Minds and Machines (2021) 31:505–534
https://doi.org/10.1007/s11023-021-09567-6

ORIGINAL ARTICLE

A Pragmatic Approach to the Intentional Stance Semantic, Empirical and Ethical Considerations for the Design of Artificial Agents

Guglielmo Papagni[1] · **Sabine Koeszegi**[1]

Received: 31 December 2020 / Accepted: 5 July 2021 / Published online: 26 July 2021

Abstract
Artificial agents are progressively becoming more present in everyday-life situations and more sophisticated in their interaction affordances. In some specific cases, like Google Duplex, GPT-3 bots or Deep Mind's AlphaGo Zero, their capabilities reach or exceed human levels. The use contexts of everyday life necessitate making such agents understandable by laypeople. At the same time, displaying human levels of social behavior has kindled the debate over the adoption of Dennett's 'intentional stance'. By means of a comparative analysis of the literature on robots and virtual agents, we defend the thesis that approaching these artificial agents 'as if' they had intentions and forms of social, goal-oriented rationality is the only way to deal with their complexity on a daily base. Specifically, we claim that this is the only viable strategy for non-expert users to understand, predict and perhaps learn from artificial agents' behavior in everyday social contexts. Furthermore, we argue that as long as agents are transparent about their design principles and functionality, attributing intentions to their actions is not only essential, but also ethical. Additionally, we propose design guidelines inspired by the debate over the adoption of the intentional stance.

Keywords Intentional stance · Robots · Virtual agents · Ethics

1 Introduction

Artificial agents (i.e., physical robots, virtual agents and embodied virtual agents) capable of taking decisions autonomously are being employed in a growing number of fields. Several applications of such agents directly influence everyday life for a growing number of people, most of whom can be considered non-expert end-users.

✉ Guglielmo Papagni
guglielmo.papagni@tuwien.ac.at

[1] Institut für Managementwissenschaften, TU Wien, Theresianumgasse 27, Vienna 1040, Austria

To properly address the challenges posed by the integration and acceptance of artificial agents within society, coordinating the contributions of various disciplinary fields is necessary and a challenge within the challenge.

Being able to correctly understand and predict the behavior of artificial agents is necessary for the development of trustworthy relationships (De Graaf & Malle, 2017; Miller, 2019). In this regard, one fundamental feature addressed by the ongoing interdisciplinary efforts concerns whether people consider artificial agents' actions and decisions as intentional. Contributions on this topic are embedded into the broader framework of discussion over the attribution of anthropomorphic and social traits to artificial agents (Nass et al., 1994; Reeves & Nass, 1996; Dreyfus et al., 2000; Nass & Moon, 2000). However, in recent times, the idea that artificial agents' actions might be interpreted as intentional is emerging as a structured and semi-autonomous debate. A significant share of research on this issue is informed by Daniel Dennett's concepts of 'intentional systems' and 'intentional stance' (Dennett, 1971, 1981, 1989). The latter represents a strategy that people can adopt to make sense of and predict the behavior of complex rational (or intentional) systems, being them human agents or machines (Dennett, 1988). As the inner complexity of artificial agents grows, researchers in the fields of human–robot and human–computer interaction (HRI and HCI respectively) are investigating from multiple perspectives how people ascribe intentions to artificial agents.

Objections to the concept of intentional stance have been raised over the years. We recognize part of this criticism, specifically that which targets Dennett's commitment to a behaviorist perspective as reasonable. However, some of these objections tend to conflate biological definitions of intentionality with the attribution of intentions to artificial entities (Thellman et al., 2017), hence overshadowing what we consider Dennett's most relevant contribution. Our aim is to build upon the originality of his intuition without committing to Dennett's behaviorism. First, this paper provides a critical analysis of Dennett's concept, particularly in light of definitions of intentionality and some of the main objections. Moving beyond Dennett's position, we emphasize that the ascription of intentionality does not necessarily imply artificial agents to have genuine mental states in the human sense. Rather, we claim that the main quality of this strategy is to help people to manage social interactions with artificial agents, and that research should focus on how to maximize this positive aspect.

Furthermore, some scholars note a pressing lack of comparative analysis (Thellman et al., 2017). We argue that only through such a systematic approach it is possible to trace the point of origin of the attribution of intentions to artificial agents. Therefore, the paper analyzes relevant cases from the HRI and HCI experimental literature. From this comparative perspective, we discuss and refine Dennett's idea that complex rational behavior is the spark that ignites the process of intention ascription.

The idea of deceptive anthropomorphism offers another criticism of features that trigger the attribution of intention. The last section investigates ethical implications of the concept, particularly in light of recent calls for machine transparency and the risks of 'deceptive design'. Our claim is that the implementation of features that make artificial agents' behavior 'seemingly intentional' is not only ethical, but can

be desirable, as it can positively contribute to the overall quality of social interactions. However, we also identify a key condition for this assumption to hold. Users must be made aware of the nature of the agent they are interacting with before the interaction unfolds. Failing to fulfil this condition, as our examples show, can have negative consequences that could jeopardize the successful societal integration of artificial agents.

2 Critical Approach to the Intentional Stance

Each of the stances discussed by Dennett represents a strategy for understanding and predicting how certain entities work (Dennett, 1981, 1988, 1989). People adopt the physical stance to make sense of and predict the future behavior of certain systems, via their knowledge of physical laws. The design stance allows predictions to be made based on the assumption that systems work as they are meant to by design. In certain cases, however, these two strategies do not suffice. In particular, it may not be possible, let alone practical, to predict how rational systems (or agents) will behave based on the two previous strategies. To address this limitation, Dennett introduces the intentional stance. This is a predictive tool that relies on the assumption that rational systems will behave in accordance with their intentions, beliefs and desires in order to achieve specific goals (Dennett, 1981, 1988, 1989).

Dennett is neither the only one, nor the first to refer to artificial agents in "anthropomorphic" or social terms. One of the most relevant contributions in this direction is represented by the pioneering study from Heider and Simmel on the attribution of social meanings to the motion of geometric shapes (Heider & Simmel, 1944). Ever since the initial establishment of computing-related disciplines, researchers have spent significant efforts in trying either to make artificial agents appear and behave more like humans, or to explain why people tend to adopt social interpretative frameworks to understand and predict artificial agents' actions (Caporael, 1986; Nass et al., 1994; Breazeal & Scassellati, 1999; Breazeal, 2002).

2.1 Complexity of Intentional Systems

After this initial contextualization of his work, it is important to note that the reason why Dennett includes (certain types of) artificial agents as targets of the intentional stance lies in the complexity of these systems. In fact, referring to a chess-playing computer, Dennett says that such systems are "practically inaccessible to prediction from either the design stance or the physical stance; they have become too complex for even their own designers to view from the design stance." Therefore, one assumes "that the computer will 'choose' the most rational move" (Dennett, 1981, p. 5). In other words, Dennett emphasizes the idea that treating certain types of agents 'as if' they had intentions might in some cases be the only fruitful strategy to understand and predict their behavior. In fact, he continues, "when one can no longer hope to beat the machine by utilizing one's knowledge of physics or programming to anticipate its responses, one might still be able to avoid defeat by

treating the machine rather like an intelligent human opponent" (Dennett, 1981, p. 5). Again, elsewhere he clarifies that we adopt the intentional stance because "it gives us predictive power we can get by no other method" (Dennett, 1997, p. 66). At the same time, he also points out the difference between "those intentional systems that really have beliefs and desires from those we may find it handy to treat as if they had beliefs and desires" (Dennett, 1997, p. 66).

However, other positions expressed by Dennett on the topic contribute to further articulating the debate and give rise to some of the main critiques. In fact, in several occasions Dennett remarks the fact that if one were to infer and attribute any mental states to another agent, doing so through the analysis of the observable behavior of the agent would be the only way to go (Dennett, 1991, 1993). Dennett claims that there is no ineffable quality of the mind and that mental states can be discerned through the recognition and analysis of behavioral patterns. Furthermore, he argues that not only robots and artificial agents can be, in principle, referred to as "philosophical zombies", but that for the concerns of consciousness and mental states, everyone is such a zombie. This fictional entity is something that is functionally, i.e., behaviorally identical to a human being, but which lacks any form of actual consciousness (Dennett, 1993, 1995).

For the purposes of this paper, such considerations represent the problematic node in Dennett's framework. It is problematic because it implies the commitment to forms of "behavioral realism". Hence, from Dennett's perspective, either both humans and artificial agents can be considered conscious, at least as long as they behave in a qualitatively comparable way, or neither of them should. Not only that, the fact that Dennett subscribes to such positions triggers several critiques, so that the debate takes, at least partially, the direction of a diatribe on the overlapping of genuine intentionality (and mental states more in general) and the attribution of intentions to artificial agents (Thellman et al., 2017). The reason for this lies in the fact that, as Dennett poses that perceiving patterns of intentionality in the behavior of an agent corresponds to saying that those patterns are the one and only real thing, the only possible consequence is to match humans and artificial agents under the sign of the intentional stance. To clarify this fundamental point, the next paragraphs briefly discuss what in the literature is referred to as genuine intentionality, to then focus on some of the main critiques proposed against Dennett's arguments.

2.2 Biological Intentionality and Objections to the Intentional Stance

John Searle, for instance, refers to intentionality as a feature of an evolution-based mind that allows people to relate in the first place to each other, but also to the environment. "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." (Searle, 1980, p. 85) Similarly, other philosophical definitions emphasize aspects of mind's relatedness to the world

(Jacob, 2019) and one's mental states as a function of their goals and aims (Miller, 2019).

Consequently, it might sound reasonable that in order to interpret and predict the behavior of even sophisticated devices such as robots, or conversational agents, it would be enough to know the purpose behind their design. As artificial entities, they are not endowed with the evolution-based features that contribute to the emergence of biological intentionality. Hence, from this perspective, it would not make sense to adopt the intentional stance when interacting with such machines. However, as it was previously noted, Dennett explicitly refers also to artificial agents in his formulations of the intentional stance and of intentional systems.

One of the shortcomings of Dennett's theory is addressed as the "ideal rationality of intentional systems". To this extent, it is argued that intentional biological agents do not always behave in full accordance with the ideal rationality implied by Stich (1985). Indeed, notes Stich, irrationality is a cornerstone of human behavior (Stich, 1985). If an intentional agent in Dennett's terms is one that always acts rationally, then intentionality and rationality necessarily go together, and if one acts irrationally, one cannot be an intentional agent, argues Stich. His main point is that this is not a valid argument to say that if one's behavior is not fully rational, then no intentions, beliefs etc. can be attributed.

