Joint Interaction and Mutual Understanding in Social Robotics

Sebastian Schleidgen · Orsolya Friedrich

Received: 30 March 2022 / Accepted: 17 September 2022 / Published online: 26 October 2022
© The Author(s) 2022

Abstract
Social robotics aims at designing robots capable of joint interaction with humans. On a conceptual level, sufficient mutual understanding is usually said to be a necessary condition for joint interaction. Against this background, the following questions remain open: in which sense is it legitimate to speak of human–robot joint interaction? What exactly does it mean to speak of humans and robots sufficiently understanding each other to account for human–robot joint interaction? Is such joint interaction effectively possible by reference, e.g., to the mere ascription or simulation of understanding? To answer these questions, we first discuss technical approaches which aim at the implementation of certain aspects of human–human communication and interaction in social robots in order to make robots accessible and understandable to humans and, hence, human–robot joint interaction possible. Second, we examine the human tendency to anthropomorphize in this context, with a view to human understanding of and joint interaction with social robots. Third, we analyze the most prominent concepts of mutual understanding and their implications for human–robot joint interaction. We conclude that it is—at least for the time being—not legitimate to speak of human–robot joint interaction, which has relevant implications both morally and ethically.

Keywords Anthropomorphism · Human–Machine interaction · Human–Robot communication · Cybernetics · Phenomenology · Social cognition

If we are to take on thinking partners – or at the least thinking servants – in the form of machines, we will be more comfortable with them, and we will relate to them more easily, if they are shaped like humans. (Asimov 1990).

Sebastian Schleidgen
sebastian.schleidgen@fernuni-hagen.de

1 FernUniversität in Hagen, Institute of Philosophy, Universitätsstrasse 33, 58097 Hagen, Germany
Introduction

This paper is dedicated to the question of whether, or rather: in what sense it is legitimate to speak of human–robot joint interaction. In what sense can we speak of humans and robots sufficiently understanding each other to account for human–robot joint interaction? Alternatively, could such joint interaction also be effectively possible by reference, e.g., to the mere ascription or simulation of understanding? To answer these questions, we, first, provide some general background on modern social robotics. Second, we go into some detail regarding the technical approaches to implement certain aspects of human–human communication and interaction in social robots with the aim of making robots accessible and understandable to humans and, therefore, human–robot joint interaction possible. Third, we examine the human tendency to anthropomorphize in this context with a view to human understanding of and joint interaction with social robots. Fourth, we analyze the most prominent concepts of mutual understanding and their implications for human–robot joint interaction. This provides the backdrop against which we, fifth, provide an answer to the question in what sense we may speak of human–robot joint interaction. As will be shown, all approaches analyzed are either implausibly reductionist or link demands to mutual understanding that make it questionable whether artificial entities like social robots can satisfy them appropriately. As a consequence, it is, at least for the time being, not legitimate to speak of human–robot communication, understanding, and joint interaction.

Background

Modern social robots are expected to interact with humans in various contexts and in ways that are accessible and understandable to humans. Accordingly, social robotics aims at developing “[…] autonomous or semiautonomous robots that are natural and intuitive for the general public to interact with, communicate with, work with as partners, and teach new capabilities” (Breazeal et al., 2016; cf. Pipitone & Chella, 2021 for a recent example). Usually, three types of social robots are distinguished (Sandry, 2015a, 2015b): human-like or android robots, animal-like as well as machine-like or “overtly other” robots. Social robots are used in different contexts and for different purposes (Table 1), and make use of different communication modalities to interact with humans, such as whole-body motion, proxemics, gestures, facial expressions, gaze behavior, head orientation, linguistic or emotive vocalization (Breazeal et al., 2016).

From the very beginning, social robotics aimed at developing robots not only entertaining interaction with the purpose of being accessible and understandable, but being capable of joint action with humans (Clodic & Alami, 2021; Clodic et al., 2017; Glasauer et al., 2010; Grigore et al., 2013; Jokinen, 2018; Mörtl et al., 2014). Joint action is generally understood as a specific form of interaction with “two or more individuals coordinat[ing] their actions in space and time to bring about a
change in the environment” (Sebanz et al., 2006).¹ Human–robot joint action, hence, essentially refers to coordinated and goal-directed action of at least two (human and robotic) individuals.

This raises the question of the conceptual presuppositions as well as the requirements of human–robot joint action. In social robotics, there are two main approaches to solve the problem of enabling robotic joint action with humans: on the one hand, attempts that refer to human–human joint action and aim at implementing certain human characteristics in robots deemed relevant for joint action (Breazeal et al., 2016). On the other hand, accounts assuming a general impossibility of such implementation and, hence, calling for an alternative understanding of human–robot joint action and opposing efforts to implement human characteristics in robots for the purpose of joint action (Peters, 1999; Sandry, 2015b). Social robotics mainstream mostly follows the first approach (Breazeal et al., 2016). This, however, shifts the question of the requirements of human–robot joint action towards the question of the requirements of human–human joint action as well as to possibilities of their implementation in robots.

Human–human joint action is usually said to at least depend “on the abilities (i) to share representations, (ii) to predict actions, and (iii) to integrate predicted effects of own and others’ actions” (Sebanz et al., 2006; cf. Gallagher, 2020). Although this has provoked extensive debates on the exact determination of the necessary

---

¹ Joint action has to be distinguished from collective action, which refers “to any action performed by an entity which is a group of several individuals” and which does not “imply anything about actions taken by individual members of a collective” (Hammond, 2016).

**Table 1** Exemplary purposes of social robots

| Robotic purpose                                | Examples                                                                 |
|-----------------------------------------------|--------------------------------------------------------------------------|
| Verbal communication                          | Mel (Sidner et al., 2005)                                                |
| Nonverbal communication                       | Pepper (Wada et al., 2006)                                               |
|                                               | Kismet (Jokinen, 2018)                                                   |
| Emotional communication                       | AIBO (Fujita, 2004)                                                      |
|                                               | Keepon (Kozima, 2006)                                                    |
|                                               | ROMAN (Berns & Hirth, 2006)                                              |
|                                               | WABIAN ( Ogura et al., 2006)                                             |
|                                               | WE-4RII, WF-4RII (Itoh et al., 2006; Solis et al., 2006)                  |
| Communication-based services                  | Valerie (Gockley et al., 2006)                                           |
| Support of people with communication problems | KASPAR (Dautenhahn et al., 2009)                                         |
| Shared attention and collaboration            | Leonardo (Breazeal et al., 2004)                                         |
| Education                                     | NAO (Pollack et al., 2002)                                               |
| Elderly Care                                  | Paro (Wada et al., 2006)                                                 |
|                                               | Pearl (Jokinen, 2018)                                                   |
| Study effects of human presence               | Geminoid (Shimada et al., 2006)                                          |
| Study social signaling                        | Nexi and Maddox (Jung et al., 2013)                                      |
conditions for joint action, particularly on the role and meaning of “shared intentions”, “joint attention”, and “task sharing”, as well as “action observation” and “coordination”, it is widely undisputed that human–human joint action requires sufficient **mutual understanding** of the involved individuals, at least with a view to the relevant aspects regarding the context and purpose of a joint action (Bratman, 1993, 2013; Breazeal et al., 2016; Butterfill, 2012; Clodic & Alami, 2021; Gallagher, 2020; Tomasello et al., 2005). Such understanding may refer to, among other things, the emotions, feelings, reasons, and intentions of others. In general, this means an understanding of certain mental states of others, to make their actions sufficiently predictable. Ultimately, only the satisfaction of this very condition of sufficient mutual understanding makes the integration of predicted action effects—and, hence, joint action as goal-directed coordinated action—possible at all (Overgaard & Michael, 2015; Robillard, 2006; Strasser, 2020). It is crucial that sufficient **mutual understanding** generally refers to, or is based on, _interactions_ of different kinds and on different levels between at least two entities. Therefore, we prefer the term of **joint interaction** over the term of joint action when it comes to mutual understanding.

