Croatian Journal of Philosophy
Vol. XXVI, No. 76, 2026
https://doi.org/10.52685/cjp.26.76.2
Received: July 18, 2025
Accepted: February 15, 2026

The Attribution of Rationality to Robots

EDOARDO DATTERI
University of Milano-Bicocca, Milan, Italy

An increasing number of studies are attempting to determine, through quantitative experimentation, whether people adopt an intentional stance towards robots. These studies mainly use questionnaires in which participants are asked to choose between mentalistic and non-mentalistic descriptions of robotic behaviours portrayed in pictures. While these methods are extremely interesting in their attempt to operationalise Dennett's theoretical constructs, they only capture one aspect of the intentional stance: the attribution of mental states to robots. They neglect the question of whether participants also attribute rationality to the system. Consequently, they are not well equipped to analyse how people form expectations about the behaviour of the robots they interact with, which is crucial for studying the dynamics of human–robot interaction. There is indeed no reason to deny that laypeople might occasionally attribute mental states to robots while believing that they can act irrationally or model the decision-making processes of the system in terms devoid of any reference to rationality. Building on these considerations, this article reflects on an emerging area of research in human-robot interaction from a philosophical perspective, identifying a potential limitation that could be overcome by referring to psychological literature on the attribution of rationality to humans.

Keywords: Human-robot interaction; mental state attribution; intentional stance.

1. Introduction

When people interact with autonomous artificial systems such as intelligent chatbots and social robots, they sometimes *talk* about them in mentalistic terms ("Hey, look! The robotic vacuum cleaner *wants* to go to the kitchen!"). They can also *attribute* beliefs and desires to the system, assuming that it will behave rationally based on them. In a

series of groundbreaking essays, the philosopher Daniel Dennett famously termed this latter phenomenon the adoption of an 'intentional stance' towards the observed system, discussing various philosophical aspects of the idea. For example, he considered what role beliefs and desires play in intentional systems theory and whether the intentional stance aligns with common-sense psychology (Dennett 1989). The notion of intentional stance pervades contemporary research on human–robot interaction. In particular, a growing community of researchers comprising roboticists, psychologists and social scientists is studying the adoption of an intentional stance towards robots using a variety of experimental techniques (see Thellman et al. 2022 for a review). This article aims to comment on a particular aspect of this scientific enterprise by suggesting that this research focuses too much on attributing beliefs and desires to robots and neglects another important aspect of the intentional stance: attributing *rationality* to robots.

The concept of rationality plays a central role in Dennett's theory of intentional systems. He presents the intentional stance as involving the attribution of mental states, such as beliefs and desires, *as well as* rationality, to the target system. Adopting the intentional stance is not just about attributing the desire to reach the kitchen and the belief that the battery needs to recharge to a robotic vacuum cleaner, for example; it also involves assuming that there is a certain relationship between the system's beliefs and desires, and its behaviour – namely, that the system will act rationally based on them. However, it has been suggested that this assumption is too strong, as we sometimes observe irrational behaviour in systems to which we attribute a mind. For instance, people sometimes appear to fail to make the most rational choice in tasks such as the Wason selection task (Wason 1968). Consequently, there has been a debate about whether, when interpreting others' behaviour in mentalistic terms, people are actually adopting the intentional stance or engaging in a form of folk psychology that does not necessarily involve the attribution of rationality (see Dennett 1989; Stich 1981). This paper does not seek to take a position in this debate. It only argues that attributing mental states to a system is independent of attributing rationality to it. This means that the attribution of mental states can happen without attributing rationality, and vice versa. It also argues that current attempts to determine whether people adopt an intentional stance towards robots tend to focus only on attributing mental states. This means they neglect attributing rationality and do not actually investigate the adoption of an intentional stance towards robots. In doing so, this article provides a philosophical reflection on certain aspects of contemporary research in the fields of social robotics and human–robot interaction.

Why is it important to study whether people attribute not only a mind, but also rationality, to robots? Our understanding of these systems clearly influences our behaviour in their presence and how we interact with them. However, while certain aspects of our behaviour can

be affected by the mere attribution of mental states to robots, others arguably depend more deeply on whether we also attribute rationality to them. Therefore, to understand certain aspects of our interactions with robots, it is beneficial, if not essential, to determine whether we regard them as rational, irrational or non-rational. One study to be discussed in Section 3, by Wiese and colleagues (2012) suggests that adopting an intentional stance towards a robot activates certain cognitive processes, such as gaze following, which are not activated when that stance is not adopted. According to the authors, this occurs in a reflexive, bottom-up way that does not necessarily involve rationalising the robot's behaviour. Studying our attribution of rationality to robots may be less important for understanding our low-level reactions to their behaviour. However, other aspects of human–robot interaction depend more crucially on our expectations of their behaviour. For instance, when the 'battery' LED on our robotic vacuum cleaner lights up and the robot heads towards the kitchen, we anticipate that it will go to the docking station, which is indeed located there. If we then recall putting a bin in front of the docking station, we will move it to allow the robot to recharge properly. Similarly, if we are crossing a street and an autonomous car is approaching (Ziemke 2020), we will decide whether to continue or step back depending on whether we expect the car to stop at the crossing. Our decision to install an assistive robot in our elderly aunt's apartment will be based on our expectations of how the robot will behave, and whether or not it will harm her.

The issue is that the dynamics of our interactions with robots are significantly shaped by our expectations of their behaviour. In order to understand people's high-level reactions to robots and how they decide to behave when interacting with them, it is important to study how people predict robot behaviour.[1] In order to study people's predictions, it is important to determine not only whether they attribute a mind or specific mental states to robots, but also whether they assume that their behaviour will be rationally influenced by their beliefs and desires. In order to understand how pedestrians interact with autonomous cars, it is important to determine whether they attribute the desire not to hurt them to those cars; however, it is also important to determine whether they assume the car will act rationally based on that desire. The most rational thing for a car with this desire would be to stop at the crossing. However, the pedestrian's mental model of the car may include the assumption that it can act *irrationally* in certain situations, possibly because rationality is considered to apply only to

[1] Thellman and Ziemke (2020) have argued that more research in this area is needed. "Despite a growing interest in the role of mental state attribution in people's mental models of robots, and the importance of perceptual belief tracking in the context of social interaction, no research has so far targeted people's ability to predict the behavior of robots based on assumptions about how they perceive the environment." Their study is one of the few that explicitly addresses this important issue.

human beings. Alternatively, the pedestrian may simply assume that the car's behaviour is lawfully connected to its beliefs and desires in a way that is not characterised in terms of rationality or irrationality. People's mental models of robots may differ greatly in the way they connect attributed beliefs and desires to overt behaviour.