Another line of argument directly targets Dennett's 'behaviorist' positions concerning intentionality and intelligence. What is criticized is the idea that if something behaves 'as if' it were intelligent (e.g., having intentions), then it should be considered as such, precisely because intelligence can be identified only through manifest behavior (Dennett, 1995; Danaher, 2020). Block makes the point that human-like behavior is not sufficient to characterize an agent as having human-like intelligence because, as a matter of fact, such behavior does not mirror actual mental states or intelligence. Rather, it is merely the manifestation of its programmers' intelligence (Block, 1981). Such machines, continues Block, lack "the kind of 'richness' of information processing requisite for intelligence" (Block, 1981, p. 28). In a similar fashion, Slors notes that Dennett never clarifies what the adoption of the intentional stance means without referring to intentions and hence winding up in a kind of circular argument (Slors, 1996).

At this point, one might note how the debate tends to be oriented towards a resolution of the contrast between genuine conceptions of intentionality (and intelligence), and the adoption of the intentional stance. In other words, it tends to relinquish the pragmatic usefulness of attributing intentions to artificial agents, as they can only display something that looks like intentionality, but lack the very substratum, processes and semantic richness that make up genuine intentionality. Whereas this might partially be Dennett's own fault in light of his behaviorism, we argue that this conflating attitude is itself part of the problem (Thellman et al., 2017).

If one aims to move beyond Dennett's behaviorism, why should a property unique to the mind be attributed to artefacts without minds? Furthermore, if certain types of machines, i.e., artificial agents, can be treated as if they were intentional, where should the line be drawn between them and devices whose behavior can be predicted only as a function of their design? And how is this boundary (if one exists) defined? These are the main questions addressed in this paper. Arguably, their relevance is

not merely philosophical. Since it is believed that artificial agents will play a central role in our society, it is fundamental to figure out appropriate design strategies to improve interactions with them and avoid ethically dangerous trends.

Based on the previous considerations, we argue that, for the purpose of social interactions with artificial agents, the usefulness of Dennett's proposal should not concern whether a machine could have genuine intentions. As the above mentioned critiques have shown, supporting the intrinsic behaviorism of the intentional stance might create more problems than it solves.

2.3 Alternative Semantics, Alternative Approach

In line with the position expressed by Thellman and Ziemke, we claim that rather than focusing on and committing to hard-to-prove ontological statements about the nature of mental states, the attention should be shifted towards the idea that, from a user perspective, treating a sophisticated agent 'as if' it were intentional might be the most appropriate strategy (if not the only one available). Attributing intentions to machines should be more about the mental states of the one doing the ascribing rather than the mental states (or lack thereof) of the machine itself (Thellman & Ziemke, 2019). To this extent, recalling Searle's definition of intentionality, Thellman and colleagues note that, in addition to reading intentionality as a function of relatedness (of subjective states to the world), Searle also refers to recognizing others as intentional agents as fundamental to predicting how they will behave (Thellman et al., 2017). Such a complementary (in Searle's terms) perspective reflects what we claim to be the most fruitful aspect of Dennett's formulation. People attribute intentions not necessarily or not exclusively to recognize conspecifics, but rather to understand and predict how people (and agents) behave (Dennett, 1988). This seems to reflect "folk-psychological" definitions which consider intentionality not only objectively (i.e., biologically) but also as a social construct that functions as a 'tool' to ease social interactions and thus also to make sense of or predict the behavior of sophisticated artificial agents (Dennett, 1988; Malle & Knobe, 1997).

In support of this idea, studies suggest that the general attitude to anthropomorphize artificial agents might be the default approach that people adopt, as a socio-cognitive construct, when they recognize human-like patterns in other agents' (human or non-human) behavior (Caporael, 1986; Nass et al., 1994; Caporael & Heyes, 1997; Breazeal & Scassellati, 1999; Nass & Moon, 2000). Within this perspective, the attribution of mental states to other agents emerges as an automatic, bottom-up process caused by the activation of brain areas responsible for social cognition as a response to the perception of human-like patterns and traits (Buckner et al., 2008; Looser & Wheatley, 2010; Spunt et al., 2015). Therefore, while it is true that such mechanisms are rooted in social cognition processes that humans developed in order to interact with each other, they are available also when interacting with artificial agents. At the same time, recognizing patterns does not necessarily imply believing that they reflect the same mental states, or any mental state at all.

Hence, we claim that for the consideration of the intentional stance to be fruitful, it should be interpreted as a strategy that people adopt, consciously or not, to

navigate the world of social interactions with other rational agents, human or artificial. To avoid the risk of conflation of the intentional stance with biological definitions of intentionality, ontological statements and commitment about artificial agents' hard-to-prove mental states should be left aside accordingly. As Breazeal states, referring to the Kismet robot, people treat it "as if it were a socially aware creature with thoughts, intents, desires, and feelings. Believability is the goal. Realism is not necessary" (Breazeal, 2002, p. 52).

Therefore, we suggest that perhaps an alternative, machine-specific semantic approach is more appropriate. One might argue that people attribute or infer intentions to other humans as well, so that attributing intentions to machines does not involve any different process. In a way, such overlap cannot be avoided, as the mental processes involved in the persons doing the ascribing are qualitatively the same, with the only difference being one of quantity. However, in light of the previous discussion about distancing from Dennett's behaviorism, what does differ qualitatively is how biological intentions and machine seemingly intentional behavior are generated. Whether or not users attribute intentions to such agents is supported and, to a certain extent, made possible by certain design strategies (as suggested in Wiese et al. (2017)).

Therefore, we claim that the emphasis should be put on the artificial and implemented nature of the features that trigger the ascription of intention in order to significantly mark the difference to biological instances. For instance, one could adopt formulations like artificial agents' 'seemingly intentional behavior'. This alternative approach should be accordingly interpreted as a pragmatic measure for researchers to work with, rather than as a specific ontological declaration. Following this premise, this strategy aims to describe the visible result, in terms of resembling intentional behavior, of specific implemented features. Since the ultimate goal of this discussion is to improve the quality of the interactions users undertake with artificial agents, if an agent's behavior appears to be intentional, we sustain that it should be possible to describe it in such terms without risking to bring out the implications of the previously discussed trend of 'conflating notions'.

On the side of users' experience, we deem the use of terms such as 'attributed' or 'ascribed' intentionality more appropriate than others like 'perceived'. 'Perceiving' intentions recalls the idea of the perceptual apparatus that people are endowed with, which collects, processes and reconstructs data from the surroundings (Malle, 2011). For instance, when an event occurs, the observable changes in the environment are perceived by our senses, recorded and processed. When it comes to social perception, perceptual information is combined so that people form impressions of each other and base their mutual judgement. Depending on the type of event or action that is perceived and processed, different types of causes (or reasons) are attributed as the result of a deliberative process (Parkinson, 2012). While the idea is in principle the same for both "object perception" and "person perception" (in which case intentions might be involved), the latter situation poses a more complex challenge, as the data and possible causes to be analyzed are more nuanced (Malle, 2011). It must be noted that such processes might not always occur on a fully conscious level. Rather, we might expect that the more people engage in relationships with artificial agents, the more specific mental models will be activated subconsciously, as

such interactions start belonging to implicit social cognitive processes (Evans & Stanovich, 2013). Such considerations aim to highlight the mental states of the people interacting with the agent. They emphasize that people resort, consciously or not, to the specific strategy of attributing intentionality in order to understand and predict the behavior of complex social machines, when other alternatives fail or cannot apply. As such, one can refer to the ascription of intentions without mentioning beliefs, desires and intentions themselves, as Slors argues (Slors, 1996).

Furthermore, we note how formulations such as 'simulated intentionality' might be proposed, in line with the idea of 'simulated social interactions' as opposed to 'fictional interactions' formulated in Seibt (2017). While it might be true that a given implemented feature aims to simulate intentional human behavior, labeling an interaction or implemented characteristics as 'simulated' suggests a potential expression of a negative bias, as expressed by Turkle when saying that simulated feelings are not real feelings (Turkle, 2010). This might in turn generate negative feelings in users who perceive their engagement as genuine leading them to scale back involvement and interactions with these agents. We propose that a similar approach could undermine the quality of social relationships with agents.

As previously mentioned, another objection refers to the 'ideal rationality' of intentional agents. What we perceive as intentional behavior in machines does not necessarily have a non-rational counterpart, as it is the case for biological intentional agents. Stich notes how intentional systems in Dennett's sense are unavoidably rational. In most cases, robots and virtual agents are not designed to have the option of irrational behavior and what is perceived as such is likely caused by errors or malfunctions. Hence, we argue that to come to terms with Stich's position, a solution must be found that allows people to treat agents as intentional when they behave rationally as long as this is beneficial for the interaction, and yet to switch to an alternative mechanistic approach if they observe (apparently) irrational behavior. This idea is supported by Wiese and colleagues, who call for the development of design strategies that support interactive flexibility. In other words, it must be possible to alternate between intentional and mechanistic mental models depending on one's specific interaction needs and contextual behavior (Wiese et al., 2017). Similarly, other authors argue that people treat robots alternatively as things or agents during different moments of an interaction (Alač, 2016).

We previously noted how adopting one or the other framework might not be always a conscious choice and that chances are that people interpret artificial agents' behavior socially as a default option. Therefore, the suggestion of developing design solutions that allow users to switch from one framework to the other should be considered under this assumption of a "primacy of the social mindset". In other words, while in certain cases it might be beneficial for the interaction to support the attribution of intentions and other mental states, other circumstances might require the opposite approach. In general, following Weick's sensemaking theory, it is important that the process of meanings co-construction (i.e., sensemaking) is lifted from the private and implicit sphere to a public and explicit level (Weick, 1995; Weick et al., 2005). This supports the notion of a more active involvement by users and the adoption of alternative semantics that support users' awareness, as we propose. This approach can also offer people a strategy to reduce the risk of wrongly adopting

one framework (e.g., the mentalistic one) instead of the other, which would lead to incorrect predictions when an artificial agent's behavior does not match the adopted mental model (Wiese et al., 2017). The next sections consider under what circumstances one or the other framework might be the most appropriate.

However, such considerations on ideal rationality do not necessarily imply that every time an agent's behavior leads to an unpredictable outcome from an intentional perspective, this is necessarily the result of a system error or malfunction. It might well be that the user simply cannot make sense of certain actions because she cannot immediately grasp the reasons behind them. It is in such cases that the user typically asks for an explanation (Miller, 2019). This can still be provided by the agent within an intentional framework, thus highlighting an "information asymmetry" according to which the agent's decision-making process was simply not obvious (Malle et al., 2007). Alternatively, the explanation could clarify whether an internal failure has occurred, thus letting users know that a mechanistic model would be more appropriate. In conclusion, this transition to non-intentional frameworks is likely to proceed more smoothly if users are made aware of the necessity to switch. We shall return and provide further support to this assumption in the last section where we discuss ethical implications.