It remains unclear, however, what consequences this reference to mutual understanding has for joint interaction of humans and robots, especially if one orients human–robot joint interaction to human–human joint interaction. A preliminary answer may be: if joint interaction requires sufficient mutual understanding of the involved parties, then sufficient mutual understanding of humans and robots (and, hence, the satisfaction of its conditions) is necessary for successful human–robot joint interaction. It follows that human agents would have to sufficiently understand their robot partners in contexts of joint interaction _and_ vice versa (Clodic & Alami, 2021). In this context, human-like and animal-like robots have proven to be particularly promising (Breazeal et al., 2016; Sandry, 2015b). This is due to the strong tendency in humans to _anthropomorphize_ or _zoomorphize_ robots, i.e., to ascribe “for instance emotional states, motivations, intentions […] by the user to the robot” (Lemaignan et al., 2014; cf. Epley et al., 2008) resulting in humans being convinced of _understanding_ robots (e.g. Banks, 2020; Fink, 2012; Reeves & Nass, 1996; Schreck et al., 2019). This ultimately led to the focus on anthropomorphic (and zoomorphic) design principles (Kiesler et al., 2008; Lemaignan et al., 2014; Powers et al., 2003), as well as to approaches to implement the above-mentioned communication modalities in social robots.\(^2\) The question arises, however, whether this follows a plausible conceptualization and may result in a corresponding realization of sufficient human understanding with a view to human–robot joint interaction.

As regards sufficient **robotic** understanding of human agents in the context of joint interaction, however, the matter is more complicated. It seems obvious, as Strasser (2020) puts it, that “if artificial agents qualify as social agents in a joint action, we have to expect [the ability to understand others] from them”. In view of

\(^2\) For the purpose of this article, we leave aside the question whether this enables communication and **human** understanding of robots (referring to, e.g., emotions, reasons, and intentions) in a plausible sense, i.e., the question of what exactly humans would have to understand in robots.
this, numerous concepts of and approaches to an artificial theory of mind (Bianco & Ognibene, 2019; Butterfill & Apperly, 2013; Chen et al., 2021; Duarte et al., 2018; Iwahashi, 2003; Scassellati, 2002; Strasser, 2020; Vanderelst & Winfield, 2018; Winfield, 2018), or, more broadly, of robotic understanding (Clodic et al., 2017; Hellström & Bensch, 2018; Jacq et al., 2018) have been developed in contexts of human–robot joint interaction. However, these approaches mostly focus exclusively on the ability of social robots to statistically predict human intentions and actions (Sciutti et al., 2018). In many contexts, this approach seems to work well and human–robot joint interaction to be successful. In analogy to the case of humans (supposedly) understanding robots, however, the question arises whether this implies a plausible conceptualization and may result in a corresponding realization of sufficient robotic understanding with a view to human–robot joint interaction. Authors like Fiebich et al. (2015) seem to be affirmative although not explicitly concerned with the issue of robotic understanding.

Others dispute the key premise of Strasser, namely that any agent to qualify as a social agent in a joint interaction must have certain capacities, e.g. the ability to understand others. According to them, this, in fact, would imply the impossibility to treat robots as social agents and, hence, to legitimately speak of human–robot joint interaction. Instead, they postulate “new conceptual tools for forms of non-reciprocal or asymmetric sociality, i.e., for social interactions where one agent lacks the capacities required for normative agency” (Seibt, 2017). This leads to the proposal that robots may simulate (joint) interaction in different ways, resulting in different degrees of sociality and, with that, joint interaction (Seibt, 2014, 2017). If this holds true, at least certain robots would indeed qualify as social agents (to some degree), and certain human–robot interactions would qualify as joint interactions (to some degree).

The question of how to conceptualize and eventually realize robotic understanding with a view to human–robot joint interaction is not merely conceptually interesting, but of key relevance for normative considerations. On the one hand, it is morally relevant in the sense that any answer to it influences our everyday moral beliefs, rules, norms and values. For instance, the way we see and talk about robots, e.g., as joint interactors, significantly influences our perceptions and moral expectations of robots (within the boundaries of our moral community). On the other hand, answering the question of robotic understanding is ethically relevant, insofar it is vital for a (philosophical or scientific) reflection on moral phenomena regarding the relationship of humans and robots. For instance, the attribution of certain properties and abilities to robots may imply certain rights of or duties towards robots. Ultimately, speaking of human–robot joint interaction implies that we are dealing with robotic social agents whom we should treat in a certain way or from whom we would expect to be treated in a certain way, who have certain responsibilities in joint interactions or are responsible for certain consequences of joint interactions. In other words: the question of whether it is legitimate to speak of human–robot joint interaction may be relevant for the question of whether it is ethically justified to treat robots as moral subjects or moral patients (Damiano & Dumouchel, 2018; Nye & Yolbas, 2021; Nyholm, 2020; Véliz, 2021). In the light of these normative dimensions, non-reciprocal or asymmetric sociality, as postulated, e.g., by Seibt (2014, 2017) seems
implausible as it would either result in lowering established moral or ethical standards in view of the capabilities of robots or lead to double-standards with a view to the concept of joint interaction.

Having laid out this, we may also, in contrast to, e.g., Hakli (2014), not “argue that if people conceive of their interaction with robots as social interaction, this should count as prima facie evidence that their interaction with robots is social interaction“. Instead, from an ethical perspective, we must conceptually analyze any claim of human–robot social, e.g., joint interaction, at least with a view to its potential moral and ethical implications. For only such analysis can provide an appropriate foundation for deciding whether robots may count as social agents and joint interactors.