Various methods have been developed to study people's mental models of robots (Rueben et al. 2021), particularly in terms of whether people adopt the intentional stance towards them. Some studies have examined how this adoption affects brain activity (Chaminade et al. 2012), gaze following (Wiese et al. 2012), gaze aversion (Desideri et al. 2021), and other cognitive processes (Ciardo et al. 2020; Marchesi et al. 2025; Roselli et al. 2022). For reviews and general reflections, see Chaminade and Cheng (2009) and Wykowska and colleagues (2016). Other studies have attempted to determine whether people adopt the intentional stance based on the robot's physical appearance, behaviour, or other contextual factors (Mandell et al. 2017; Marchesi et al. 2019; Martini et al. 2015, 2016; Terada et al. 2007). These studies have made significant contributions to our understanding of how people perceive robots. However, they do not establish whether participants fully adopt the intentional stance towards robots since they only investigate the attribution of mental states to systems, not rationality. To demonstrate this, it is helpful to provide a brief overview of the intentional stance and argue that attributing beliefs and desires does not necessarily imply attributing rationality: a robot can be perceived as having beliefs and desires (and other propositional attitudes) while acting irrationally, or non-rationally, with respect to them.

2. Rationality and the intentional stance

2.1 The intentional stance

Dennett's theory of intentional systems is well known and has been the subject of extensive discussion since the publication of his seminal article (Dennett 1971). This section aims to recap some of its key features in preparation for the subsequent discussion. In this oft-quoted passage from (Dennett 1989), he defines the intentional stance (IS) as follows:

> Here is how it works: first you decide to treat the object whose behavior is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs. A little practical reasoning from the chosen set of beliefs and desires will in many – but not all – instances yield a decision about what the agent ought to do; that is what you predict the agent *will* do. (17)

The IS is often presented by the author as an explanatory strategy (e.g. Dennett 1971, 2009). Suppose the 'battery' LED on a robotic vacuum cleaner lights up and the robot turns towards the kitchen. An observer

adopting the IS might explain this behaviour by attributing the following beliefs and desires to the robot: the belief that the battery is running out of charge; the belief that the docking station is in the kitchen; and the desire to recharge. Given these beliefs and desires, they would assume that the most rational thing for the robot to do would be to turn towards the kitchen. In this way, the IS provides a rational explanation for the robot's behaviour. The IS can also be regarded as a predictive strategy: the observer might predict that the most rational course of action, after turning right, would be to proceed straight towards the kitchen. However, it is important to stress that, in Dennett's framework, attributing beliefs and desires to the system in the IS does not mean believing that the system actually possesses them. The observer only treats the system as if it possesses them for the sole purpose of explaining and predicting its behaviour. Intentional systems theory is based on an instrumentalist conception of beliefs, desires, and other propositional attitudes. In this theory, the conditions under which the target system can be said to have a belief (or desire) with a particular content do not have to be the same as those adopted in other psychological frameworks based on the attribution of so-called propositional attitudes.

Rationality plays a central role in Dennett's theory of intentional systems. Firstly, the rationality of the system forms an integral part of the mental model constructed by the observer. It characterises the relationship between the system's beliefs and desires, and its behaviour (meaning that, according to the IS, the system will act rationally relative to its beliefs and desires). Secondly, rationality arguably influences the process by which the observer identifies the beliefs and desires that explain the system's behaviour. Why did the robot turn right? One first assumes the rationality of the system – in this case, that turning right was the most rational response to the situation. Then, one assumes that the system has beliefs and goals that make the right-turning behaviour rational, i.e. one makes "adjustments in the information-processing conditions" (Dennett 1971: 94). In this case, the aforementioned beliefs and desires (the belief that the battery is low and that the docking station is in the kitchen, and the desire to recharge) do indeed render this behaviour rational. According to intentional systems theory, explanation is akin to rationalisation: the assumption that the system is rational informs the observer's decision regarding the beliefs and desires attributed to the system.

Note that predictive tasks differ from explanations in that the behaviour of the system is unknown. To predict how the robot will act after turning right, the observer must first attribute desires and beliefs to the system and then assume that the resulting action is the most rational given these premises. However, as the action is still unknown, it is not possible to determine the system's beliefs and desires based on the assumption that the behaviour produced was rational, as in an explanatory context. According to Dennett (see the chapter "Three

Kinds of Intentional Psychology" in Dennett 1989), one ascribes the beliefs that the system ought to have "given its perceptual capacities, its epistemic needs, and its biography", as well as the desires that the system ought to have "given its biological needs and the most practicable means of satisfying them". Arguably, an assumption of rationality plays a role here too, albeit perhaps more covertly, because the system is assumed to have beliefs and desires that are rational, given its epistemic and biological needs.

In summary, when adopting an IS towards a system such as the robotic vacuum cleaner used as an example in this section, both mentalistic and rationality assumptions are made. The mentalistic assumption is that the system has beliefs, desires, and other mental states with content (conceived as "idealised fictions in an action-predicting, action-explaining calculus" Dennett 1978: 30), while the rationality assumption is that the system acts rationally based on them. It is through this rationality assumption that the observer can connect the system's beliefs and desires to its actual behaviour; the rationality assumption plays a key role in generating expectations about the robot's future behaviour.

It is worth noting that Dennett does not provide a clear definition of rationality. Stich (1981) notes that Dennett himself "has no illusions on the point" and "portrays intentional-systems theory – the general normative theory of rationality – as a discipline in its infancy" (42). Indeed, Dennett clearly acknowledges this in his work (Dennett 1989): "What then do I say rationality is? I don't say" (94). He also claims to have "good reasons for cautiously resisting the demand for a declaration on the nature of rationality" (94) and merely indicates what, in his view, rationality is not: it is neither deductive closure nor logical consistency. He never connects the intentional systems theory to classical decision theory, which involves ordering the set of possible actions by assigning probabilities to their potential outcomes and determining their expected utility in relation to the system's goals (Mele and Rawling 2004; Peterson 2017). Even if he did, there would be many possible ways to implement a classical decision-making process. Unless additional details were specified, including the definition of the set of possible actions and the criteria for calculating their consequences and assigning their expected utilities with respect to the goals, the attribution of rationality that accompanies the IS would not enable the observer to make precise predictions about the system's behaviour. This suggests that Dennett's presentation of the IS conceptual framework provides poor grounds for the analysis of people's behavioural expectations about robots unless some aspects of it are more precisely worked out.

2.2 Non-mentalistic rationalization

The next section will argue that, although tasks for investigating the attribution of rationality to systems have been developed in non-ro-

botic psychological literature (see Gergely et al. 1995), contemporary research on adopting an IS towards robots focuses narrowly on attributing mental states to them and neglects the phenomenon of rationality attribution. In support of this claim, it is worth noting that these two types of attribution are relatively independent of one another; one can rationalise a robot's behaviour without ascribing mental states to it, and vice versa. Without this further assumption, one might dispute the main assertion of this article, arguing that methods for studying the attribution of mental states to robots also determine whether participants attribute rationality to them.