2.4 AlphaGo: A Case Study

Here, we briefly discuss how the elements discussed so far are not only a matter of theoretical debate or experimental testing, but also apply to real life. To do so, we analyze a few aspects related to the case of Deep Mind's AlphaGo. One of the main reasons to reflect upon this case is that it offers an 'updated' direct comparison with Dennett's original example of the chess-playing computer. However, we can expect the future to provide further examples as this type of technology becomes more broadly present in our society.

Go is a very old board game and among the most complex ones, where 'human intuition' plays a fundamental role. Deep Mind's Go-playing system is not preprogrammed by expert players to perform a set of specific moves. Rather it is trained (or trains itself) through reinforcement learning. Through mimicking human strategies first, and then playing against different versions of itself (Silver et al., 2017), the system is able to improve and adapt its strategies autonomously. When challenged by some of the best human players, AlphaGo has repeatedly proved its efficacy in the game (Andras et al., 2018; Curran et al., 2019).

Curran and colleagues conducted a content analysis of how the Chinese and American press approached AlphaGo's games. Beside the predictable cultural differences, they also highlight how it is not unusual to attribute qualities such as 'intuition' and 'creativity' to the system (Curran et al., 2019). Furthermore, they argue, if such qualities "are no longer the sole domain of humans, there is a demand for a reconceptualization first and foremost of what it means to be human" (Curran et al., 2019, p. 733). In other words, they note, observing traits typical of human intelligence in a machine (whose nature is always transparent)

might even lead to an ontological reconsideration of what it means to be human (Severson & Carlson, 2010; Kahn, et al., 2011).

Another interesting aspect is the fact, that "some moves are made that are novel and inexplicable to human Go-playing experts and yet are effective, leading to more wins and new insights into the game" (Andras et al., 2018, p. 79). Few things can be noted. Heider differentiates intentional actions from unintentional events by saying that the former exhibit 'equifinality' (Heider, 1983). While AlphaGo can employ new and unpredictable moves, the apparent intention to win the game remains the same, i.e., oriented towards the same goal, i.e., 'equa-final' (Heider, 1983).

A second remark emerges that concerns and further explains the previously discussed distinction between perceiving and attributing intentions, particularly with those AlphaGo's moves that are inexplicable and yet effective (especially move 37 of the second match against Lee Sedol (Metz, 2016)). The reason why certain moves were difficult to predict is that human players would have hardly ever used them in those circumstances. The commentators of the game even wondered whether move 37 was a mistake. In pragmatic terms, what AlphaGo did was to opt for a very uncommon (among human players) strategy, whose outcome was a almost certain victory, although with a very small margin. Beyond uncovering new possible approaches to the game, the point we aim to make here concerns the fact that a move initially perceived as possibly erroneous turned out to be a winning one. In other words, the audience attributed the 'equifinality' of winning the game only in hindsight, while the initial perception was unclear. Furthermore, recalling the previous considerations on systems' ideal rationality, part of the audience was prone to attributing move 37 to a system mistake, highlighting the persistence of mechanistic interpretations of the system's behavior. However, such a possibility was later discarded as the move proved to be successful, although initially hard to predict and explain.

Finally, it can be noted that it would probably be very difficult for laypeople to obtain new insights into the game and learn new gaming strategies from a mathematical (i.e., design) perspective (for a similar analysis see Ling et al. (2019)). It might, however, be possible to interpret (or predict) those moves as if they had been made by a human whose goal is to win. In this regard, Curran and colleagues report a professional player commenting on one such move by saying that "almost no human would've thought of it" (Curran et al., 2019, p. 733).

In conclusion, Curran and colleagues observe how the type of narrative used by the media was influenced by the journalists' lack of domain-specific expertise, which allegedly led them "to relay on broad and undifferentiated frames" (Curran et al., 2019, p. 734). However, as we argued previously and partially in line with Dennett, a more plausible explanation is that the complexity of systems like AplhaGo does not leave laypeople (included most journalists and professional players) much room for interpreting the moves as the result of programming strategies. Rather, attributing to AlphaGo the intention (and desire) to win the match, the rationality needed to achieve this goal and the belief that a specific strategy would have been successful as it would be done with human players, is the only viable strategy for non-experts to understand, predict and perhaps learn from the system's behavior.

3 Tracing the Point of Origin of Intention Ascription in Artificial Agents: A Comparative Analysis of HCI and HRI

Dennett's original formulation of the intentional stance hinges on the complexity of sophisticated systems and the apparent rationality of their behavior as the main trigger for the ascription of intentions. Today's AI-based technologies are far more complex and advanced compared to those described by Dennett. Therefore, the issue of whether or not to treat today's machines as intentional agents is extremely pressing and relevant for their successful introduction into our society. How can complex rational behavior be unpacked and articulated to come up with implementable strategies? And how is this issue interpreted and studied in the empirical literature?

We approach this issue considering that attributing intentionality and other mental states to artificial agents is a flexible process (Abu-Akel et al., 2020). Furthermore, we acknowledge the intrinsic nuances of the process, which can very sensibly according to the already existing variety of artificial agents. According to Thellman and colleagues, how the attribution of intentions and other mental states varies depending on the type of agent represents an open challenge (Thellman et al., 2017). To this extent, they note, "there has been very little comparative research on how people actually interpret the behavior of different types of artificial agents" (Thellman et al., 2017, p. 1). Therefore, in this section, we analyze relevant examples from the experimental literature on virtual and embodied agents to investigate these assumptions from a comparative perspective. Based on this analysis, we argue that only by adopting a transversal approach does it become possible to grasp the nuances of this flexibility.

However, it is important to acknowledge that situations in real life might soon become even more nuanced, especially as the number of typologies and the diversification of artificial agents to interact with increase. Whereas we circumscribe the analysis to virtual and embodied agents, variants of each category already exist (e.g., anthropomorphic and machine-looking robots) that are worth examining individually and in comparison with other forms of social presence (Cassell, 2000; Bartneck, 2003; Kiesler et al., 2008; Li, 2015). Furthermore, there is in the literature a lack of long-term studies, which would help building a better understanding of how processes such as the attribution of intentions and other mental states evolve over time.

An initial distinction that emerges from the comparative analysis highlights how the phenomenon depends on intrinsic features of the agents, people's dispositions and external and contextual conditions. Furthermore, how the combination of these factors influences the overall process is rarely taken into consideration (Marchesi et al., 2019; Schellen & Wykowska, 2019).

3.1 Contextual Conditions

Several contextual elements that contribute to triggering the attribution of intentions can be identified. For instance, as illustrated by the example of AlphaGo, a society's cultural background can have repercussions for its perception of artificial agents

(Haring et al., 2014; Curran et al., 2019). Even more, attribution of anthropomorphic traits appears to be influenced by whether people perceive artificial agents to be members of the same in-group (e.g., in terms of nationality or gender) rather than of out-groups (Eyssel & Kuchenbrandt, 2012; Eyssel et al., 2012; Kuchenbrandt et al., 2013). Another relevant avenue of research investigates how attribution of anthropomorphic traits to physically present artificial agents can influence people's success in carrying out social and cognitive tasks (Riether et al., 2012; Spatola et al., 2019, 2019).

However, one element that occupies a central position is the type of tasks involved in the interaction (Epley et al., 2007; Marchesi et al., 2019). The fact that many artificial agents are meant to be employed in social contexts makes this aspect particularly relevant. For instance, Chaminade et al. (2012) conducted an fMRI study involving a competitive scenario (rock-paper-scissors) to compare attitudes towards the competitors—a human, an 'intelligent' robot and a 'random agent' (that did not base its moves on any strategy). Their results show that while participants treated the human competitor as being intentional, their reactions towards the robot were not significant in terms of intention attribution (Chaminade et al., 2012). As a possible explanation for this, the authors point to participants' lack of a clear cognitive strategy to interact with the agent, which resulted in them relying mostly (or exclusively) on individual opinions about the robot's inner mechanisms (Chaminade et al., 2012).

However, a different explanation is provided by Thellman et al. (2017), who suggest that the simple experimental scenario was the reason why no significant attribution of intentions was detected. While it is true that individual expectations do indeed play a central role (as discussed later), considering that a game like rock-paper-scissors does not involve much strategy (unlike other games, such as chess or Go), it becomes clear that the same type of agent might be treated as being either intentional or mechanical depending on the interaction affordances. This interpretation further refines the idea of people adopting a default social mindset when interacting with artificial agents. Specifically, the last consideration suggests that people tend to adopt a mentalistic approach as a default option when other cognitive processes are involved (e.g., strategic thinking and social cognition) (Spunt et al., 2015). However, the different interpretations of the results obtained by Chaminade et al. (2012) highlight another contextual factor we ought to consider.

Researchers' attitudes when investigating this phenomenon (or any phenomenon) might play a part in influencing how participants perceive an agent. This aspect seems to be largely underestimated in the literature. Perhaps this is because researchers' attitudes are not believed to have a direct impact on real social interactions. However, the way a researcher approaches a topic surely influences what can be found (i.e., when a researcher studies a phenomenon, the divergences and biases introduced by his or her unique point of observation may go unnoticed). Consequently, as researchers are among the people in charge of designing artificial agents, their approach findings can influence the societal perception of a specific topic in an indirect way, particularly in the longer term.

Another example stems from the analysis by Lim and Reeves (2010). The authors discuss levels of engagement in gaming experiences when playing with or against

'avatars' and 'agents'. Based on several studies, they state that when people believe they are interacting with a digital avatar of a real person, perceived social presence is higher compared to when interacting with an artificial agent. While in principle this might be a sound assumption, the authors hypothesize that a negative attitude towards the agents arises because players cannot 'mentalize' their opponent when this is an agent (rather than an avatar) (Lim & Reeves, 2010). In particular, their assumptions rest on a description of agents rooted in their (lack of) biological intentionality (Lim & Reeves, 2010). However, as we discussed previously, attributing intentions to an agent does not necessarily imply biological forms of intentionality. In conclusion, whereas relying on biological definitions of intentionality might explain negative dispositions towards agents, if engagement and ascription of intentions are detected and measured, this implies that people can and do mentalize artificial agents. Hence, the explanation provided by Lim and Reeves (2010) does not hold and, to the contrary, shows an underlying bias in addressing the topic.