The Technology Behind “Understanding” and “Understandable” Social Robots

When it comes to promoting the accessibility as well as understandability of robots and, hence, human–robot joint interaction, the mainstream of social robotics focuses on the implementation of certain human, or at least: human-like, characteristics and abilities in robots (Brinck & Balkenius, 2020). With a view to the understandability of robots, communicative characteristics and abilities are deemed particularly vital. In this regard, two main issues arise: first, the question of an appropriate robotic communication repertoire, i.e., the set of abilities considered necessary for adequate communication with humans. And second, the issue of implementing corresponding communication architectures, i.e., the concrete technical realization of these very abilities in robots. As regards the communication repertoire, social-emotional intelligence as well as socio-cognitive skills are at the center of considerations (Breazeal et al., 2016; cf. Clodic et al., 2017; Goswami & Vadakkepat, 2019; Nyholm, 2020; Von Braun et al., 2021): depending on the intended use and context, it is deemed necessary that robots are able to recognize and interpret affective signals from humans, that they possess internal models of emotions (usually based on certain psychological theories of emotion), and that they are able to communicate affective states. Furthermore, to allow for successful human–robot joint interaction, social robots should be able to recognize, understand, and predict human behavior with a view to its underlying mental states (such as beliefs, intents, desires, and feelings) and react accordingly.

As regards the technical realization of these abilities, social robotics strongly focusses on possible ways to—verbally and non-verbally—exchange information (cf., e.g., Bisk et al., 2016; Edwards et al., 2020). For verbal exchange, natural language processing (NLP) modules are implemented to communicate primary information either in turn-based or in dynamic conversations that may account for social signals and other background information relevant for human–robot dialogue interactions in a dynamically changing world (cf. Jokinen, 2018). As regards non-verbal information-exchange, different types of regulators for gaze and pose behavior, displays indicating affective, cognitive, or conversational states, illustrators allowing for deictic and iconic gestures, as well as sensors and mechanisms for tactile
interaction are used to communicate secondary information intended to supplement verbally exchanged information (Breazeal et al., 2016).

Besides allowing robots to understand humans, the overarching goal of these efforts consists in making robots understandable to humans and, hence, in increasing human willingness to engage with robots, to interact cooperatively and effectively, and, ultimately, to jointly interact with them (Breazeal et al., 2016; Sandry, 2015b). In this regard, anthropomorphic design principles play an important role, which refer to the much-discussed human tendency to anthropomorphize non-human beings and entities. This, however, requires elaborating what the concept of anthropomorphism and the tendency to anthropomorphize alludes to. In particular, such an elaboration is necessary concerning the question of sufficient mutual understanding as a necessary condition for legitimately speaking of human–robot joint interaction.

**Questioning the Significance of Anthropomorphizing Social Robots for Mutual Understanding and Joint Interaction**

Generally, anthropomorphism is understood as “the human tendency to attribute human traits to non-human entities” (Damiano & Dumouchel, 2018; cf. Epley et al., 2007; Złotowski et al., 2015), or to interpretate “non-human behavior as motivated by human feelings and mental states” (Airenti, 2015). Hence, the concept does not only refer to the human-like appearance of non-human beings or entities, but also to the attribution of certain human competencies and characteristics.

As regards appearance, the tendency to anthropomorphize has, for instance, been demonstrated for robots with humanlike faces and voices (Nass & Brave, 2005; Yee et al., 2007), or other bodily and behavioral similarities or commonalities (Ames, 2004). As regards the attribution of competencies and characteristics, in view of certain robotic features people tend to assume that robots have, e.g., certain levels of knowledge regarding specific facts and phenomena, that they have certain characters, or attitudes towards humans, objects, as well as social phenomena (Lee et al., 2005; Powers et al., 2005). This often implies projecting one’s attitudes onto robots as well as developing specific attitudes towards robots, sometimes culminating in cooperating with robots “as […] with a real person” (Kiesler et al., 2008). Therefore, while certain similarities or commonalities of humans and robots seem to be a necessary condition for anthropomorphizing robots, ongoing human–robot interactions may even increase the degree to which humans perceive robots as being similar or sharing commonalities and, hence, the degree of accepting robots as cooperators (Sandry, 2015a).

The projecting of one’s attitudes is of particular interest in this context, as it shows the crucial role of pre-existing attitudes for the ways in which robots are anthropomorphized, as well as for the degree of human willingness to cooperation and joint interaction: this willingness decisively depends on the basic attitudes and values one holds, especially when it comes to interaction with non-human beings and entities (Richert et al., 2018). This is what Kiesler et al. (2008) have in mind when stating that “anthropomorphism is partly a value prescription process that facilitates potential interaction” and, ultimately, communication and joint interaction. This is
one of the reference points for understanding “anthropomorphism as a fundamental tool in successful human–robot relations” (Damiano & Dumouchel, 2018) and, accordingly, for introducing anthropomorphic design principles in social robotics. For “then face-to-face interaction with a humanlike machine should prompt greater anthropomorphism of the machine” (Kiesler et al., 2008) and, in turn, support further interaction, communication, and joint interaction. According to this position, the tendency to anthropomorphize is not understood as a cognitive mistake, as is often argued, but rather as a basic tool that can be used to improve human–robot interaction: “humanlike […] social robot designs are based on the assumption that familiar appearance and modes of communication […] are the best way to support easy and effective human–robot interactions” (Sandry, 2015a). However, this has not remained uncriticized. From an ethical perspective, for instance, a number of authors is concerned that anthropomorphizing robots could result in cognitive and psychological damage or even in reduced quality of life (Damiano & Dumouchel, 2018; cf. Allen & Wallach, 2012; Calo, 2012; Whitby, 2012).

More important for our purpose, however, are critiques that focus on the connection of anthropomorphizing robots on the one hand and understanding them on the other: in this regard, one line of fundamental criticism is of particular interest. Following the Levinasian tradition, “the other always retains a level of alterity and therefore cannot be completely comprehended by the self” (Sandry, 2015a). Communication with and understanding the Other, according to this position, necessarily points toward, or even is reliant on, the self. Consequently, binding communication, understanding and joint interaction to certain commonalities with humans—as anthropomorphic design principles do—would only result in “[c]ommunication as reduplication of the self or its thoughts in the Other” (Peters, 1999). While, according to this position, the persisting difference between the self and the Other would make communication necessary in the first place, creating merely external similarities between the self and the Other per se does not allow for communication or even mutual understanding, but rather only for projecting one’s own attitudes, beliefs and intentions in the Other (here: the robot) (Sandry, 2015a).

A second line of argument holds that making use of the human tendency to anthropomorphize robots could be deceptive and, hence, lead to false expectations. According to this argument, the anthropomorphic design of robots would present them to humans as “artifacts that have inner states of mind”. Therefore, human–robot interaction could promote the illusion of “understanding these states of mind” (Turkle, 2005; cf. Damiano & Dumouchel, 2018) and, as a consequence, maybe even expectations of robots being capable of understanding humans or following human social norms and reciprocity conditions (Syrdal et al., 2008; cf. Złotowski et al., 2015). This is not necessarily to say that the human tendency to anthropomorphize is, in principle, a cognitive mistake (even though Turkle seems to take this position), but rather that it is misleading when exploited in designing robots and, as such, having many important implications for, e.g., human self-understanding.