Can a mental model of a robot incorporate the assumption that the robot is rational without also assuming that it has mental states? This depends on how the system is regarded as rational. According to the IS, the system acts rationally *with respect to its beliefs and desires*. Thus, in Dennett's framework, rationality cannot be attributed without also attributing mental states. However, a robot may be considered rational in ways that differ from the IS. For instance, an observer might assume that the robot adheres to a specific standard of rationality based solely on its behaviour: this concept is known as *rational analysis* (Anderson 1991). As Anderson presents it, rational analysis belongs to a long tradition of trying to understand behaviour as an adaptation to the system's environment.

> A rational analysis is an explanation of an aspect of human behavior based on the assumption that it is optimized somehow to the structure of the environment. […] As in economics, the term does not imply any actual logical deduction in choosing the optimal behavior, only that the behavior is optimized. (471)

Anderson outlines the various steps of rational analysis as follows: First, the goals of the system are specified. As any behaviour can be viewed as optimising a potential goal, external constraints must be considered in order to select the system's actual goal. Note that the system is assumed to have *goals*, not *desires*. This is evident from the rational analysis of the various systems that Anderson uses as examples. For instance, the goal of memory systems is to access necessary information from the past, while categorisation systems aim to predict features of objects. Anderson does not characterise these systems as having desires, conceived as intentional mental states. According to this approach, it is appropriate to say that the goal of a thermostat is to maintain the environmental temperature within a certain range, without implying that the thermostat desires this. The second step is to develop a formal model of the environment to which the system is adapted. Anderson specifies that this must be done from the system's perspective; the model must include environmental features that are accessible to the system rather than taking the observer's point of view as a reference. The third step is to make some minimal assumptions about the computational limitations of the system under analysis. For example, we might assume that the system's ability to process alterna-

tives is limited, that processing incurs a cost, or that its short-term memory is finite. The fourth step is to derive the system's optimal behaviour given these assumptions. This involves predicting behaviour that will maximise expected utility when considering the goals identified in step one, the environmental constraints defined in step two and the computational costs defined in step three. These predictions are then evaluated against the existing literature or experimental results. If necessary, the theory is revised iteratively.

Anderson points out that identifying the variables that the system is optimising may constrain the identification of the internal mechanisms responsible for its behaviour. Some authors in the Commentary on (Anderson 1991) have claimed that identifying the computational limitations to which the system is subject at step three requires one to consider possible internal mechanisms. Anderson's approach to rational analysis occasionally suggests that it can inform the discovery of mechanisms. For example, he states that his rational analysis of problem solving "seems more like the actual problem solving that people face daily" (481), and he outlines a procedure for selecting the partial plan with the greatest probability of success. However, he also states that "the structure driving explanation in a rational theory is that of the environment" rather than the mind of the system (471). The goal of rational analysis is "to predict behavior from the structure of the environment rather than the structure of the mind" (474) and "it is in the spirit of a rational analysis to prescribe what the behavior of a system should be rather than how to compute it". The question of how the system succeeds in producing optimal behaviour – whether through the interaction of beliefs, desires, intentions and other propositional attitudes, or by virtue of a different mechanism – is not one that rational analysis is intended to solve. The rational analysis of problem solving presented in his article is objective in the sense that it is not necessarily "in the subject's head" (482). Furthermore, he does not claim "that the human system actually goes through the relatively complex Bayesian analysis used to establish what the optimal behavior was" (483). The term 'rational' "is used in the economist's sense, which is that the output of the system is optimal and no claim is made about the mental processes by which this output is computed" (510).

Clearly, rational analysis can be applied to the behaviour of robots. In this case, the observer does not make any assumptions about the mental state of the system, even instrumentally. The observer only needs to make assumptions corresponding to the steps of rational analysis, as previously illustrated, and these steps do not include anything concerning the system's mental state. Rational analysis can particularly be combined with adopting a design stance towards the system based on the assumption that its functional and computational processes are adapted to its intended purpose. Although Anderson's rational analysis would not lead to the adoption of a rationality assumption as intended by Dennett (since the latter refers to the system's beliefs and desires),

the considerations made in this section suggest that it is, in principle, possible to consider a robot to be rational and to use this assumption to predict its future behaviour without necessarily assuming that it possesses any mental states.

2.3 Non-rational mentalization

Can mental states be attributed to a robot without also assuming that it always acts in the most rational way? There is no obvious reason why an observer should be prevented from assuming that the robot's decision-making system is occasionally or always irrational. Moreover, an alternative assumption could be made: that the behaviour of the system is neither rational nor irrational, but simply *non-rational*. This would mean that the relationship between mental states and system behaviour is modelled without any reference to rationality. For example, suppose the observer attributes the following beliefs and desires to the robotic vacuum cleaner: the belief that its battery is running low and that the docking station is in the kitchen, and the desire to recharge. As previously mentioned, these assumptions are idle for prediction purposes in themselves, unless the observer makes additional assumptions about how these mental states contribute to behaviour. The rationality assumption incorporated in the IS can be expressed as follows:

> (A) The system performs the most rational action with respect to its desires and beliefs.

As previously mentioned, it is unclear how this assumption can yield precise predictions of the robot's behaviour unless accompanied by a detailed account of what constitutes rational behaviour. As such, it offers only a generic representation of how the system's behaviour is influenced by its desires and beliefs. For the purposes of this discussion, the important point is that ascribing those beliefs and desires to the system does not imply that the observer must assume (A) or make equivalent attributions of rationality. It is not inconsistent to ascribe certain beliefs and desires to the system while assuming that it will behave irrationally relative to them. Indeed, the observer might well assume (instead of A) something along the following lines:

> (B) The system performs the least rational action with respect to its desires and beliefs.

Now, it is one thing to assume that the system acts *irrationally* and another to assume that it acts *non-rationally*. The latter occurs when the observer does not characterise the relationship between mental states and behaviour in terms of rationality or irrationality, but assumes that the robot's mental states are connected to its behaviour via law-like generalisations, for example:

> (C) If the system desires to recharge and believes that the docking station is in the kitchen, it will turn towards the kitchen.

Or, more generally:

(D) If the system desires to recharge and believes that the docking station is in room X, then it will turn towards room X.