3.2 Human Attitudes

We have previously noted how, alongside objective biological interpretations, intentionality (and the ability to infer and ascribe intentions to others' behavior) can also be read as a socio-cognitive construct (that makes social interactions possible). The intertwining of these two processual levels starts at very early stages of life. In fact, "when infants follow others in adopting the intentional stance, they acquire better interpretational resources, which increases their incorporation into the adult environment, and this, in turn, furthers the process of enculturalization." (Perez-Osorio & Wykowska, 2019). Furthermore, the ability to mentalize others might be part of a cerebral network labeled "the brain's deafult network", which has been shown to activate when one tries to mentally anticipate and explore social scenarios (Buckner et al., 2008). Consequently, people are trained to mentalize and recognize intentional patterns (Frith & Frith 1999, 2006; Chaminade et al., 2012; Perez-Osorio & Wykowska, 2019) and, more generally, to attribute anthropomorphic traits to non-human entitites (Caporael & Heyes, 1997; Nass & Moon, 2000; Nass et al., 1994), meaning that these strategies are widely available if necessary, according to the interaction affordances. We shall now discuss what it means for a strategy to be available if necessary.

Importantly, it can be noted that for people, it still makes a difference whether they interact with conspecifics or with artificial agents. In other words, brain activation is stronger in human–human interactions. However, reported differences can vary greatly from case to case (Thellman et al., 2017; Marchesi et al., 2019; Perez-Osorio & Wykowska, 2019). To this extent, an interesting perspective is provided by Bossi and colleagues, who conducted a study analyzing how brain activity in the 'resting state', i.e., when not engaged in a task, biases the perception of robots during interaction. They found that if mentalizing processes are present during the resting state, people are more likely to treat robots mechanistically later when interacting with them (Bossi et al., 2020). They explain these counterintuitive results by arguing that "if participants were involved in thinking about other people, and their

intentions or mental states in general, before they took part in the task, the contrast with a robotic agent might have been larger" (Bossi et al., 2020, p. 4). Hence, although the attribution of intentions is a strategy that is always available, its adoption (or lack thereof) might be affected by the preceding neural activity, showing a non-linear correlation with other variables such as the type of activity or the general disposition of individuals towards artificial agents.

Despite quantitative differences, there seems to be a certain degree of agreement on the possible cause for this differential activation of mentalistic schemata. The idea is that it is fairly easy for people to interpret certain artifacts as material objects and humans as intentional agents. Everything in between lacks a specific ontological categorization, forcing people to adopt a familiar framework, which often turns out to be the intentional one (Davidson, 1999; Thellman et al., 2017; Marchesi et al., 2019; Abu-Akel et al., 2020). Consequently, as previously argued, people adopt this strategy when it proves to be the most efficient or reliable (Perez-Osorio & Wykowska, 2019). To this extent, according to Weick, if previously adopted sense-making strategies have been successful, they will be retained and reenacted in future interactions (Weick, 1995). However, this is likely to not always be the most fruitful approach, but only in cases where tasks require social cognition (Epley et al., 2007; Spunt et al., 2015; Wiese et al., 2018; Ohmoto et al., 2018; Schellen & Wykowska, 2019). It is in such cases that treating artificial agents as intentional tends to improve the quality of the interaction (Wiese et al., 2017; Schellen & Wykowska, 2019).

Furthermore, in line with Dennett's idea of complexity, it should also be considered that approaching artificial agents from a mechanistic perspective is generally difficult for many people (Dennett, 1981). The reason is that it may be cognitively too demanding, especially for non-expert users, to try to make sense of artificial agents behavior from a mathematics-based, design stance. This highlights a connection between the concept of systems' complexity and the idea of necessity. However, such a relationship should be read in light of the type of tasks and interactions involved, as previously discussed. Before drawing any conclusion, the next paragraph will consider the last set of relevant elements that can influence the attribution of intentions and other mental states. Additionally, it should still be considered that cultural or personal dispositions towards technology might still override the availability of an intentional framework, encouraging people to adopt either a mechanistic or an anthropomorphic approach (Waytz et al., 2010; Haring et al., 2014).

3.3 Intrinsic Features

Analyses of internal or exhibited features of agents that factor into the attribution of intentions (and more broadly of social skills) converge towards two categories of qualities: appearance and behavior (Wiese et al., 2017). This, supported by advances in neurosciences, particularly the availability of fMRI techniques, has led some researchers, especially in the field of human–robot interaction, to emphasize the importance of anthropomorphic embodiment as a trigger for the use of mentalistic descriptions (Marchesi et al., 2019). Of particular interest, in this direction is the discovery of mirror neurons (Rizzolatti & Craighero, 2004). These appear to play a

role in the processes of attributing intentionality based on embodiment, so that similarity to human physical presence triggers higher activation.

Consequently, research on physical appearance has focused on endowing agents (particularly robots) with human-like features. Some features, such as a face (Johnson, 2003; Looser & Wheatley, 2010; Balas & Tonsager, 2014), gazing eyes (Khalid et al., 2016; Willemse et al., 2018), a non-symmetric ratio between the head and the body, smooth bodily transformations as opposed to rigid and linear changes, (Johnson, 2003), and the visibility of the entire body (Chaminade & Cheng, 2009) have been highlighted as preeminent. This approach is rooted in the idea that "humans might be able to understand the behavior of human-like robots more easily than, for example, the behavior of autonomous lawnmowers or automated vehicles" (Thellman et al., 2017, p. 2). This is a plausible explanation, as feature similarity is likely to more effectively and quickly activate the brain areas involved in mentalizing and motor resonance (Chaminade et al., 2007; Wiese et al., 2017).

However, if the attribution of intentions mostly depends of appearance, this would not explain positive results in the absence of a body or with very different forms of embodiment. Interestingly, Ziemke (2020) reports one such case in relation to a road accident involving an autonomous vehicle. The accompanying report by the U.S. National Transportation Safety Board notes how some people were surprised that "Uber's self-driving car didn't know pedestrians could jaywalk" (Ziemke, 2020, p. 1). This is explained as "an expectation-probably shared by many people that driverless cars should have a human-like common sense understanding of human behavior." (Ziemke, 2020, p. 2) More generally, this implies that behavioral elements might be at least as important as appearance, if not more (Terada et al., 2007; Wiese et al., 2017). Among qualities in this category, studies have focused on the reciprocity and contingency of the behavior in relation to the environment and to other agents (Johnson, 2003; Pantelis et al., 2014, 2016). Furthermore, autonomous, rational and seemingly biological motion play a central role (Castelli et al., 2000; Gazzola et al., 2007; Oberman et al., 2007; Pantelis et al., 2016; Abu-Akel et al., 2020).

In this respect, experiments in the tradition of a well-known study by Heider and Simmel make an important contribution. Heider and Simmel argued how people attribute social skills to geometric shapes in motion (Heider & Simmel, 1944). Developing this concept further, Pantelis and colleagues analyzed the relationship between the goal-directed motion of similarly simple, autonomous geometric objects in a two-dimensional virtual environments and the attribution of mental states (Pantelis et al., 2014, 2016). The results of the first study show that people tend to estimate agents' states (e.g., when they are 'attacking' another agent, or 'fleeing' from it) correctly and coherently with one another (Pantelis et al., 2014). Perhaps more revealing are the results of a follow-up study, where an evolutionary factor is introduced into the agents' behavior. In fact, the authors hypothesize that the ascription of mental properties is at least partially related to how artificial agents adapt to their environment (Pantelis et al., 2016). One of their main arguments is that people's ability to correctly infer the agents' states increases concurrently with the rationality of the agents' behavior. Their results show that people tend to infer more accurately the mental states of agents that adapt their behavior. These studies not only

corroborate the relevance of motion for appropriate judgements of the behavioral information that motion itself conveys. Arguably, they support the idea of a primacy of 'pro-social rational content' over the means of conveyance (i.e., in this case, motion itself). Thus, in the absence of such minimal rationality, artificial agents' behavior does not communicate any mental state (Pantelis et al., 2016).

This last proposition is supported by the fact that appearance and biological motion alone (or combined) cannot explain the ascription of intentions and the activation of brain areas responsible for mentalizing in cases where both features are missing. The study conducted by Abu-Akel et al. (2020) is in line with this position. The authors investigate the ascription of intentions to a virtual agent in a competitive scenario, with the participants not able to see their competitor. They hypothesize that the activation of brain areas involved in mentalizing operations does not require motion. Instead, they claim that abstract information about the opponent is sufficient as long as it is considered an intentional and rational agent of either natural or artificial nature. Thus, their results show how "activation of the 'mentalizing network' might be specific to mentalizing, but it is not specific to mentalizing about humans." (Abu-Akel et al., 2020, p. 8). Interestingly, conclude the authors, "such flexibility in the attribution of intentionality (whether to active or passive, human or computer agents) can be manipulated volitionally and even strategically" (Abu-Akel et al., 2020, p. 8). As it will be addressed in the next section, this has have ethical implications.

Another study pointing in a similar direction was conducted by Pinchbeck (2008). Here, the authors analyze how to enhance gaming experiences by implementing simple behavioral tricks in non-player characters, rather than relying on more complex AI techniques to drive more nuanced individual behaviors. Referring to a group of non-player characters (human mercenaries), they describe a 'breakdown of intentionality' as a consequence of the characters' incoherent behavior. Under certain conditions, these characters enter a 'combat state' (i.e., 'seemingly intentional' behavior of actively seeking enemies). Instead, when they are in the water they engage in a sort of 'rest state' ('the pool party effect', in the terminology used in the paper), making themselves vulnerable to attacks. Furthermore, the characters seem to be uninterested in solving this issue, an attitude that the authors identify as totally irrational. This negatively affects the attribution of intentions and rationality to the characters (and the gaming experience). This, the authors note, happens because people tend to grant intentionality when actions are recognized as ecologically valid (Pinchbeck, 2008). By contrast, another kind of non-player character (i.e., mutated monkeys called Trigens) display more ecologically valid behavior by avoiding entering the water, which would cause them to drown. This rational attitude appears to suggest a higher degree of intentionality despite the characters' less human appearance.

In conclusion, as the examples show, our analysis identifies a few concepts that are useful for the design of artificial agents. Implementing features that support the attribution of intentions can be a desirable strategy, as it may enhance the overall quality of the interactions. In this way, manifest behavior that conveys a message of contextual, pro-social rationality serves as the main spark that ignites the processes of attributing mentalistic qualities to artificial agents. This is supported by the fact

that the ascription of intentions and other mental states is a widely available mental process that people are trained to engage in beginning at an early age. Hence, this also clarifies our previous point on necessity. While the ontological classification of most objects is not problematic, as soon as more sophisticated devices' behavior appears as minimally rational and pro-social, the combination of the availability of a mentalistic approach and the likelihood of a cognitive overload that might derive from trying to make sense of such machines from a mechanistic perspective result in the default adoption of mentalistic schemata.