In summary, it seems to be true that an anthropomorphic design of robots affords the attribution of certain traits, abilities, and characteristics (e.g., knowledge, intelligence, understanding) (Lee et al., 2005; Powers et al., 2005) and, hence, fosters human willingness to engage with robots, to interact cooperatively and effectively,
and, ultimately, to jointly interact with them (Bailenson & Yee, 2005; Kiesler et al., 2008; Parise et al., 1999; Torrey et al., 2006). At the same time, however, this attribution and willingness do only seem to establish a certain kind of one-way relation: attributing certain traits, characteristics and abilities to robots ultimately is the result of projecting human values, attitudes, characteristics, and abilities onto robots, which, in turn, promotes willingness to interact with a view to certain human goals (for a more detailed discussion of this point cf., e.g., Brinck & Balkenius, 2020; Damiano & Dumouchel, 2018; Duffy, 2003).

At this point, recall the necessary conditions for joint interaction: it requires successful communication with the goal of sufficient mutual understanding, at least with a view to the relevant aspects regarding the context and purpose of a joint interaction. In this regard, successful communication may presuppose, among other things, sufficient understanding of the emotions, feelings, reasons, and intentions, in general: of the mental states of others, to make their actions sufficiently predictable with a view to goal-directed coordinated action. Therefore, it remains questionable whether mere attribution or projection (e.g., of robots communicating with and understanding humans) is sufficient to appropriately speak of humans understanding robots. Vice versa, it is also questionable whether existing robotic communication repertoires and architectures (such as the ones outlined above) are sufficient to appropriately speak of robots understanding humans. Hence, the question remains whether, and if so, in what sense, an anthropomorphic design of robotic communicative characteristics and abilities allows for sufficient human–robot mutual understanding as a necessary condition for joint interaction. Ultimately, this alludes to the question of whether mere attribution of certain characteristics and abilities as communicative is sufficient, or whether these properties must in fact be realized in robots to foster (mutual) understanding and possibly joint interaction. To answer this question, it is worth taking a more detailed look at the most important approaches and theories dealing with communication and (mutual) understanding.

**Communication and Mutual Understanding**

Craig (1999) identified seven traditions of communication theory: first, rhetorical approaches, which theorize communication as “the practical art of discourse”; second, semiotic approaches defining communication as “intersubjective mediation of signs”; third, phenomenological approaches which grasp communication as “experience of otherness” and dialogue; fourth, cybernetic approaches conceptualizing communication as “information processing”; fifth, sociopsychological approaches, which capture communication as “expression, interaction & influence”; sixth, sociocultural approaches conceiving communication as “(re)production of social order”; and seventh, critical approaches framing communication in terms of “discursive reflection”. Such different approaches to communication result in highly differing conceptions of understanding. Although almost all these approaches share certain aspects and can in principle be used to analyze and evaluate different aspects of social robots and human–robot interaction, the cybernetic and the phenomenological
accounts seem particularly interesting concerning our question of (mutual) understanding and joint interaction with robots.

**Communication and Mutual Understanding in the Cybernetic Tradition**

It is often claimed that the cybernetic tradition initiated modern communication theory, following the work of, e.g., Claude E. Shannon, Warren Weaver, John von Neumann, and Alan Turing (Craig, 1999; Heims, 1991; Krippendorff, 1989). Simultaneously, the influence of cybernetics is still observable in current theories in, e.g., systems and information science, cognitive science or artificial intelligence (Craig, 1999). Thus, it is still an important point of reference in information technology and social robotics development. Cybernetics theorizes communication as *mathematically quantifiable and predictable information processing and exchange*, which can take place in and between all kinds of complex systems. Accordingly, not only human individuals may communicate with each other, but also humans and artificial entities (like, e.g., robots), groups, organizations, and whole societies. Furthermore, individual thought of living and non-living entities is framed as intrapersonal communication, resulting from the idea that thought would be “nothing more than information processing” (Craig, 1999).

As a consequence, successful communication in and between entities “is linked with the clear transmission or exchange of accurate information” in this model (Sandry, 2015a). Semantic aspects of information, however, are excluded from this model (Shannon, 1948). This is due to the aim of Shannon (and other proponents of the transmission model) to mathematically quantify and predict information and communication: as meaning depends on semantic interpretation of information, which, in turn, is essentially dependent on subjective preconditions of individuals involved in communication, it would seem not appropriate to quantify and predict it mathematically and, hence, to include it in the transmission model (Kimmel, 2020).

Concerning our question of (mutual) understanding and human–robot joint interaction, it could be stated that, according to the transmission model, successful communication of some information from a robot to a human (or vice versa) would result in some sense of understanding in this very human (or robot). However, such successful communication is solely measured in terms of the mathematical predictability for information being appropriately reproduced in the receiver: “[t]he fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point” (Shannon, 1948). Even mental states would then have to be regarded as or as a result of information processing, and would have to be grasped and transferred mathematically accurate with a view to communication and understanding. This is probably the reason why cybernetics and the transmission model continue to play an important role in machine learning (ML), robotics and particularly in the development of anthropomorphic robots (Sandry, 2015b; cf., e.g., Li et al., 2016 for a mathematical model of conversational

---

3 Although, e.g., Weaver (1949) suspected that the aspect of meaning in principle could be included and various attempts have been made to do so (Bao et al., 2011).
agents). We doubt, however, that this model is appropriate to make the actions of joint interactors sufficiently predictable with a view to goal-directed coordinate action: its inherent reduction of communication to mathematically quantifiable and predictable information processing and exchange does not seem to account for any plausible version of human communication, understanding and joint interaction.

Communication and Mutual Understanding in Phenomenology

A rather different account of communication and (mutual) understanding follows from the phenomenological tradition. Interestingly, its founder Edmund Husserl (2012 [1931]) emphasized the importance of commonalities between individuals involved in communicative processes. In view of this, and similar to the cybernetic tradition, his theory is sometimes said to be “the very foundation of modern theories of communication” (Chang, 1996) as well as to be reflected in the assumption of modern social robotics “that humanlike robots are likely to make more effective social robots” (Sandry, 2015a). In fact, there seems to be some similarity between Husserl’s approach and the cybernetic tradition, although Husserl did not allude to a mathematical reduction of communication, but rather to commonalities of form and behavior in communicating individuals.

In contrast, phenomenologists like Emmanuel Levinas focus on the value of difference between individuals, which ultimately would make it impossible to completely understand the Other. As a consequence, accounts of communication presupposing similarities or commonalities of individuals in terms of successful communication and understanding would fall prey to inauthenticity (Levinas, 1989). This is the reason why phenomenologists of this tradition often criticize social robotics and particularly the development of anthropomorphic robots. Indeed, authors like Pinchevski (2005) evaluate traditional communication theories as, for instance, those of cybernetics, to be violent because their “descriptions of communication can be understood to support the elimination of otherness and difference, resulting in a level of ‘violence’ to the other” (Sandry, 2015b).