The striking difference between option (A) and options (C) and (D) is that the latter ones do not explicitly refer to rationality. Here, the system's beliefs, desires and other propositional attitudes interact in a law-like manner to produce behaviour, with no assumption as to whether these regularities are rational or irrational. It should be noted that the point is not that these generalisations *cannot*, in principle, be said to be rational or irrational. In fact, turning towards the kitchen is intuitively rational with respect to the desire to recharge, and options (C) and (D) may be said to guide the system towards rational behaviour. However, the notion of rationality plays no essential role in predicting the behaviour of the system based on these two generalisations. To illustrate this, suppose that one of the law-like regularities attributed by the observer to the system has the following form:

(E) If the system has a belief with content *a* and a belief with content (if *a* then *b*), then the system forms a belief with content *b*.

This generalisation can be used to predict the presence of specific beliefs within a system's knowledge base. For instance, if an observer attributes two beliefs to the system – one stating that today is Friday, and the other stating that, if today is Friday, it is laundry day – the above regularity would enable the observer to predict that the system would possess the belief that today is laundry day. Other generalisations attributed to the robot could allow the observer to link the system's possession of this belief to its subsequent motor actions (e.g. doing the laundry). One could argue that this generalisation takes the form of a *modus ponens* inference rule, which would make the system's decision-making process appear rational according to a traditional view of rationality. In other words, one could assume that the system is regulated by (E) *and* that (E) is a rational decision-making rule. However, this additional assumption does not play a crucial predictive role. An observer unaware that (E) incorporates a principle of rational thinking, or who conceives of rationality differently, would still predict that the robot believes 'Today is laundry day'. Assumption (E) would support this prediction even if *modus ponens* were the hallmark of irrationality. In this case, the notion of rationality plays no role in the observer's mental model of the system. This non-rational approach to understanding the robot's behaviour could result in the following generalisation being attributed to it:

(F) If the system has beliefs with contents *b* and (if *a*, then *b*), then it forms a belief with the content *a*.

The fact that (F) is a fallacious rule of inference would likely prevent it from being included in IS-style rationalisations of the system's behaviour. However, nothing actually prevents one from modelling the inter-

action between the system's beliefs, desires and behaviour in terms of this law-like regularity, without any reference to rationality.

In terms of prediction, how different are these two strategies? While they may converge in some cases, in others they will diverge significantly. Suppose two observers both attribute the beliefs and desires mentioned before to a robotic vacuum cleaner. Observer 1 adopts an IS approach, assuming that the robot will act in the most rational way given its current beliefs and desires. Observer 2 assumes that the robot is subject to (D) rather than to the rationality principle. In this case, both observers will probably have the same expectations regarding the robot's behaviour. However, this will not always be the case. While the predictive engine of the IS is based on just one rule, albeit underdetermined, observer 2's approach may ascribe additional or different law-like regularities to the system. For example, they might assume that, when certain conditions are met, the robot will move in the opposite direction to the kitchen, move more slowly in the direction it was previously travelling, or make a complete turn and stop. The first regularity could be considered irrational[2], whereas the last two options are neither clearly rational nor clearly irrational. However, the point is that there are "simply" law-like regularities which make no reference to rationality. They are neither inherently rational nor irrational.

These considerations suggest that an observer can have a mental model of a robot that attributes mental states, such as beliefs and desires, to it without assuming that the robot will act in the most rational way based on these states. Mentalising a system does not imply rationality attribution, but can be accompanied by a variety of non-rational models of the relationship between mental states and behaviour.[3] Although the predictions of two observers (one adopting the IS and the other a non-rationalistic modelling style) may sometimes converge, they will diverge in many other cases. To understand how people form their expectations of the robot's behaviour, it is essential to study not

[2] The idealising assumption made here is that the robot has no beliefs or desires other than those mentioned previously. Assuming that the system believes only that the battery is running out of charge and that the docking station is in the kitchen, and desires only to recharge the battery, turning in the direction opposite the kitchen would be an irrational decision according to many accounts of rationality. However, one might save the rationality assumption by inserting an additional belief: for example, that there is a more powerful docking station in the dining room, which is opposite the kitchen.

[3] It follows from previous claims that this also holds for the IS. The IS (at least as presented by Dennett) appears to be a rather underdeveloped predictive strategy. This is because, unless it is specified exactly what it means for a system to produce rational actions given its current beliefs and desires, it is unclear how the IS can provide a comprehensive strategy for predicting the system's next action. Therefore, stating that one adopts the IS towards the robot does not reveal much about the action they will expect the robot to perform. In the absence of a clear account of rationality attached to the IS, it is reasonable to assume that different observers may have different opinions on how the most rational action should be calculated and, accordingly, make different behavioural predictions.

only whether they attribute mental states to it, but also their view of the connection between the robot's behaviour and its beliefs and desires.

Before concluding this section, it is worth reflecting on the stance adopted when modelling the relationship between a robot's mental states and behaviour using non-rational, law-like regularities. Since this modelling strategy does not involve attributing rationality, it cannot be classified as an intentional stance. It is also not the physical stance, since it does not use the language of physics. The third option in Dennett's framework is the design stance. However, it is the IS, not the design stance, that involves attributing beliefs and desires to the system. When taking the design stance, predictions are made solely from knowledge or assumptions about the system's design. Should this then be regarded as a fourth stance, located somewhere between the intentional and design stances? The rather vague way in which Dennett's stances have been defined in the literature does not help to answer this question. One hypothesis is that attributing mental states in terms of law-like regularities rather than rationality constitutes a kind of mentalistic design stance that Dennett does not explicitly discuss.

Is this stance likely to be adopted by ordinary people? This is an empirical question, but there are reasons to suggest that it will be. After all, it is reasonable to expect laypeople to view robots as designed systems that operate in a mechanical, law-like manner. However, it is also likely that they will consider robots to have content-bearing states and internal representations of goals that correspond to beliefs and desires in commonsense psychology. The possibility that laypeople adopt a mentalistic design stance similar to the IS in the attribution of mental states but differing from it in assuming that the interaction between these mental states is mechanical or law-like is not to be excluded and should merit empirical investigation. It should be noted that a mentalistic design stance towards artificial intelligence systems, which is more akin to the explanatory approach of Marr-like cognitivism than to that of propositional-attitude psychology, has been proposed by Larghi and Datteri (2024) on different grounds.

3. Testing the intentional stance in human-robot interaction

The time has come to apply the considerations made so far to contemporary research in human-robot interaction. As expected, many empirical studies have attempted to determine people's mental models of robots and the factors that shape them over time. Although the literature is relatively new, it is impressively rich in terms of methods and results. Some studies have employed qualitative methods: for instance, Rueben et al. (2021) observed and interviewed six individuals interacting with a robotic shoe rack for six weeks in a yoga class, with the

aim of studying their mental models of the robot and how these evolve. The interview questions were deliberately chosen to be very general in order to elicit the users' perceptions in the most neutral way possible. This exploratory study mostly resulted in interesting hypotheses and research questions being generated, such as: how does the model change over time? Why do users sometimes decide to avoid experimenting with the robot?