Whenever they can be implemented, features like a human-like appearance or biological motion are fundamental tools to support the process, and as such, they should always be considered as a possible design strategy. Nevertheless, human-like appearance and motion alone are not sufficient conditions (e.g., a highly anthropomorphic robot that does not act in a contextually rational way is likely to be treated as a sophisticated mannequin). They need to be accompanied by (appearance) or convey (motion) some sort of rational message with ecological validity. Accordingly, artificial agents that do not display either of these qualities (appearance or motion) can still be treated as intentional, as is for instance the case with AlphaGo or conversational agents. As the perceivable rationality of agents' behavior increases, the attribution of intentions becomes more likely. Additionally, as the interactions with artificial agents increase in number and variety, the attribution of intentions and other mental states may become part of implicit social cognition processes. Referring to models of the mind proposed within the context of "dual processes" and "dual systems" theories, this would further reduce the cognitive load, and make the process more automatic (Evans & Stanovich, 2013).

However, it is important to note that depending on the interaction context, the tasks to be carried out and the type of machine, a mentalistic approach might not always be the most appropriate. If no social cognition is involved (e.g., as with autonomous vacuum cleaners or lawnmowers), more mechanistic mental models are adequate. Referring to design features that allow users to switch from one interpretative framework to another, it is fundamental that such design strategies consider said transition in both directions. For instance, if a robotic vacuum cleaner crashes or malfunctions, adopting a mentalistic approach is counterproductive. More generally, even when such machines function properly treating them as intentional agents would likely not be very beneficial to the interaction. This further supports the emphasis on social cognition and the idea of seemingly rational behavior as triggers. Similar considerations are to be accounted for in the design phase of artificial agents, also in light of the fact that people tend to attribute human traits to machines even when, in principle, mechanistic approaches would be more appropriate (Carpenter, 2013).

4 Ethical Considerations: Attribution of Intentions and Deception

This section of the paper takes a twofold approach to the ethical aspects of adopting the intentional stance. On the one hand, we acknowledge that, as artificial agents' presence in a growing number of everyday contexts increases, it is important that

interactions with them become progressively more efficient, pleasant and trustworthy. Accordingly, design strategies that support users' adoption of the intentional stance in contexts that involve social cognition might be sought after (Spunt et al., 2015; Schellen & Wykowska, 2019), particularly with respect to mutual connections, joint human–agent efforts, and, more generally, social acceptance of artificial agents (Wiese et al., 2017). On the other hand, these efforts aimed at improving the overall quality of human–agent interactions should not translate into design strategies that make the categorization of artificial agents ambiguous (Hackel et al., 2014; Mandell et al., 2017). In fact, extreme anthropomorphic attributions might have a negative impact on the quality of the interaction (Mandell et al., 2017; Ziemke, 2020) if, for instance, people start perceiving artificial agents as a threat rather than valuable resources (Spatola & Normand, 2020). It is therefore important to find a proper balance between these two necessities in advance, so as to not leave the burden of evaluation entirely on the people interacting with the agents.

4.1 Layers of Deception in Human–Agent Interaction

Before analyzing whether the implementation of features that resemble intentional behavior should be labeled a deceptive design strategy, it is first necessary to briefly consider what deception means in the first place. According to Danaher, at the lowest level of analytical granularity, "deception involves the use of signals or representations to convey a misleading or false impression. Usually the deception serves some purpose other than truth, one that is typically to the advantage of the deceiver." (Danaher, 2020, p. 118) According to this interpretation, deception centers around three main elements: the person being deceived, the agent directly responsible for perpetrating the deception, and the signal or misleading information. Another layer must be considered. It is represented by the interests of what we call a 'third party', which typically is the entity or set of entities (e.g., companies, designers, malicious users etc.) that act from 'behind the curtain' to provide the conditions necessary for the agent to perform deceptive acts. These are often the actors that ultimately gain the greatest advantage, for instance, in terms of data use (Kaminski et al., 2016; Hartzog, 2016) or for unethical commercial or even criminal purposes (O'Leary, 2019). It is important to acknowledge this aspect within the context of human–agent interaction, for reasons that are mostly related to responsibility distribution, as it will be addressed further on.

The issues with attributing intentions to artificial agents and the implementation of strategies meant to trigger such attributions are at the heart of the debate on deception. The first consideration in this regard is quite nuanced. As Danaher notes, it has to do with the fact that what exactly constitutes a deceptive act among humans is defined by the intentions, desires and beliefs of the deceiver. Arguably, taking such a perspective in human–agent interaction is problematic, since whether agents have intentions and other mental states or not is itself part of the debate on deception (Danaher, 2020). Most of the debate concentrates on interpretations of intentionality that are in line with or similar to Searle's. Often, this type of criticism is also directed at features that express emotional engagement (such as care

or love). Therefore, anthropomorphic cues that do not reflect actual qualities (e.g., mental states) are fundamentally seen as deceptive. Some authors argue that anthropomorphic behavior and 'simulated qualities' are designed to trick and fool people precisely because they let people believe robots (and other agents) have those qualities that they lack (Sparrow & Sparrow, 2006; Sharkey & Sharkey, 2010; Turkle, 2010; Elder, 2016). More broadly, it is argued that the implementation of features that express seemingly intentional behavior could trigger categorical uncertainty (in ontological terms) and therefore undermine social interaction (Hackel et al., 2014; Mandell et al., 2017).

Not only researchers but also institutions that oversee the development of AI and robotics have highlighted the potentially negative aspects of excessive anthropomorphization/personification of artificial agents. The EU High Level Group's call for trustworthy AI is one such example, but several other bodies have moved in a similar direction (Coeckelbergh, 2019; Floridi, 2019; HLEG, 2021). For instance, The UK Engineering and Physical Sciences Research Council's (EPSRC) Principles of Robotics clarifies that robots should not be designed to deceive users and should always be clear and transparent about their artificial nature (Theodorou et al., 2016; Boden et al., 2017). Similar efforts highlight that robots and other artificial agents should not pose as humans (Shahriari & Shahriari, 2017; Heaven, 2018).

However, it is important to note that not all researchers agree with these positions. In fact, some consider at least some forms of deception to be an intrinsic feature of robotics and AI, as they offer the best possibility of successfully developing socially integrated artificial agents. As such, deception is seen as an acceptable, even desirable phenomenon (Wagner & Arkin, 2011; Shim & Arkin, 2012; Isaac & Bridewell, 2017; de Oliveira et al., 2020). Indeed, one might argue that deception even lies at the foundation of the Turing test, the many versions of which share the assumption that, in order to pass the test, a machine must succeed in convincing a human jury that they are actually interacting with another human.

Danaher highlights a further possible distinction. He regards what he calls 'hidden state deception' as the most dangerous layer. This form of deception occurs when agents hide capacities they possess by means of deceptive signals (Danaher, 2020). Collecting personal data without users knowing it or, even worse, pretending it is not happening falls into this category. While it is reasonable to share such concerns, this paper primarily aims to discuss another level of potential deception, what Danaher calls 'superficial state deception'. This entails that an agent "uses a deceptive signal to suggest that it has some capacity or internal state that it actually lacks." (Danaher, 2020, p. 121) Indeed, implementing features that resemble intentionality is a form of superficial state deception, although the main beneficiaries of the two levels are roughly the same (i.e., the aforementioned third parties).

4.2 Seemingly Intentional Behavior is not (Necessarily) Deception

The thesis we defend here is a twofold one. First, we argue that in principle, it is not unethical to opt for design strategies that support the adoption of the intentional stance. However, to avoid feelings of deception, a fundamental prerequisite is that

users are put in a position to, if not consciously decide, at least be aware of the nature of the agent. We approach this discussion based on the aspects introduced in the first part of this section and in light of the alternative approach previously proposed. Regarding the recursive argument pinpointed by Danaher, we identify two levels of interpretation. If to say that someone is a deceiver implies the intention to deceive, then we argue that this is not the case with artificial agents. By speaking of 'seemingly intentional behavior', we mean to emphasize the conscious attempt to emulate human behavioral traits for the sake of interactional quality. As such, the term specifically seeks to avoid conflation with biological, evolution-based interpretations of intentionality. Artificial agents do not have genuine intentions (e.g., to deceive) in the biological sense, which implies that they cannot be genuinely deceptive.

However, we also introduced the idea of a possible involvement of third parties. Artificial agents can, in principle, stand for the interests of said actors. Accordingly, in addition to what was previously reported, Jacob states that if "a speaker utters words from some natural language or draws pictures or symbols from a formal language for the purpose of conveying to others the contents of her mental states, these artifacts used by a speaker too have contents or intentionality." (Jacob, 2019, p. 1) We consider it problematic if these actors seek to deceive users through the agents (i.e., the artifacts that convey the actor's intention). As such, artificial agents could well be 'tools of deception' by these third parties.

For this reason, we deem it fundamental to make an ethical distinction between the promotion of illusions such as personification and the implementation of features that trigger the ascription of intentions. Recalling that intentionality is not only an objective quality, but also a social construct that makes interactions possible and even increases the overall quality of the relationships, we cannot consider the implementation of features that aim to resemble intentionality in itself as ethically problematic. Accordingly, the attribution of intentions should be seen as a strategy that people can adopt to better predict and interpret agents' behavior and to navigate social interactions with them. From this perspective, design strategies that facilitate this process should be regarded as worth striving for. Furthermore, one may even argue that if users feel more at ease treating agents as intentional in certain cases (as the experimental literature shows), telling them that this feeling is part of a deception could negatively affect their social interactions with the agents. In other words, similarly to what we argue about terms such as 'simulated intentionality', 'ethically exclusivist' positions could dissuade users from engaging in meaningful interactions with the agents if the only interpretative and predictive framework they can use or that seems to work is the intentional one. However, whether artificial agents should employ emotional terminology (such as 'I care'), for instance, is open to debate. In fact, such solutions may lead to the perception of ontological ambiguity (Hackel et al., 2014; Mandell et al., 2017).