It is important, however, that, according to Levinas, this difference between individuals is the very reason for the necessity of communication in the first place (Levinas, 1969, 1989; cf. Sandry, 2015a). As regards our question of (mutual) understanding as necessary condition for human–robot joint interaction, according to this concept of communication, understanding the other and, hence, mutual understanding is principally problematic. With, e.g., Stegmaier (2020), the strongest possible claim seems to be that understanding the other consists of our “permanent testing of interpretations we expect others may have”. The implications of this for human–robot joint interaction remain, however, open. However, it could also be followed that robotic communication and understanding are not readily

---

4 In the following, we leave aside post-phenomenological approaches (cf., e.g., Ihde, 1990; Rosenberger and Verbeek, 2015) as they are mainly concerned with human-world relations mediated by technology and not so much with phenomena like communication, understanding and joint interaction as we discuss them here.
comparable to human–human communication and understanding and, hence, that a completely different approach is needed for human–robot joint interaction (Damiano & Dumouchel, 2018; Duffy, 2003; Stegmaier, 2020; Złotowski et al., 2015). It still remains an open question how such an approach could be conceptualized. Such a speculative position therefore does not allow for an answer to our question regarding human–robot joint interaction.

Communication and Mutual Understanding in Social Cognition Theories

Regarding the abilities of understanding the mental states of others as well as anticipating or predicting their behavior (possibly with the aim of goal-directed coordinate and, hence, joint interaction), a third line of theories comes into play, which relates these abilities to so-called “mindreading” or “Theory-of-Mind” (Fodor, 1992). Roughly, theories of this kind can be divided into cognitivist and interactionist theories (De Jaegher et al., 2010; Gallotti & Frith, 2013). The most important variants of the cognitivist theory are the well-known approaches of theory-theory and simulation theory.

Theory-theory (as developed, e.g., by Gopnik & Wellman, 1992; Leslie, 1994) assumes that mindreading and, hence, the ability to understand others’ mental states, requires a Theory-of-Mind. Accordingly, mindreading is based on implicit knowledge of a set of “folk” laws or principles that link mental states to sensory stimuli, behavioral responses, and other mental states (Barlassina & Gordon, 2017). Hence, the ability to represent and think about the mental states of others is said to be based on a tacit psychological theory, and mindreading is defined as an information-rich and a theory-driven process. However, approaches of theory-theory differ as to where this information stems from. According to Overgaard and Michael (2015), it is either.

(a) [a] “theory” that is formed on the basis of observation, testing, and learning […];

or:

(b) [c]ontained in a “module” that is activated at some point in development […].

In contrast, proponents of simulation theory (e.g., Gordon, 1986; Heal, 1995) hold mindreading to be a matter of taking others’ perspectives, where mental simulation plays a central role. Hence, when we mindread, we represent the mental states of others by simulating their mental states in our own minds, “by putting ourselves in other people’s ‘shoes,’ using our own mind to work out what we would do, think, or feel in their situation—and then attributing those intentions, thoughts, or emotions to those other people” (Overgaard & Michael, 2015). However, if this holds true, simulation theorists continue, i.e., if our cognitive resources can be modified to function as representations of other people’s mental states, then we do not need to store general information about how people function (Barlassina & Gordon, 2017). Accordingly, mindreading is defined as information-poor and purely process-driven.

As regards our question of (mutual) understanding as necessary condition for human–robot joint interaction, it becomes clear that referring to either
theory-theory or simulation theory would set fairly high standards for the development of social robots. First, “a wide range of complex mental states” (Strasser, 2020) would be required, including, e.g., desires and emotions. Second, a psychological theory (in a theory-theory-driven approach) —either acquired through learning or implemented in some module—, or the ability to simulate mental states of others (in a simulation-theory-driven account) would be necessary. In order to qualify as mindreaders, robots (and humans) would not only “have to notice that another agent is noticing something relevant for the joint action, but [...] also [have to] recognize whether the other agent is desiring something or is afraid of something” (Strasser, 2020). Although, in fact, much research is done on these subjects, they still are a particularly challenging aspect in the design and development of social robots.

A solution to this problem could be to elaborate certain minimal conditions of mindreading and, consequently, of (mutual) understanding and joint interaction, as, for instance Butterfill and Apperly (2013) have presented in their account of a Minimal Theory of Mind. The authors propose tying the ability of artificial agents to mindread to two “mental states”, namely their encountering (i.e., a kind of simple perception) and their registration (i.e., a rudimentary form of believing). Some authors find this approach to be “a promising starting point to characterize mindreading abilities of artificial agents” (Strasser, 2020). However, this line of thinking seems in some sense to be similar to the proposal of new conceptual tools for forms of non-reciprocal or asymmetric sociality by Seibt (2014, 2017), which we already rejected due to moral and ethical considerations.

Another solution would be to provide an alternative to theory-theory and simulation theory as a basis for (mutual) understanding. This brings into play interactionist theories of mindreading as developed, among others, by Shaun Gallagher (2001, 2008, 2020). In contrast to theory-theory and simulation theory, Gallagher focuses on (mutual) understanding in embodied and enactive processes of interaction and assumes that mental states of others are not necessarily hidden away. Rather, they are accessible through the actions, bodily expressions, etc. of others. This is essentially based on two theses: first, that thinking is not necessary or essentially an internal matter; and second, that thinking is socially situated. As regards the former, Gallagher, by referring to Merleau-Ponty (1964), argues that at least a part of our thinking would take place through speech acts. Even more, thinking could be co-constituted by speech, gestures, and facial expressions. This would mean that which speech acts one performs, as well as their meaning, necessarily depends on the social context. However, this does not allude to the position that speech acts would only be social interventions. Rather, they are inherently social, insofar as speech acts can only be understood regarding their embeddedness in a certain social context.

From these two theses Gallagher concludes that mental states of others can be perceived directly in the context of social interaction. In fact, we can see the emotions and intentions of others if we were to understand them as patterns of situated physical movements and expressive properties in normative social contexts, and we can hear the beliefs of others when they communicate them to us. This would mean to hear the other thinking, since at least some of her beliefs are realized in the context of their linguistic expression in the first place.
Regarding our question of (mutual) understanding as necessary condition for human–robot joint interaction, Gallagher’s focus on embodied and enactive processes of interaction is of particular importance, as it shifts the question of necessary conditions for joint interaction towards the question of necessary conditions of such embodied and enactive processes. Consequently, if it was possible to interact with robots in embodied and enactive processes, it would also be possible to jointly interact with them. This again would set fairly high standards for the development of social robots, for then robots would have to be designed in a way that would render them embodied and enactive in the very same sense as humans are, according to Gallagher. Whether this can be accomplished is, however, an open question despite numerous research activities on this matter.