In this growing body of literature, there are some studies that explicitly refer to Dennett's intentional systems (IS) theory. These studies attempt to determine the factors that influence people's adoption of IS towards robots and the consequences of this adoption on other phenomena. This section considers this category of studies exclusively (see Thellman et al. 2022 for a systematic review). The underlying question is whether these studies regard the adoption of an IS as a phenomenon involving the attribution of mental states and rationality to the system. While these studies make significant contributions to the analysis of people's mental models of robots, it will be argued that they neglect the issue of rationality to a certain extent. Some methods have been devised in the non-robotic literature to study the attribution of rationality (e.g. Gergely et al. 1995). As emphasised throughout this article, overcoming this limitation would be beneficial: understanding how people form their expectations about robots' behaviour requires determining the nature of the connection they perceive between mental states and behaviour. For instance, does this connection adhere to principles of rationality or is it merely based on nomic interactions between the system's various mental states?

Some studies explicitly adopting Dennett's framework focus on the consequences of adopting an IS towards robots in relation to other cognitive, neural, or behavioural phenomena. Examples include studies on how IS adoption affects gaze behaviour. Gaze cueing phenomena are integral to social interaction between people. For instance, they occur when a person's gaze directs another person's visual attention. The aforementioned study by Wiese and colleagues (2012) investigated whether people can be cued by the gaze of robots. The researchers designed a task in which participants had to discriminate between two stimuli appearing on the left or right side of a robotic face. They investigated whether the gaze direction of the robot (pointing left or right) affected the error rate. Specifically, they asked whether the error rate would change when participants adopted an IS towards the agent providing the gaze cues. The working hypothesis was that it would, as people's attention cannot be directed by the gaze of things that they do not perceive as intentional systems. Desideri et al. (2021) studied gaze aversion instead. People often tend to "look away" from potentially distracting stimuli when thinking in order to save cognitive resources. In particular, people look away more often when facing social stimuli, as these are cognitively demanding. Therefore, the hypothesis underlying

the study was that adopting an IS towards a robot would increase the effects of gaze aversion when facing the robot. To determine whether an IS affects gaze cueing and gaze aversion in the two studies, the researchers needed to manipulate the participants' mental models of the agent with which they were interacting in the task. They needed to create experimental conditions in which participants either took or failed to take an intentional stance towards the robot. They did so *indirectly*. In some sessions, participants interacted with either a human or a robotic face, and the effects of gaze cuing or gaze aversion were compared. In other sessions, participants only interacted with a robotic face and were explicitly told whether it was controlled by a human or an algorithm. They manipulated not only the perceptual features of the stimulus (human vs. robotic face) but also attempted to shape the participants' mental model of the robot by explicitly informing them that the robotic face was controlled either by a human or by an artificial agent. While the results supported the authors' hypotheses, showing that gaze-cueing and gaze-aversion effects increased when an intentional stance had been induced, it is clear that the authors did not directly manipulate the participants' mental model of the robot, but only the likelihood of the participants adopting an IS, without ensuring that the manipulation was successful. And their analysis was restricted to the attribution of mental states to the robot, which, as previously mentioned, is only 'half' of the IS.

Other studies have adopted a neuroscientific approach to investigate what happens in the brain when people adopt an IS towards a robot. For example, Chaminade and colleagues (2012) analysed the brain activity of subjects interacting with different agents – a human, a small humanoid robot with artificial intelligence and a random number generator – during a game of rock, paper, scissors. As in previous studies, the different conditions involved variations in the perceptual characteristics of the agent and the specific instructions given to participants. Participants were informed that in the second condition, the humanoid robot was intelligent and had a strategy to win the game, whereas in the third condition, the agent simply acted randomly. The results suggested that certain areas (the medial prefrontal cortex and the temporoparietal junction) responded only to the human, while other parietal and frontal areas responded more to the humanoid robot than to the random number generator, but less than to the human. Crucially, "brain areas involved in adopting an intentional stance in a social interaction were not recruited when interacting with an artificial intelligence" (8). While other studies have complemented and partially revised these results (see Özdem et al. 2017, for example; see also Perez-Osorio and Wykowska 2020; Wiese et al. 2017; Wykowska 2020; Wykowska et al. 2016) the key issue here is methodological. Comparing fMRI activity when participants interact with humans or robots, or when they are informed whether their partner has a strategy, can

significantly contribute to analysing the brain processes involved in interacting with different kinds of agents. However, as in previous studies, presenting a human or robotic face or instructing participants that the face stimulus is governed by a human or random number generator does not reveal whether participants attribute mental states and rationality to the agent. While these studies are extremely interesting from a scientific perspective, it is unclear how they can inform our understanding of how people form their expectations about the behaviour of the robots they are interacting with.

While the studies discussed so far have attempted to determine the consequences of adopting an IS towards a robot, other studies have aimed to establish whether people's adoption of the IS depends on the robot's morphology, behaviour, or other contextual factors. As IS adoption is the dependent variable in these studies, rather than an experimental condition, tools are needed to assess whether participants adopt the IS or not. Terada and colleagues (2007) devised a solution consisting of describing the three stances in Dennett's intentional systems theory (intentional, design, and physical) to the participants and letting them choose which they preferred. Interestingly, the authors found that the intentional stance was preferred when the robot (a motorised wheelchair) displayed reactive behaviour rather than non-reactive periodic behaviour. The article does not specify how the intentional stance was described to the participants, nor does it address the question of whether they attributed rationality to the robot. Mandell and colleagues (2017) are more explicit about the tool they used. In an attempt to study the relationship between morphological features and IS adoption, they presented participants with pictures of faces displaying different "degrees of physical humanness" by morphing a human face into a robot face in small increments. The questionnaire included questions such as "Rate how much this face looks like it has a mind" and "Do you think this agent would feel pain if it tripped and fell on hard ground?". Similarly, in (Martini et al. 2015, 2016), participants were asked to ask questions about whether agents with varying degrees of physical human-likeness possessed a mind and emotions. However, none of the questions concerned the attribution of rationality, and the questions were too general to determine how the mental states attributed to the robot were connected to its behaviour in the participants' mental models.