4.3 Users' Awareness

The other point we highlight is the concerning possibility of artificial agents acting deceptively on behalf of third parties. As previously mentioned, we consider it

among the most pressing aspects of the present debate. Therefore, here we address the idea of 'promoting the illusion of personification'. In accordance with regulatory institutions that have called for greater transparency, we argue that users' awareness should be a *conditio sine qua non* for the design phase. In other words, for the implementation of specific features to be ethical and successful in fostering social engagement, people should be made aware of the nature of the agents they interact with. We also consider this a prerequisite for people to be given the possibility to switch to mechanistic approaches when necessary. Furthermore, we claim that building such awareness is the specific responsibility of the third parties in charge of the design of artificial agents. Coeckelbergh (2018) takes a relevant position with respect to this issue stating that although deception can be seen as a co-created performance, designers and other third parties have the responsibility to ultimately reveal the 'trick' behind it. In other words, they are responsible for the performative affordances of the agents they introduce into society (Coeckelbergh, 2018).

In accordance with Coeckelbergh, we consider performance as a co-construction (Coeckelbergh, 2018). An agent behaves in a seemingly intentional manner; a person then attributes intentionality to the performed act, and together, the two co-construct the interaction. However, we further claim that the tricks performed by an artificial agent should be revealed beforehand. It makes it possible to still be meaningfully engaged without necessarily thinking that the agent has intentions (in the biological sense) or, even worse, being surprised that one is interacting with an artificial agent. This last point is fundamental for a simple reason. The more sophisticated artificial agents become, the more difficult it will be for people to tell the difference (between a human and a machine) in advance. In order to shed greater light on our position, we now provide an example concerning the well-known topic of the 'uncanny valley'.

4.3.1 The 'Uncanny Valley' Case

The 'uncanny valley' hypothesis posits that extreme anthropomorphism can trigger negative reactions in people (Mori et al., 2012). In this view, the 'valley' of the curve represents the point where human-like appearance and behavior do not quite reach total resemblance, but are still enough to trigger rejection mechanisms. Many hypotheses have been proposed to explain the phenomenon. One argument in line with our position centers around the idea of rejection as the expression of violated expectations (Saygin et al., 2012; Urgen et al., 2018). Urgen and colleagues conducted a study with a highly anthropomorphic android. Whereas the android was human-like enough to generate high initial expectations through its appearance alone (e.g., by looking at a picture or a video of it), as soon as the android began to move its artificial nature became evident, triggering feelings of rejection. This connection between movement and eeriness or uncanniness was previously raised by Mori et al. (2012), but not investigated thoroughly.

Urgen and colleagues argue that, although participants generate an initial mental model (and expectations) based on appearance alone, when the robot's movements reveal its true nature, the established model fails to hold, contributing (in this case) to an increase in uncanny feelings (Urgen et al., 2018). A similar mismatch between

appearance and motion is reported in Saygin et al. (2012). Also in this case, participants were exposed to videos of different types of agents, while their neural activity was monitored. Respectively, the agents were a robot with mechanical appearance, an android and a human. They conclude that while "the android used in our study is often mistaken for a human at first sight, longer exposure and dynamic viewing has been linked to the uncanny valley" in reason of the participants' prediction error that the mismatch generates (Saygin et al., 2012, p. 420). Interestingly, while in Saygin et al. (2012), Urgen et al. (2018), participants are shown only videos of the robot, in other studies with similar androids people interact directly with them (Bartneck et al., 2009). Importantly, in this last case people are always aware of the fact that, regardless of how extremely anthropomorphic, the agent is artificial. In the previously cited cases, they are not. They only become aware of this when the robot moves.

The different records in terms of feelings of uncanniness reported by the studies can be explained by the abrupt failure of expectations (reported in the first two studies). When the conditions for behaving 'as if' are not made explicit, people (in the considered case) are likely to simply behave as they would with other humans. But when the illusion is broken, so are the mental models, generating feelings of rejection.

4.4 'Third Parties" Responsibility

These considerations support our claim that users need to be made aware of the nature of the agent and that this awareness contributes to the quality of the interaction. Thus, our final point for reflection concerns our last claim, that responsibility for ensuring users' awareness of the nature of an artificial agent should fall on the 'third parties' in charge of designing what kind of performance the agent is capable of providing. This issue, notes Coeckelbergh, is seldomly considered in robotics, because "the designer (and especially the company) needs to sell the device as magic" (Coeckelbergh, 2018, p. 80). Fundamentally, in order to have the chance to switch to behaving 'as if', people must be aware of the type of performance that has been created. At the same time, treating agents 'as if' does not mean that there is nothing real to be gained from interacting with them. The thoughts, impressions and feelings one experiences are real, rather than 'simulated' (Turkle, 2010; Seibt, 2017).

Furthermore, not only should ensuring such awareness be third parties' responsibility but, as we noted, the underlying trick should be revealed before the interaction takes place. In fact, as artificial agents become progressively more sophisticated, the decision to automatically behave 'as if' might become less obvious. This is particularly the case with virtual technologies like conversational agents and chatbots. In fact, in most cases it is still possible and fairly easy to determine the artificial nature of a physical robot. No matter how well crafted modern androids might be, it is particularly difficult to flawlessly replicate an extremely human-like appearance, the smoothness of human movements, non-verbal cues, etc., so that, as in the considered example, only through an indirect medium (i.e., pictures or videos) and in

absence of movement could said androids be mistaken for humans. The same cannot be taken for granted when the interaction occurs in a fully virtual environment, as the next examples show.

4.4.1 Google Duplex and GPT-3

Google Duplex is a conversational agent endowed with natural language features to handle tasks such as making reservations and appointments (O'Leary, 2019). The most relevant aspect here is that the system does not only engage in natural language conversations. It also incorporates what (O'Leary, 2019) calls 'speech disfluencies', conversational elements that break the flawless pace of a conversation (such as 'hmm's' and 'uh's'). These kinds of 'interruptions' are very common in human–human interactions, because people do it often as they try to gather their thoughts (Leviathan & Matias, 2018). Nevertheless, including such disfluencies has attracted criticism, precisely because such behavior can be interpreted as deceptive. Consequently, Google's design choices have generated ethical concerns (Lomas, 2018). However, as we have already argued, conducting conversations by employing fluent natural language capabilities or even displaying sophisticated 'speech disfluencies' are not themselves the problem. As previously noted, such features could be worth striving for, as they can improve the overall quality of interactions. However, what is being debated here is the lack of a specific form of transparency that supports users' awareness of the kind of agent they are interacting with. Therefore, having the agent (Duplex, in this case) identify itself at the beginning of an interaction seems to be a reasonable solution (Bay, 2018; O'Leary, 2019). It should then be up to users to decide whether they want to continue under the specified conditions, i.e., once they have been put in the condition to behave 'as if'.

Another technology holding similar potential is GPT-3 (Generative Pre-trained Transformer 3) (Damassino & Novelli, 2020). This deep learning-based natural language processing model can generate text that is often indistinguishable from something a human would write. In their commentary, Floridi and Chiratti show the possibilities and limits of this tool (Floridi & Chiriatti, 2020). They note that in many cases, people might not recognize or even care whether a piece of text has been written by an artificial agent. While this might certainly be true, at least in the very near future, we believe that regulations should not only apply to ideal cases, but also to extraordinary ones. In other words, regulations requiring such artificial agents to make their nature clear in advance are likely to be necessary, especially in cases that could create ambiguity. Consequently, we agree with (Floridi & Chiriatti, 2020)'s conclusions that people should be able (i.e., put in the conditions) to discern what is what. One early example of the dual nature of this point is the use of GPT-3-powered bots on Reddit, a popular online platform. In fact, one of these bots was active under a normal username with almost no one noticing it. While most of its comments were reportedly unharmful, the bot also engaged in conversations about sensitive topics, such as suicide (Heaven, 2020). The bot's real nature was discovered when its text outputs were compared to those of the so-called 'Philosopher AI', another GPT-3 based bot (Heaven, 2020). The main difference between the two

is that in the case of the 'Philosopher AI' the artificial nature of the bot has been made clear from the beginning, allowing users to engage in entertaining question and answer sessions with the bot (including about the nature and the coding of the bot itself).

As a closing remark, a relevant concept for our last claim is represented by the process of dehumanization, as opposed to that of anthropomorphization. Dehumanizing is typically intended as failing to attribute human-like traits to other humans (Haslam, 2006; Epley et al., 2007; Haslam & Loughnan, 2014). Hence, in human–human interaction it often comes with negative connotations. However, in the context analyzed here (i.e., human–agent interaction), the idea of dehumanizing agents has positive consequences, at least as long as it is meant to counter the default anthropomorphizing trend. To this extent, Haslam and Bain note how a concrete (rather than abstract) mental 'construal' of other people could help reducing such a dehumanization process (Haslam & Bain, 2007; Haslam & Loughnan, 2014). Additionally, we previously noted how the lack of an agent-specific ontological categorization is among the main causes that trigger the attribution of anthropomorphic traits. Therefore, we assume that, in specific circumstances, supporting a more concrete mental 'construal' of artificial agents will let people engage in a specular process of dehumanizing artificial agents. To this extent, we claim that specific design features, such as endowing the agents with machine-like traits, having them identifying themselves as artificial entities, or pointing out the 'mechanical' nature of malfunctions, would support this process of dehumanization. In turn, this will help the switch from a mentalistic approach to a mechanistic one, when the latter is more pragmatically or ethically appropriate.

Hence, whereas some researchers' concerns about what is unethical for artificial agents may be too extreme, this last reflection leads us to agree with regulatory attempts to require companies and other third parties to adopt an approach that makes people aware of the type of performances displayed and the artificial nature of the performer. In principle, this paper calls for pre-performance forms of transparency, i.e., the nature of the agents' 'tricks' should be revealed before they are performed. This is at least partially in line with those researchers and institutions that promote higher transparency and, consequently, a distribution of responsibility that calls for explicit commitments by third parties. However, the important role played by machine seemingly intentional behavior in enhancing the quality of social interactions must be acknowledged. Therefore, we conclude that, as long as users are made aware of the nature of the agent they are interacting with, the implementation of strategies that support the attribution of intentions to those artificial agents meant to be employed in contexts that involve social cognition and skills should be considered not only ethically acceptable, but also ethically desirable. On the other hand, mentalistic frameworks appear to be the default approach that people resort to while interacting with seemingly rational artificial agents that do not clearly fall into objectual ontological categorizations. When no social cognition is involved, an opposite dehumanizing approach is more adequate. This could be pursued by, for instance, emphasizing the artificial nature of the agents and their machine-like traits.