**Conclusion**

At the center of our considerations was the question in what sense it could be legitimate to speak of human–robot joint interaction and in what sense we could speak of humans and robots sufficiently understanding each other to account for human–robot joint interaction. The discussion of modern approaches to the implementation of certain communicative properties of humans in robots turned out to be insufficient for enabling mutual understanding, if only because their reference to the human tendency to anthropomorphize merely establishes a one-way relation, in which certain human values, attitudes, characteristics, and abilities are projected onto robots.

The cybernetic transmission model of communication also was shown to be insufficient with a view to mutual understanding and joint interaction. Although it is, in fact, compatible with many technical approaches in social robotics, its inherent reduction of communication to mathematically quantifiable and predictable information processing and exchange does not seem to account for any plausible version of human communication, understanding and joint interaction. Following phenomenologists like Levinas, on the other hand, it could be stated that communication and understanding is not readily comparable to human communication and understanding, which would mean that a completely different approach is needed for human–robot joint interaction. However, it remained an open question how such an approach could be conceptualized. Finally, defining mutual understanding in terms of theory-theory, simulation theory and interactionism proved to be too rich in preconditions to be appropriately implemented in social robots.

To summarize, all approaches analyzed here are either implausibly reductionist or link demands to mutual understanding that make it questionable if artificial agents like social robots can satisfy them appropriately. Simultaneously, we face the fact that joint interaction with robots seems to be possible in certain contexts (Breazeal et al., 2016). In principle, two answers seem possible to this observation: first, it could be stated that human–robot communication, understanding, and joint interaction have to be understood in a less demanding sense than human–human communication, understanding and joint interaction, e.g., in accordance with a Minimal Theory of Mind (Butterfill & Apperly, 2013), or with a view to the proposal of non-reciprocal or asymmetric sociality (Seibt, 2014, 2017). This alternative, however,
is to be rejected due to moral and ethical considerations. Second, at least if any of the discussed accounts of mutual understanding would be sufficiently plausible and would, thus, have to be adapted to robots to legitimately speak of human–robot joint interaction, it follows that human–robot communication, understanding, and joint interaction are, at least for the time being, not possible. Consequently, humans and robots may cause events in the world together, but joint interaction—in a more sophisticated understanding of the term—is only seemingly possible at the moment. It is then morally as well as ethically important to not raise expectations of human–robot communication, understanding, and joint interaction too high. It is not legitimate to speak of human–robot joint interaction in the sense we speak of human–human joint interaction. Hence, on a moral level it is not legitimate to have any moral expectations of robots. And it is not ethically justified to treat robots as moral subjects or moral patients.

Acknowledgements This work was supported by the German Research Foundation (DFG) [Grant Number 418201802]. We are grateful to Nicole Falkenhayner for proofreading the manuscript. We thank three anonymous reviewers for their helpful comments on an earlier version of the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Airenti, G. (2015). The cognitive bases of anthropomorphism: From relatedness to empathy. International Journal of Social Robotics, 7(1), 117–127. https://doi.org/10.1007/s12369-014-0263-x
Allen, C., & Wallach, W. (2012). Moral machines: Contraction in terms or abdication of human responsibility. In P. Lin, K. Abney, & G. A. Bekey (Eds.), Robot ethics: The ethical and social implications of robotics (pp. 55–68). MIT Press.
Ames, D. R. (2004). Inside the mind reader’s tool kit: Projection and stereotyping in mental state inference. Journal of Personality and Social Psychology, 87(3), 340. https://doi.org/10.1037/0022-3514.87.3.340
Asimov, I. (1990). Robot visions. ROC.
Bailenson, J. N., & Yee, N. (2005). Digital chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments. Psychological Science, 16(10), 814–819. https://doi.org/10.1111/j.1467-9280.2005.01619.x
Banks, J. (2020). Theory of mind in social robots: Replication of five established human tests. International Journal of Social Robotics, 12, 403–414. https://doi.org/10.1007/s12369-019-00588-x
Bao, J., Basu, P., Dean, M., Partridge, C., Swami, A., Leland, W., & Hendler, J. A. (2011). Towards a theory of semantic communication (extended technical report). Rensselaer Polytechnic Institute Troy NY.

Barlassina, L., & Gordon, R. M. (2017). Folk psychology as mental simulation. Retrieved March 28, 2022, from https://plato.stanford.edu/archives/sum2017/entries/folkpsych-simulation/

Berns, K., & Hirth, J. (2006). Control of facial expressions of the humanoid robot head ROMAN. In: 2006 IEEE/RSJ international conference on intelligent robots and systems, pp. 3119–3124. https://doi.org/10.1109/IROS.2006.282331

Bianco, F., & Ognibene, D. (2019). Transferring adaptive theory of mind to social robots: Insights from developmental psychology to robotics. In M.A. Salichs, S. Shuzhi, E.I. Barakova, J.-J. Cabibihan, A.R. Wagner, A. Castro-González, & H. He (Eds), Social Robotics. ICSR 2019. Lecture Notes in Computer Science (Vol. 11876, pp. 77–87). Springer.

Bisk, Y., Yuret, D., & Marcu, D. (2016). Natural language communication with robots. In K. Knight, A. Nenkova, & O. Rambow (Eds.), Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies, 2016 (pp. 751–761). Association for Computational Linguistics (ACL).

Bratman, M. E. (2013). Shared agency: A planning theory of acting together. Oxford University Press.

Bratman, M. E. (1993). Shared intention. Ethics, 104(1), 97–113. https://doi.org/10.1086/293577

Breazeal, C., Brooks, A., Gray, J., Hoffman, G., Kidd, C., Lee, H., Lieberman, J., Lockerd, A., & Chilongo, D. (2004). Tutelage and collaboration for humanoid robots. International Journal of Humanoid Robotics, 1(2), 315–348. https://doi.org/10.1142/S0219843604000150

Breazeal, C., Dautenhahn, K., & Kanda, T. (2016). Social robotics. In B. Siciliano & O. Khatib (Eds.), Springer handbook of robotics (pp. 1935–1972). Springer.

Brinck, I., & Balkenius, C. (2020). Mutual recognition in human-robot interaction: A deflationary account. Philosophy & Technology, 33(1), 53–70. https://doi.org/10.1007/s13347-018-0339-x

Butterfill, S. (2012). Joint action and development. The Philosophical Quarterly, 62(246), 23–47. https://doi.org/10.1111/j.1467-9213.2011.00005.x

Butterfill, S., & Apperly, I. (2013). How to construct a minimal theory of mind. Mind & Language, 28(5), 606–637. https://doi.org/10.1111/mila.12036

Calo, R. (2012). Robots and privacy. In P. Lin, K. Abney, & G. A. Bekey (Eds.), Robot ethics: The ethical and social implications of robotics (pp. 187–202). MIT Press.

