemke, T. (2022). Embodiment in cognitive science and robotics. In A. Cangelosi & M. Asada (Eds.), *Cognitive robotics* (pp. 213–229). MIT Press. https://doi.org/10.7551/mitpress/13780.003.0016

Ziemke, T., & Sharkey, N. (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), 701–746. https://doi.org/10.1515/semi.2001.050

Ziemke, T., & Thellman, S. (2023). How puzzling is the social artifact puzzle? *Behavioral and Brain Sciences*, *46*, e50. https://doi.org/10.1017/S0140525X22001571, PubMed: 37017073

Zlatev, J. (2001). The epigenesis of meaning in human beings, and possibly in robots. *Minds and Machines*, *11*(2), 155–195. https://doi.org/10.1023/A:1011218919464

Zlatev, J. (2002). Meaning = Life (+ Culture): An outline of a unified biocultural theory of meaning. *Evolution of Communication*, *4*(2), 253–296. https://doi.org/10.1075/eoc.4.2.07zla