5 Conclusions and Limitations

The main aim of this paper is to discuss why the attribution of intentions is effective and desirable and to identify corresponding design suggestions. To do so, we propose an analysis in three main directions, corresponding to the three main sections of the paper. First, we discussed semantic implications of the concept, in light of definitions of intentionality and of some objections directed at Dennett's idea, with particular attention to the behaviorism that informs it. We emphasized how the notion that the intentional stance is a strategy to understand and predict the behavior of sophisticated artificial agents represents the most useful aspect of Dennett's formulation. This led us to suggest the adoption of alternative terminology, in order to reduce the risk of conflation between the attribution of intentions to artificial agents and biological approaches to intentionality.

Furthermore, we traced the point of origin of the process of intention attribution by examining experimental literature about robots and virtual agents. Our conclusion is that contextually valid rationality represents the most important feature in order for agents to be treated as intentional. However, we also identified how this can and should be supported by other features and contextual conditions. Additionally, we considered how a mentalistic approach is not the most appropriate when no social cognition is involved and suggested possible strategies to counter excessive anthropomorphization accordingly.

Finally, we discussed possible ethical implications of the attribution of intentionality to artificial agents. While acknowledging the possible benefits of an intentional framework for social engagement with agents, we also identified a prerequisite for the ethical acceptability of such a framework. In line with most regulatory institutions, we argue that is necessary for users to be made aware of the agents' artificial nature and provide examples to support our claim.

One question that we leave open concerns to what extent the actual implementation of features that trigger the ascription of intentions and other mental states should be pushed. Referring to the case of Google Duplex, we said that the speech disfluencies employed by the system are not themselves the problem. Is the same true for the use of 'more sensitive' and openly mentalistic terms like 'I understand', 'I think', or their emotional counterparts like 'care' 'love', etc.? The risk we identify in this case is an extreme and perhaps 'deceptive' form of anthropomorphism. Where should the line be drawn? We question the necessity to employ such emotionally and semantically rich terminology for the ascription of intentions and other mental states to be successful. Perhaps, a critical analysis of such issues in light of other, related concepts, such as that of a "phenomenal stance" will help shed greater light on the debate.

Funding Open access funding provided by TU Wien (TUW).

References

Abu-Akel, A. M., Apperly, I. A., Wood, S. J., & Hansen, P. C. (2020). Re-imaging the intentional stance. *Proceedings of the Royal Society B, 287*(1925), 20200244.

Alač, M. (2016). Social robots: Things or agents? *AI & Society, 31*(4), 519–535.

Andras, P., Esterle, L., Guckert, M., Han, T. A., Lewis, P. R., & Milanovic, K. (2018). Trusting intelligent machines: Deepening trust within socio-technical systems. *IEEE Technology and Society Magazine, 37*(4), 76–83.

Balas, B., & Tonsager, C. (2014). Face animacy is not all in the eyes: Evidence from contrast chimeras. *Perception, 43*(5), 355–367.

Bartneck, C. (2003). Interacting with an embodied emotional character. In *Proceedings of the 2003 International Conference on Designing Pleasurable Products and Interfaces* (pp. 55–60).

Bartneck, C., Kanda, T., Ishiguro, H. & Hagita, N. (2009). My robotic doppelgänger-A critical look at the uncanny valley. In *RO-MAN 2009-The 18th IEEE international symposium on robot and human interactive communication* (pp. 269–276). tex.organization: IEEE.

Bay, M. (2018). *Am I speaking to a human?*, Retrieved May 10, 2018, from https://slate.com/technology/2018/05/google-duplex-can-make-phone-calls-for-you-but-it-should-have-to-identify-itself (tex.jorunal:slate).

Block, N. (1981). Psychologism and behaviorism. *The Philosophical Review, 90*(1), 5–43.

Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., & Kember, S. (2017). Principles of robotics: Regulating robots in the real world. *Connection Science, 29*(2), 124–129.

Bossi, F., Willemse, C., Cavazza, J., Marchesi, S., Murino, V., & Wykowska, A. (2020). The human brain reveals resting state activity patterns that are predictive of biases in attitudes toward robots. *Science Robotics, 5,* 46.

Breazeal, C. L. (2002). *Designing sociable robots*. MIT Press.

Breazeal, C., & Scassellati, B. (1999). How to build robots that make friends and influence people. In *Proceedings 1999 IEEE/RSJ international conference on intelligent robots and systems. Human and environment friendly robots with high intelligence and emotional quotients (cat. No. 99CH36289)* (Vo.l 2, pp. 858–863). (tex.organization: IEEE).

Buckner, R., Andrews-Hanna, J., & Schacter, D. (2008). The brain's default network: Anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences, 1124,* 1–38.

Caporael, L. R. (1986). Anthropomorphism and mechanomorphism: Two faces of the human machine. *Computers in Human Behavior, 2*(3), 215–234.

Caporael, L. R., & Heyes, C. M. (1997). Why anthropomorphize? Folk psychology and other stories. Anthropomorphism, anecdotes, and animals, 59. State University of New York Press

Carpenter, J. (2013). The Quiet Professional: An investigation of US military Explosive Ordnance Disposal personnel interactions with everyday field robots (Unpublished doctoral dissertation).

Cassell, J. (2000). Embodied conversational interface agents. *Communications of the ACM, 43*(4), 70–78.

Castelli, F., Happé, F., Frith, U., & Frith, C. (2000). Movement and mind: A functional imaging study of perception and interpretation of complex intentional movement patterns. *Neuroimage, 12*(3), 314–325.

Chaminade, T., & Cheng, G. (2009). Social cognitive neuroscience and humanoid robotics. *Journal of Physiology-Paris, 103*(3–5), 286–295.

Chaminade, T., Hodgins, J., & Kawato, M. (2007). Anthropomorphism influences perception of computer-animated characters' actions. *Social Cognitive and Affective Neuroscience, 2*(3), 206–216.

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

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*(2), 71–85.
Coeckelbergh, M. (2019). Artificial intelligence: Some ethical issues and regulatory challenges. *Technology and Regulation,* 31–34.
Curran, N. M., Sun, J., & Hong, J. W. (2019). Anthropomorphizing AlphaGo: A content analysis of the framing of Google DeepMind's AlphaGo in the Chinese and American press. AI & Society, 1–9. Springer
Damassino, N., & Novelli, N. (2020). *Rethinking, reworking and revolutionising the turing test*. Springer.
Danaher, J. (2020). Robot Betrayal: A guide to the ethics of robotic deception. Ethics and Information Technology, 1–12. Springer.
Davidson, D. (1999). The emergence of thought. *Erkenntnis, 51*(1), 511–521.
De Graaf, M. M., & Malle, B. F. (2017). How people explain action (and autonomous intelligent systems should too). In *2017 AAAI Fall Symposium Series*.
de Oliveira, E., Donadoni, L., Boriero, S., & Bonarini, A. (2020). Deceptive actions to improve the attribution of rationality to playing robotic agents. *International Journal of Social Robotics,* 1–15. Springer.
Dennett, D. C. (1971). Intentional systems. *The Journal of Philosophy, 68*(4), 87–106.
Dennett, D. C. (1981). *Brainstorms: Philosophical essays on mind and body*. MIT Press.
Dennett, D. C. (1988). Précis of the intentional stance. *Behavioral and Brain Sciences, 11*(3), 495–505.
Dennett, D. C. (1989). *The intentional stance*. MIT Press.
Dennett, D. C. (1991). Real patterns. *The Journal of Philosophy, 88*(1), 27–51.
Dennett, D. C. (1993). *Consciousness explained*. Penguin.
Dennett, D. C. (1995). The unimagined preposterousness of zombies.
Dennett, D. C. (1997). True, believers: The intentional strategy and why it works. *Mind Design,* 57–79.
Dreyfus, H., Dreyfus, S. E., & Athanasiou, T. (2000). *Mind over machine*. Simon and Schuster.
Elder, A. (2016). False friends and false coinage: A tool for navigating the ethics of sociable robots. *ACM SIGCAS Computers and Society, 45*(3), 248–254.
Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review, 114*(4), 864.
Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8*(3), 223–241.
Eyssel, F., De Ruiter, L., Kuchenbrandt, D., Bobinger, S., & Hegel, F. (2012). 'If you sound like me, you must be more human': On the interplay of robot and user features on human-robot acceptance and anthropomorphism. In *2012 7th ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 125–126). tex.organization: IEEE.
Eyssel, F., & Kuchenbrandt, D. (2012). Social categorization of social robots: Anthropomorphism as a function of robot group membership. *British Journal of Social Psychology, 51*(4), 724–731.
Floridi, L. (2019). Establishing the rules for building trustworthy AI. *Nature Machine Intelligence, 1*(6), 261–262.
Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines,* 1–14.
Frith, C. D., & Frith, U. (1999). Interacting minds: A biological basis. *Science, 286*(5445), 1692–1695.
Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. *Neuron, 50*(4), 531–534.
Gazzola, V., Rizzolatti, G., Wicker, B., & Keysers, C. (2007). The anthropomorphic brain: The mirror neuron system responds to human and robotic actions. *Neuroimage, 35*(4), 1674–1684.
Hackel, L. M., Looser, C. E., & Van Bavel, J. J. (2014). Group membership alters the threshold for mind perception: The role of social identity, collective identification, and intergroup threat. *Journal of Experimental Social Psychology, 52,* 15–23.
Haring, K. S. , Silvera-Tawil, D., Matsumoto, Y., Velonaki, M., & Watanabe, K. (2014). Perception of an android robot in Japan and Australia: A cross-cultural comparison. In *International conference on social robotics* (pp. 166–175). (tex.organization: Springer)
Hartzog, W. (2016). Et tu, Android? Regulating dangerous and dishonest robots. *Journal of Human-Robot Interaction, 5*(3), 70–81.
Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review, 10*(3), 252–264.