The Instance tool, devised by Marchesi and colleagues (2019) is perhaps the most advanced questionnaire developed so far for studying people's adoption of an intentional stance towards robots. The authors deserve credit for directly addressing the problem of operationalising the IS as a philosophical construct. For this reason, the questionnaire has been used in a large number of subsequent studies (Bossi et al. 2020; Roselli et al. 2023; Spatola et al. 2022). The questionnaire comprises 34 fictional scenarios in which the iCub robot performs simple

activities while interacting with people and objects. Each scenario is illustrated with a sequence of three photographs. Participants must choose between two possible descriptions of each scenario by moving a slider: one formulated in mentalistic terms and the other in non-mentalistic terms. For example, one of the scenarios shows iCub gazing at a ball in different positions on a table with a pyramid and a cube also present. The participant must choose between the descriptions "iCub categorises objects by their shape" and "iCub likes round objects", which are considered non-mentalistic and mentalistic, respectively. Another example is a sequence of three photographs showing iCub playing cards with a human. In the middle photograph, the human looks distractedly away from the robot and iCub leans towards his cards. In the final photograph, iCub returns to the initial position. The mentalistic description is "iCub was trying to cheat by looking at his opponent's cards", and the non-mentalistic description is "iCub was unbalanced for a moment". In half of the scenarios, the mentalistic sentence is on the right side of the slider and the non-mentalistic sentence is on the left. In the other half of the scenarios, it is the other way around. The order in which the scenarios are presented to each participant is randomised. To calculate the score, the non-mentalistic–mentalistic scale is converted into a numerical scale from 0 to 100, and the corresponding scores for each answer are averaged.

Like all questionnaires in psychology, the Instance suffers from the obvious limitation: selecting the mentalistic option does not imply that one's beliefs about the robot conform to this choice. People may give answers that do not fully reflect their beliefs. This is a general problem affecting all questionnaires. Indeed, when the questionnaire is used to study the IS, as with the Instance, the situation becomes even more complicated. This is because the intentional stance is an instrumentalist rationalisation of the behaviour of the target system. The *attribution* of a belief to a robot does not imply that the person *believes* the robot has this belief (see Datteri 2025 for a discussion of the concept of "attribution" in this literature). In general, scientists researching the adoption of an IS towards robots are well aware of the distinction between what Thellman and Ziemke (2019) refer to as the "attribution question" (what mental states do people attribute to robots?) and the "belief question" (what mental states do people really believe robots have?). In this instance, it must be acknowledged that the participant's choice of "iCub was trying to cheat by looking at the opponent's cards" does not imply that they *attributed* the desire to cheat to iCub, nor that they *believed* iCub wanted to cheat the opponent. Therefore, to make sensible use of the Instance test, it is necessary to assume that the participants' answers are sincere and not influenced by factors such as the desire to please or avoid disappointing the experimenter ("I would like to choose the mentalistic option, but I am embarrassed because the experimenter might think I believe robots have minds and laugh

at my naivety"). Assuming this can be done, the participants' choices can reflect their instrumental attributions *or* their 'real', inner beliefs; the Instance does not discriminate between the two. This is arguably a problem only if this distinction is relevant to the research question for which the Instance is used, which seems rarely, if ever, to be the case.

The Instance Test is currently the most elaborate quantitative and explicit tool for studying whether people adopt an IS towards robots. The idea that the choice of mentalistic description signals the adoption of an IS towards iCub is intuitively plausible. In one scenario, for example, a human points at a ball and iCub picks it up and gives it to her. The mentalistic description is "iCub understood that the girl wants the ball", while the non-mentalistic description is "iCub tracked the girl's hand movements". These two descriptions seem very different at first glance: 'understanding' is undoubtedly a mentalistic term, whereas 'tracking' alludes to a behavioural reaction to an external stimulus, conceived in terms of stimulus-response mechanisms without the mediation of beliefs and goals. It is worth noting that the descriptions included in the Instance questionnaire were pre-tested with several participants who had a philosophical background to ensure that they could sensibly be regarded mentalistic or not, depending on the case. Nevertheless, regardless of the philosophers' opinions during pretesting, it could be argued that verbs such as 'tracking' *are* in a sense mentalistic. This is because, even if 'tracking' is interpreted as referring only to the movements of the robot's head, it also implicitly refers to an *internal representation* of these movements: iCub's tracking of a movement consists of it having an internal representation whose content changes consistently as the object moves. Assuming that the robot moves its head according to an internal representation of a vehicle is dangerously similar to assuming that it is in a functional relationship with a representation with content, or in other words, that it possesses a certain propositional attitude. This observation is reminiscent of Pylyshyn's comment on behaviourism (Pylyshyn 1989). Pylyshyn recalled that behaviourists built a psychology out of notions such as stimuli, responses and reinforcements in an attempt to eliminate mentalistic terms. However,

> such categories are cognitive: What serves as the functional stimulus depends on how a person interprets the situation (for example, the stimulus in the pedestrian-automobile example is *accident*; but, of course, if that person is told it is a rehearsal for a television show, the stimulus is no longer *accident* but *rehearsal* and engages the habits appropriate for that category). Similarly, what constitutes the response is also implicitly cognitive. Some particular bit of movement (accidentally bumping into a telephone while in a booth keeping out of the rain) does not count as a "response", only movements intended a certain way are counted. (9)

These considerations show that it is not clear that, when a participant chooses "iCub tracked the girl's hand movement", they are not

attributing mental states to the system, as the study by Marchesi and colleagues seems to assume. This is not (only) because questionnaires generally do not detect people's inner beliefs or instrumental attributions, but (also) because attributing a 'tracking system' is, from a certain perspective, attributing internal states with content. This is consistent with the thesis of Larghi and Datteri (2024) that people may form 'cognitivist' mental models of robots that differ structurally from IS and propositional-attitude psychology while still being mentalistic.

Granted that at least some of the 34 questions in the Instance test offer only mentalistic alternatives, can the Instance questionnaire distinguish between participants who do and do not attribute rationality to the system? Interestingly, the term 'mechanistic' is used by the authors to indicate non-mentalistic descriptions in the questionnaire. For example, "iCub tracked the girl's hand movement" is considered a mechanistic description, whereas 'iCub was trying to cheat by looking at the opponent's cards' is considered a mentalistic description. The use of these terms presupposes that mentalistic descriptions cannot be mechanistic. However, this assumption would require justification, taking into account the extensive literature on the structure of mechanistic explanations that has dominated philosophy of science since the beginning of the XXIst century (see Glennan and Illari 2015 for a comprehensive essay; see Bechtel, 2008 for a specific discussion of mental mechanisms). In a sense, attributing *mentalistic* law-like regularities to systems such as those expressed by assumptions (C) and (D) contributes to formulating a mental model of the robot that is more mechanistic (and mentalistic) than not. Mechanisms operate regularly from triggering to termination conditions, and this is precisely what robotic vacuum cleaners are designed to do when governed by these mentalistic regularities. In principle, this does not preclude the possibility that a rational system also operates mechanically. If all the relevant details of a decision-making mechanism are specified and there are reasons to believe that, according to a particular theory of rationality, the mechanism will consistently make the most rational decision, then a mental model that attributes the mechanism to the robot will be both rational and mechanistic (but see Searle 2001, for an alternative view that asserts rationality necessitates free will). In summary, mentalistic descriptions can be mechanistic, thus calling into question the mentalistic vs. mechanistic distinction made in the Instance test. It is not obvious that, by choosing "iCub was trying to cheat by looking at the opponent's cards", a participant is not treating the system as a mechanism, nor is it obvious that, by choosing "iCub tracked the girl's hand movements", they are not treating the system as non-mentalistic. The Instance test is, thus, not well-equipped to determine the nature and structure of the relationship that participants see between the robot's mental states and its overt behaviour. This could be a rational decision-making system (analysable in mechanistic terms or not), an irrational one, or simply a mechanism made up of law-like regularities connect-

ing the various mental states of the system with its overt behaviour. While the Instance test is one of the most advanced tools for exploring people's mental models of robots, it suffers from the same limitation as other attempts to determine people's adoption of an IS towards robots in that it only addresses 'half' of the issue. Understanding the missing half would be very helpful in terms of gauging people's expectations about the future behaviour of the robots they interact with.