Chang, B. G. (1996). Deconstructing communication: Representation, subject, and economics of exchange. University of Minnesota Press.

Chen, B., Vondrick, C., & Lipson, H. (2021). Visual behavior modelling for robotic theory of mind. Scientific Reports, 11(1), 1–14. https://doi.org/10.1038/s41598-020-77918-x

Clodic, A., & Alami, R. (2021). What is it to implement a human-robot joint action? In J. Von Braun, M. S Archer, G. M. Reichberg, & M. Sánchez Sorondo (Eds.), Robotics, AI, and humanity: Science, ethics, and policy (pp. 229–238). Springer.

Clodic, A., Pacherie, E., Alami, R., & Chatila, R. (2017). Key elements for human-robot joint action. In R. Hakli & J. Seibt (Eds.), Sociality and normativity for robots (pp. 159–177). Springer.

Craig, R. T. (1999). Communication theory as a field. Communication Theory, 9(2), 119–161. https://doi.org/10.1111/j.1468-2885.1999.tb00355.x

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

Dautenhahn, K., Nehaniv, C. L., Walters, M. L., Robins, B., Kose-Bageci, H., Mirza, S., & Blow, M. (2009). KASPAR—a minimally expressive humanoid robot for human–robot interaction research. Applied Bionics and Biomechanics, 6(3–4), 369–397. https://doi.org/10.1080/11762320903123567

De Jaegher, H., Di Paolo, E., & Gallagher, S. (2010). Can social interaction constitute social cognition? Trends in Cognitive Sciences, 14(10), 441–447. https://doi.org/10.1016/j.tics.2010.06.009

Duarte, N. F., Raković, M., Tasevski, J., Cocom, M. I., Billard, A., & Santos-Victor, J. (2018). Action anticipation: Reading the intentions of humans and robots. IEEE Robotics and Automation Letters, 3(4), 4132–4139. https://doi.org/10.1109/LRA.2018.2861569

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

Edwards, A., Edwards, C., & Gambino, A. (2020). The social pragmatics of communication with social robots: Effects of robot message design logic in a regulative context. International Journal of Social Robotics, 12(4), 945–957. https://doi.org/10.1007/s12369-019-00538-7
Epley, N., Waytz, A., Akalis, S., & Cacioppo, J. T. (2008). When we need a human: Motivational determinants of anthropomorphism. *Social Cognition, 26*(2), 143–155. https://doi.org/10.1521/soco.2008.26.2.143

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review, 114*(4), 864. https://doi.org/10.1037/0033-295X.114.4.864

Fiebich, A., Nguyen, N., & Schwarzkopf, S. (2015). Cooperation with robots? A two-dimensional approach. In C. Messelhorn (Ed.), *Collective agency and cooperation in natural and artificial systems. Explanation, implementation and simulation* (pp. 25–43). Springer.

Fink, J. Anthropomorphism and human likeness in the design of robots and human-robot interaction. In S.S. Ge, O. Khatib, J.-J. Cabibihan, R. Simmons, & M.-A. Williams (Eds.), *International conference on social robotics, 2012* (pp. 199–208). Springer.

Fodor, J. A. (1992). A theory of the child’s theory of mind. *Cognition, 44*(3), 283–296. https://doi.org/10.1016/0010-0277(92)90004-2

Fujita, M. (2004). On activating human communications with pet-type robot AIBO. *Proceedings of the IEEE, 92*(11), 1804–1813. https://doi.org/10.1109/JPROC.2004.835364

Gallagher, S. (2001). The practice of mind. Theory, simulation or primary interaction? *Journal of Consciousness Studies, 8*(5–6), 83–108.

Gallagher, S. (2008). Direct perception in the intersubjective context. *Consciousness and Cognition, 17*(2), 535–543. https://doi.org/10.1016/j.concog.2008.03.003

Gallagher, S. (2020). *Action and interaction*. Oxford University Press.

Gallotti, M., & Frith, C. D. (2013). Social cognition in the we-mode. *Trends in Cognitive Sciences, 17*(4), 160–165. https://doi.org/10.1016/j.tics.2013.02.002

Glasauer, S., Huber, M., Basili, P., Knoll, A., & Brandt, T. (2010). Interacting in time and space: Investigating human-human and human-robot joint action. In: *19th international symposium in robot and human interactive communication, 2010*, (pp. 252–257). https://doi.org/10.1109/ROMAN.2010.5598638

Gockley, R., Simmons, R., & Forlizzi, J. (2006). Modeling affect in socially interactive robots. *ROMAN 2006-The 15th IEEE international symposium on robot and human interactive communication, 2006* (pp. 558–563). https://doi.org/10.1109/ROMAN.2006.314448

Gopnik, A., & Wellman, H. M. (1992). Why the child’s theory of mind really is a theory. *Mind & Language, 7*, 145–171.

Gordon, R. M. (1986). Folk psychology as simulation. *Mind & Language, 1*(2), 158–171. https://doi.org/10.1111/j.1468-0017.1986.tb00324.x

Goswami, A., & Vadakkepat, P. (2019). *Humanoid robotics: A reference*. Springer.

Grigore, E. C., Eder, K., Pipe, A. G., Melhuish, C., & Leonards, U. (2013). Joint action understanding improves robot-to-human object handover. In: *2013 IEEE/RSJ international conference on intelligent robots and systems, 2013*, (pp. 4622–4629). https://doi.org/10.1109/IROS.2013.6697021

Hakli, R. (2014). Social robots and social interaction. In J. Seibt, R. Hakli, & M. Nørskov (Eds.), *Social robots and the future of social relations. Proceedings of robo-philosophy 2014* (pp. 105–114). IOS Press.

Hammond, P. (2016). Distinguishing joint actions from collective actions. *Synthese, 193*(9), 2707–2720. https://doi.org/10.1007/s11229-015-0876-0

Heal, J. (1995). How to think about thinking. In M. Davis & T. Stone (Eds.), *Mental simulation* (pp. 33–52). Blackwell.

Heims, S. J. (1991). *The cybernetics group*. MIT Press.

Hellström, T., & Bensch, S. (2018). Understandable robots-what, why, and how. *Paladyn, Journal of Behavioral Robotics, 9*(1), 110–123. https://doi.org/10.1515/pjbr-2018-0009

Husserl, E. (2012). [1931]). Felix Meiner Verlag.

Ilde, D. (1990). Technology and the lifeworld: From garden to earth. Indiana University Press.

Itoh, K., Miwa, H., Zecca, M., Takanobu, H., Roccella, S., Carrozza, M. C., Dario, P., & Takanishi, A. (2006). Mechanical design of emotion expression humanoid robot we-4rii. In *CISM international centre for mechanical sciences, courses and lectures* (pp. 255–262). Springer.