Haslam, N., & Bain, P. (2007). Humanizing the self: Moderators of the attribution of lesser humanness to others. *Personality and Social Psychology Bulletin, 33*(1), 57–68.
Haslam, N., & Loughnan, S. (2014). Dehumanization and infrahumanization. *Annual Review of Psychology, 65,* 399–423.
Heaven, W. D. (2018). Robot laws. *New Scientist, 239*(3189), 38–41.
Heaven, W. D. (2020). A GPT-3 bot posted comments on Reddit for a week and no one noticed. MIT Technology Review. Retrieved November 24, 2020, from https://www.technologyreview.com/2020/10/08/1009845/a-gpt-3-bot-posted-comments-on-reddit-for-a-week-and-no-one-noticed/
Heider, F. (1983). *The psychology of interpersonal relations*. Psychology Press.
Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American journal of psychology, 57*(2), 243–259.
HLEG. (2021). Ethics guidelines for trustworthy AI - FUTURIUM - european commission. FUTURIUM - European Commission, . Retrieved from https://ec.europa.eu/futurium/en/ai-alliance-consultation
Isaac, A. M., & Bridewell, W. (2017). Why robots need to deceive (and how). *Robot Ethics, 2,* 157–172.
Jacob, P. (2019). Intentionality. In E. N. Zalta (eds.) *The Stanford Encyclopedia of Philosophy (Winter 2019 ed.). Metaphysics Research Lab, Stanford University*. Retrieved from https://plato.stanford.edu/archives/win2019/entries/intentionality/
Johnson, S. C. (2003). Detecting agents. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 358*(1431), 549–559.
Kahn, P. H., Reichert, A. L., Gary, H. E., Kanda, T., Ishiguro, H., Shen, & S..Gill, B. (2011). The new ontological category hypothesis in human-robot interaction. In*2011 6th*, *ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 159–160). tex.organization: IEEE.
Kaminski, M. E., Rueben, M., Smart, W. D., & Grimm, C. M. (2016). *Averting robot eyes. Md. L. Rev., 76,* 983.
Khalid, S., Deska, J. C., & Hugenberg, K. (2016). eye gaze triggers the ascription of others' minds The eyes are the windows to the mind: Direct eye gaze triggers the ascription of others' minds. *Personality and Social Psychology Bulletin, 42*(12), 1666–1677.
Kiesler, S., Powers, A., Fussell, S. R., & Torrey, C. (2008). Anthropomorphic interactions with a robot and robot-like agent. *Social Cognition, 26*(2), 169–181.
Kuchenbrandt, D., Eyssel, F., Bobinger, S., & Neufeld, M. (2013). When a robot's group membership matters. *International Journal of Social Robotics, 5*(3), 409–417.
Leviathan, Y., & Matias, Y. (2018). Google duplex: An AI system for accomplishing real-world tasks over the phone. Retrieved from https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation
Li, J. (2015). The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies, 77,* 23–37 (**Publisher: Elsevier**).
Lim, S., & Reeves, B. (2010). Computer agents versus avatars: Responses to interactive game characters controlled by a computer or other player. *International Journal of Human-Computer Studies, 68*(1–2), 57–68 (**Publisher: Elsevier**).
Ling, Z., Ma, H., Yang, Y., Qiu, R. C. , Zhu, S. C., & Zhang, Q. (2019). Explaining AlphaGo: Interpreting contextual effects in neural networks. arXiv:1901.02184
Lomas, N. (2018). *Duplex shows Google failing at ethical and creative AI design*. Retrieved May 10, 2018, from https://techcrunch.com/2018/05/10/duplex-shows-google-failing-at-ethical-and-creative-ai-design
Looser, C. E., & Wheatley, T. (2010). The tipping point of animacy: How, when, and where we perceive life in a face. *Psychological Science, 21*(12), 1854–1862.
Malle, B. F. (2011). Attribution theories: How people make sense of behavior. *Theories in Social Psychology, 23,* 72–95.
Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of experimental social psychology, 33*(2), 101–121.
Malle, B. F., Knobe, J. M., & Nelson, S. E. (2007). Actor-observer asymmetries in explanations of behavior: New answers to an old question. *Journal of Personality and Social Psychology, 93*(4), 491.
Mandell, A. R. , Smith, M., & Wiese, E. (2017). Mind perception in humanoid agents has negative effects on cognitive processing. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 61, pp. 1585–1589). Number: 1 tex.organization: SAGE Publications.
Marchesi, S., Ghiglino, D., Ciardo, F., Perez-Osorio, J., Baykara, E., & Wykowska, A. (2019). Do we adopt the intentional stance toward humanoid robots? *Frontiers in Psychology, 10,* 450.

Metz, C. (2016). In two moves, AlphaGo and lee sedol redefined the future. Wired. Retrieved 2016–03-16 from https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future. WIRED.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267,* 1–38.

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine, 19*(2), 98–100.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*(1), 81–103.

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In: Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 72–78).

Oberman, L. M., Pineda, J. A., & Ramachandran, V. S. (2007). The human mirror neuron system: A link between action observation and social skills. *Social Cognitive and Affective Neuroscience, 2*(1), 62–66.

Ohmoto, Y. , Karasaki, J., & Nishida, T. (2018). Inducing and maintaining the intentional stance by showing interactions between multiple agents. In: *Proceedings of the 18th*, *International Conference on Intelligent Virtual Agents* (pp. 203–210).

O'Leary, D. E. (2019). GOOGLE's duplex: Pretending to be human. *Intelligent Systems in Accounting, Finance and Management, 26*(1), 46–53.

Pantelis, P. C., Baker, C. L., Cholewiak, S. A., Sanik, K., Weinstein, A., Wu, C. C., & Feldman, J. (2014). Inferring the intentional states of autonomous virtual agents. *Cognition, 130*(3), 360–379.

Pantelis, P. C., Gerstner, T., Sanik, K., Weinstein, A., Cholewiak, S. A., Kharkwal, G., & Feldman, J. (2016). Agency and rationality: Adopting the intentional stance toward evolved virtual agents. *Decision, 3*(1), 40.

Parkinson, B. (2012). *Social perception and attribution*. Hewstone, M.; Stroebe, W.; Jonas, K. (red.), An Introduction to Social Psychology, 55–90.

Perez-Osorio, J., & Wykowska, A. (2019). Adopting the intentional stance towards humanoid robots. In *Wording robotics* (pp. 119–136). Springer.

Pinchbeck, D. (2008). Trigens can't swim: Intelligence and intentionality in first person game worlds. In: *Proceedings of the*, *Philosophy of Computer Games*, 2008 (pp. 242–260). Potsdam University Press.

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people*. Cambridge University Press.

Riether, N., Hegel, F., Wrede, B., & Horstmann, G. (2012). Social facilitation with social robots? In: *2012 7th*, *ACM/IEEE International Conference on Human-Robot Interaction* (HRI) (pp. 41–47). tex.organization: IEEE.

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Reviews Neuroscience, 27,* 169–192.

Saygin, A. P., Chaminade, T., Ishiguro, H., Driver, J., & Frith, C. (2012). The thing that should not be: Predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Social Cognitive and Affective Neuroscience, 7*(4), 413–422.

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

Searle, J. (1980). Intrinsic intentionality. *Behavioral and Brain Sciences, 3*(3), 450–457.

Seibt, J. (2017). Towards an ontology of simulated social interaction: varieties of the "As If" for robots and humans. In *Sociality and normativity for robots* (pp. 11–39). Springer.

Severson, R. L. & Carlson, S. M. (2010). Behaving as or behaving as if? Children's conceptions of personified robots and the emergence of a new ontological category. *Neural Networks, 23*(8–9), 1099–1103.

Shahriari, K,. & Shahriari, M. (2017). IEEE standard review–Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. In *2017 IEEE Canada International Humanitarian Technology Conference (IHTC)* (pp. 197–201). tex.organization: IEEE.

Sharkey, N., & Sharkey, A. (2010). The crying shame of robot nannies: An ethical appraisal. *Interaction Studies, 11*(2), 161–190.

Shim, J., & Arkin, R. C. (2012). Biologically-inspired deceptive behavior for a robot. In *International conference on simulation of adaptive behavior* (pp. 401–411). tex.organization: Springer.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., & Guez, A. (2017). Mastering the game of Go without human knowledge. *Nature, 550*(7676), 354–359.

Slors, M. (1996). Why Dennett cannot explain what it is to adopt the intentional stance. *The Philosophical Quarterly, 46*(182), 93–98.

Sparrow, R., & Sparrow, L. (2006). In the hands of machines? The future of aged care. *Minds and Machines, 16*(2), 141–161.

Spatola, N., Belletier, C., Chausse, P., Augustinova, M., Normand, A., Barra, V., & Huguet, P. (2019). Improved cognitive control in presence of anthropomorphized robots. *International Journal of Social Robotics, 11*(3), 463–476.

Spatola, N., Monceau, S., & Ferrand, L. (2019). Cognitive impact of social robots: How anthropomorphism boosts performances. *IEEE Robotics & Automation Magazine, 27*(3), 73–83.

Spatola, N., & Normand, A. (2020). Human vs. machine: The psychological and behavioral consequences of being compared to an outperforming artificial agent. *Psychological Research,* 1–11.

Spunt, R. P., Meyer, M. L., & Lieberman, M. D. (2015). The default mode of human brain function primes the intentional stance. *Journal of Cognitive Neuroscience, 27*(6), 1116–1124.

Stich, S. P. (1985). Could man be an irrational animal? Some notes on the epistemology of rationality. *Synthese,* 115–135.

Terada, K., Shamoto, T., Ito, A., & Mei, H. (2007). Reactive, Movements of Non-Humanoid Robots Cause Intention Attribution in Humans. In *2007 IEEE/RSJ international conference on intelligent robots and systems* (pp. 3715–3720). tex.organization: IEEE.

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.

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).

Theodorou, A., Wortham, R. H., & Bryson, J. J. (2016). Why is my robot behaving like that? Designing transparency for real time inspection of autonomous robots. In: *AISB workshop on principles of robotics*. tex.organization: University of Bath.

Turkle, S. (2010). In good company?: On the threshold of robotic companions. In *Close engagements with artificial companions* (pp. 3–10). Benjamins.

Urgen, B. A., Kutas, M., & Saygin, A. P. (2018). Uncanny valley as a window into predictive processing in the social brain. *Neuropsychologia, 114,* 181–185.

Wagner, A. R., & Arkin, R. C. (2011). Acting deceptively: Providing robots with the capacity for deception. *International Journal of Social Robotics, 3*(1), 5–26.

Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science, 5*(3), 219–232.

Weick, K. E. (1995). *Sensemaking in organizations* (Vol. 3). Sage.

Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (2005). Organizing and the process of sensemaking. *Organization Science, 16*(4), 409–421.

Wiese, E., Buzzell, G. A., Abubshait, A., & Beatty, P. J. (2018). Seeing minds in others: Mind perception modulates low-level social-cognitive performance and relates to ventromedial prefrontal structures. *Cognitive, Affective, & Behavioral Neuroscience, 18*(5), 837–856.

Wiese, E., Metta, G., & Wykowska, A. (2017). Robots as intentional agents: Using neuroscientific methods to make robots appear more social. *Frontiers in Psychology, 8,* 1663.

Willemse, C., Marchesi, S., & Wykowska, A. (2018). Robot faces that follow gaze facilitate attentional engagement and increase their likeability. *Frontiers in Psychology, 9,* 70.

Ziemke, T. (2020). Understanding robots. *Science Robotics, 5,* 46.