4. Concluding remarks

This article addresses an emerging area of research in human-robot interaction from the perspective of the philosophy of science. Recently, attempts have been made to operationalise a construct originating from the philosophy of mind to deepen our understanding of how people perceive robots in everyday interactions. While other theoretical frameworks are occasionally adopted in this literature – most frequently the so-called 'theory of mind', as in (Thellman and Ziemke 2020) – Dennett's intentional systems theory has attracted significant attention among social roboticists. These researchers have developed experimental methods to detect when and if people adopt an intentional stance towards robots. These methods have supported the hypothesis that certain morphological and behavioural characteristics of robots greatly influence whether people 'see' beliefs, desires and other mental states behind their behaviour. Quantitative tools are now available to explore the idea that familiarity with technology affects the attribution of mental states to robots and to establish correlations between the attribution of mental states and other human-robot interaction phenomena, such as the sense of agency (Roselli et al. 2022). These studies facilitate dialogue between roboticists and philosophers, with the former paying increasing attention to the contributions of philosophy of mind and science to our understanding of the nature and structure of the mind.

This article has attempted to convey the message that these methods are generally not yet well equipped to study the adoption of an intentional stance towards robots. Rather than studying the adoption of an intentional stance towards robots, they should be regarded as methods of studying the attribution of mental states, such as beliefs and desires, to machines. The intentional stance incorporates the assumption that robots will act rationally based on these mental states, which goes beyond mere attribution. People may mentalise other entities by attributing mental states to them without assuming that their behaviour is rational. This may not concern roboticists if their only aim is to determine how people's low-level, reflexive, bottom-up reactions to robots depend on their immediate understanding of their 'inner life', prior to any form of rationalisation. However, if the aim is to analyse people's expectations of robot behaviour and their reactions to it, it is important to determine how people think the robots' mental states are connected to their behaviour. Rationality is not the only option. People may as-

sume that the robot's behaviour is rational, occasionally irrational, or they may identify a mental mechanism defined by law-like regularities between the robot's mental states and behaviour without any reference to rationality. Admitting the possibility of these further options means acknowledging that Dennett's intentional systems theory, with its key reference to rationality, is a somewhat limited framework for studying people's understanding of robots.

What methods could be employed to study the attribution of rationality to robots? At the very least, a precise but potentially limited definition of rationality would be required. Gergely and colleagues (1995) devised an experiment in which different groups of 12-month-old participants were habituated to a ball displaying rational and non-rational behaviour, respectively. In both cases, the ball started at point A and had to reach point B. In the 'rational' condition, an obstacle was placed between A and B, and the ball followed a curved path to circumvent it. In the 'irrational' condition, the obstacle was placed *behind* the ball so that it did not constitute an obstacle; however, the ball followed the curved path anyway, displaying irrational behaviour. Thus, in both cases, the ball followed a curved path; the only difference was the presence of an obstacle between the two points. After habituation, both groups were presented with a different situation in which there was no obstacle between A and B. In this situation, the ball either followed a straight path (rational behaviour) or a curved path (non-rational behaviour). It was found that participants who had been habituated to rational behaviour were more surprised when the ball followed a curved path than when it followed a straight path. Note that they had not been habituated to the linear path; in the habituation phase, they observed the ball following *a curved path* to circumvent an obstacle; they were in fact exposed to the very same trajectory shown in the test situation. Their increased level of surprise when the ball displayed the same behavioural trajectory as in the habituation phase suggests that they had interpreted the ball as not only having the goal of reaching B, but also as a rational entity in that phase. In their own words, the authors state that "the results of the … habituation study provide independent empirical support for the general conjecture that by the end of the first year infants are indeed capable of taking the intentional stance (Dennett 1987) in interpreting the goal-directed behavior of rational agents" (184). The intentional stance is an appropriate reference here, as the authors test both the mentalistic and rationality assumptions. The notion of rationality adopted here is clearly relatively narrow: to be rational is to follow the shortest path while avoiding obstacles. Can this task, or others from the psychological literature, be adapted to study the attribution of rationality to robots? This question is challenging and, for the reasons shown here, important for understanding what people expect robots to do and how they interact with robots in ethically sensitive situations.

References

Anderson, J. R. 1991. "Is Human Cognition Adaptive?" *Behavioral and Brain Sciences* 14 (3): 471–485.

Bechtel, W. 2008. *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. London: Routledge.

Bossi, F., C. Willemse, J. Cavazza, S. Marchesi, V. Murino, and A. Wykowska. 2020. "The Human Brain Reveals Resting State Activity Patterns that Are Predictive of Biases in Attitudes toward Robots." *Science Robotics* 5 (46): eabb6652.

Chaminade, T., and G. Cheng. 2009. "Social Cognitive Neuroscience and Humanoid Robotics." *Journal of Physiology-Paris* 103 (3–5): 286–295.

Chaminade, T., D. Rosset, D. Da Fonseca, B. Nazarian, E. Lutcher, G. Cheng, and C. Deruelle. 2012. "How Do We Think Machines Think? An fMRI Study of Alleged Competition with an Artificial Intelligence." *Frontiers in Human Neuroscience* 6.

Ciardo, F., F. Beyer, D. De Tommaso, and A. Wykowska. 2020. "Attribution of Intentional Agency towards Robots Reduces One's Own Sense of Agency." *Cognition* 194: 104109.

Datteri, E. 2025. "Folk-Ontological Stances Towards Robots and Psychological Human Likeness." *International Journal of Social Robotics* 17 (2): 257–276.

Dennett, D. C. 1971. "Intentional Systems." *The Journal of Philosophy* 68 (4).

Dennett, D. C. 1978. *Brainstorms: Philosophical Essays on Mind and Psychology*. Cambridge: MIT Press.

Dennett, D. C. 1989. *The Intentional Stance*. Cambridge: MIT Press.

Dennett, D. C. 2009. "Intentional Systems Theory." In A. Beckermann, B. P. McLaughlin, and S. Walter (eds.). *The Oxford Handbook of Philosophy of Mind*. Oxford: Oxford University Press, 339–350.

Desideri, L., P. Bonifacci, G. Croati, A. Dalena, M. Gesualdo, G. Molinario, A. Gherardini, L. Cesario, and C. Ottaviani. 2021. "The Mind in the Machine: Mind Perception Modulates Gaze Aversion During Child–Robot Interaction." *International Journal of Social Robotics* 13 (4): 599–614.

Gergely, G., Z. Nádasdy, G. Csibra, and S. Bíró. 1995. "Taking the Intentional Stance at 12 Months of Age." *Cognition* 56 (2): 165–193.

Glennan, S., and P. Illari (eds.). 2015. *The Routledge Handbook of Mechanisms and Mechanical Philosophy*. London: Routledge.

Larghi, S., and E. Datteri. 2024. "Mentalistic Stances Towards AI Systems: Beyond the Intentional Stance." In A. Aldini (ed.). *Lecture Notes in Computer Science 14568*. Cham: Springer, 28–41.

Mandell, A. R., M. Smith, and E. Wiese. 2017. "Mind Perception in Humanoid Agents Has Negative Effects on Cognitive Processing." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 61 (1): 1585–1589.

Marchesi, S., D. Ghiglino, F. Ciardo, J. Perez-Osorio, E. Baykara, and A. Wykowska. 2019. "Do We Adopt the Intentional Stance Toward Humanoid Robots?" *Frontiers in Psychology* 10: 450.

Marchesi, S., K. Kompatsiari, D. De Tommaso, and A. Wykowska. 2025. "Adopting the Intentional Stance Affects Social Attention when Interacting with a Humanoid Robot." *Technology, Mind, and Behavior* 6 (2).

Martini, M. C., G. A. Buzzell, and E. Wiese. 2015. "Agent Appearance Modulates Mind Attribution and Social Attention in Human–Robot Interaction." In A. Tapus, E. André, J.-C. Martin, F. Ferland, and M. Ammi (eds.). *Social Robotics*. Cham: Springer, 431–439.

Martini, M. C., C. A. Gonzalez, and E. Wiese. 2016. "Seeing Minds in Others – Can Agents with Robotic Appearance Have Human-Like Preferences?" *PLOS ONE* 11 (1): e0146310.

Mele, A. R., and P. Rawling (eds.). 2004. *The Oxford Handbook of Rationality*. Oxford: Oxford University Press.

Özdem, C., E. Wiese, A. Wykowska, H. Müller, M. Brass, and F. Van Overwalle. 2017. "Believing Androids – fMRI Activation in the Right Temporo-Parietal Junction is Modulated by Ascribing Intentions to Non-human Agents." *Social Neuroscience* 12 (5): 582–593.

Perez-Osorio, J., and A. Wykowska. 2020. "Adopting the Intentional Stance toward Natural and Artificial Agents." *Philosophical Psychology* 33 (3): 369–395.

Peterson, M. 2017. *An Introduction to Decision Theory*. 2nd ed. Cambridge: Cambridge University Press.

Pylyshyn, Z. W. 1989. *Computation and Cognition: Toward a Foundation for Cognitive Science*. Cambridge: MIT Press.

Roselli, C., F. Ciardo, D. De Tommaso, and A. Wykowska. 2022. "Human-likeness and Attribution of Intentionality Predict Vicarious Sense of Agency over Humanoid Robot Actions." *Scientific Reports* 12 (1): 13845.

Roselli, C., S. Marchesi, D. De Tommaso, and A. Wykowska. 2023. "The Role of Prior Exposure in the Likelihood of Adopting the Intentional Stance toward a Humanoid Robot." *Paladyn: Journal of Behavioral Robotics* 14 (1).

Rueben, M., J. Klow, M. Duer, E. Zimmerman, J. Piacentini, M. Browning, F. J. Bernieri, C. M. Grimm, and W. D. Smart. 2021. "Mental Models of a Mobile Shoe Rack: Exploratory Findings from a Long-term In-the-Wild Study." *ACM Transactions on Human–Robot Interaction* 10 (2): 1–36.

Searle, J. R. 2001. *Rationality in Action*. Cambridge: MIT Press.

Spatola, N., S. Marchesi, and A. Wykowska. 2022. "Different Models of Anthropomorphism across Cultures and Ontological Limits in Current Frameworks: The Integrative Framework of Anthropomorphism." *Frontiers in Robotics and AI* 9: 863319.

Stich, S. P. 1981. "Dennett on Intentional Systems." *Philosophical Topics* 12 (1): 39–62.

Terada, K., T. Shamoto, A. Ito, and H. Mei. 2007. "Reactive Movements of Non-humanoid Robots Cause Intention Attribution in Humans." In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 3715–3720.

Thellman, S., M. De Graaf, and T. Ziemke. 2022. "Mental State Attribution to Robots: A Systematic Review of Conceptions, Methods, and Findings." *ACM Transactions on Human–Robot Interaction* 11 (4): 1–51.

Thellman, S., and T. Ziemke. 2019. "The Intentional Stance Toward Robots: Conceptual and Methodological Considerations." In A. K. Goel, C. M. Seifert, and C. Freska (eds.). *Proceedings of the 41st Annual Conference of the Cognitive Science Society*. Austin: Cognitive Science Society, 1097–1103.

Thellman, S., and T. Ziemke. 2020. "Do You See what I See? Tracking the Perceptual Beliefs of Robots." *iScience* 23 (10): 101625.

Wason, P. C. 1968. "Reasoning about a Rule." *Quarterly Journal of Experimental Psychology* 20 (3): 273–281.

Wiese, E., G. Metta, and A. Wykowska. 2017. "Robots as Intentional Agents: Using Neuroscientific Methods to Make Robots Appear More Social." *Frontiers in Psychology* 8.

Wiese, E., A. Wykowska, J. Zwickel, and H. J. Müller. 2012. "I See What You Mean: How Attentional Selection Is Shaped by Ascribing Intentions to Others." *PLoS ONE* 7 (9): e45391.

Wykowska, A. 2020. "Social Robots to Test Flexibility of Human Social Cognition." *International Journal of Social Robotics* 12 (6): 1203–1211.

Wykowska, A., T. Chaminade, and G. Cheng. 2016. "Embodied Artificial Agents for Understanding Human Social Cognition." *Philosophical Transactions of the Royal Society B: Biological Sciences* 371 (1693): 20150375.

Ziemke, T. 2020. "Understanding Robots." *Science Robotics* 5 (46): eabe2987.