Iwahashi, N. (2003). A method for forming mutual beliefs for communication through human-robot multi-modal interaction. In *Proceedings of the fourth SIGdial workshop of discourse and dialogue, 2003*, (pp. 79–86).

Jacq, A. D., Magnan, J., Ferreira, M. J., Dillenbourg, P., & Paiva, A. (2018). Sensitivity to perceived mutual understanding in human-robot collaborations. In: *Proceedings of the 17th international conference on autonomous agents and multiagent systems, 2018* (pp. 2233–2235)
Jokinen, K. Dialogue models for socially intelligent robots. In S.S. Ge, J.-J. Cabibihan, M.A. Salichs, E. Broadbent, H. He, A.R. Wagner, & Á. Castro-González (Eds.), *International conference on social robotics, 2018* (pp. 127–138). Springer.

Jung, M. F., Lee, J. J., DePalma, N., Adalgeirsson, S. O., Hinds, P. J., & Breazeal, C. (2013). Engaging robots: Easing complex human-robot teamwork using backchanneling. In *Proceedings of the 2013 conference on Computer supported cooperative work, 2013* (pp. 1555–1566). https://doi.org/10.1145/2441776.2441954

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. https://doi.org/10.1521/soco.2008.26.2.169

Kimmel, M. R. (2020). A realist model of communication. Applications for informational technology and artificial cognitive systems. *International Journal on Information Theory, 9*(3/4), 1–16. https://doi.org/10.5121/ijit.2020.9401

Kozima, H. (2006). An anthropologist in the children’s world: A field study of children’s everyday interaction with an interactive robot. In *Proceedings of the international conference on development and learning, ICDL-2006*, Bloomington, IN, USA, 2006.

Krippendorff, K. (1989). Cybernetics. In E. Barnouw, G. Gerbner, W. Schramm, T. Worth, & L. Gross (Eds.), *International encyclopedia of communication* (pp. 443–446). Oxford University Press.

Lee, S.-I., Lau, I. Y.-m., Kiesler, S., & Chiu, C.-Y. (2005). Human mental models of humanoid robots. In *Proceedings of the 2005 IEEE international conference on robotics and automation, 2005* (pp. 2767–2772). https://doi.org/10.1109/ROBOT.2005.1570532

Lemaignan, S., Fink, J., & Dillenbourg, P. (2014). The dynamics of anthropomorphism in robotics. In *2014 9th ACM/IEEE international conference on human-robot interaction (HRI), 2014* (pp. 226–227).

Leslie, A. M. (1994). Pretending and believing: Issues in the theory of ToMM. *Cognition, 50*(1–3), 211–238. https://doi.org/10.1016/0010-0277(94)90029-9

Levinas, E. (1969). *Totality and infinity*: Duquesne University Press.

Levinas, E. (1989). The other in Proust. In S. Hand (Ed.), *The Levinas Reader* (pp. 160–165). Blackwell.

Li, J., Monroe, W., Ritter, A., Galley, M., Gao, J., & Jurafsky, D. (2016). Deep reinforcement learning for dialogue generation. *arXiv preprint*. https://doi.org/10.48550/arXiv.1606.01541

Merleau-Ponty, M. (1964). *Signs*. Northwestern University Press.

Mörtl, A., Lorenz, T., & Hirche, S. (2014). Rhythm patterns interaction-synchronization behavior for human-robot joint action. *PLoS ONE, 9*(4), e95195. https://doi.org/10.1371/journal.pone.0095195

Nass, C. I., & Brave, S. (2005). *Wired for speech: How voice activates and advances the human-computer relationship*, MIT Press.

Nye, H., & Yolbas, T. (2021). Artificial moral patients: Mentality, intentionality, and systematicity. *The International Review of Information Ethics*. https://doi.org/10.29173/irie418

Nyholm, S. (2020). *Humans and robots: Ethics, agency, and anthropomorphism*. Rowman & Littlefield.

Ogura, Y., Aikawa, H., Shimomura, K., Kondo, H., Morishima, A., Lim, H.-o., & Takanishi, A. (2006). Development of a new humanoid robot WABIAN-2. In *Proceedings 2006 IEEE international conference on robotics and automation, 2006*. ICRA 2006 (pp. 76–81). https://doi.org/10.1109/ROBOT.2006.1641164

Overgaard, S., & Michael, J. (2015). The interactive turn in social cognition research: A critique. *Philosophical Psychology, 28*(2), 160–183. https://doi.org/10.1080/09515089.2013.827109

Parise, S., Kiesler, S., Sproull, L., & Waters, K. (1999). Cooperating with life-like interface agents. *Computers in Human Behavior, 15*(2), 123–142. https://doi.org/10.1016/S0747-5632(98)00035-1

Peters, J. D. (1999). *Speaking into the Air: A history of the idea of communication*. University of Chicago Press.

Pinchevski, A. (2005). *By way of interruption: Levinas and the ethics of communication*. Duquesne University Press.

Pipitone, A., & Chella, A. (2021). What robots want? Hearing the inner voice of a robot. *IsScience, 24*(4), 102371. https://doi.org/10.1016/j.isci.2021.102371

Pollack, M. E., Brown, L., Colby, D., Orosz, C., Peintner, B., Ramakrishnan, S., Engberg, S., Matthews, J. T., Dunbar-Jacob, J., McCarthy, C. E., Thrun, S., Montemerlo, M., Pineau, J., & Roy, N. (2002). Pearl: A mobile robotic assistant for the elderly. *AAAI workshop on automation as eldercare, 2002* (Vol. 2002)

Springer
Powers, A., Kiesler, S., & Goetz, J. (2003). Matching robot appearance and behavior to tasks to improve human-robot cooperation. In: Proceedings of the 12th IEEE international workshop on robot and human interactive communication, 2003. ROMAN 2003 (pp. 55–60). https://doi.org/10.1109/ROMAN.2003.1251796

Powers, A., Kramer, A. D., Lim, S., Kuo, J., Lee, S.-I., & Kiesler, S. (2005). Eliciting information from people with a gendered humanoid robot. In: ROMAN 2005. IEEE international workshop on robot and human interactive communication, 2005 (pp. 158–163). https://doi.org/10.1109/ROMAN.2005.1513773

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

Richert, A., Müller, S., Schröder, S., & Jeschke, S. (2018). Anthropomorphism in social robotics: Empirical results on human–robot interaction in hybrid production workplaces. AI & SOCIETY, 33(3), 413–424. https://doi.org/10.1007/s10458-017-0756-x

Robillard, J. (2006). Philosophy of communication: What does it have to do with Philosophy of social sciences. Cosmos and History: The Journal of Natural and Social Philosophy, 1(2), 245–260.

Rosenberger, R., & Verbeek, P.-P. (2015). A field guide to postphenomenology. In R. Rosenberger & P.-P. Verbeek (Eds.), Postphenomenological investigations: Essays on human-technology relations (pp. 9–41). Rowman & Littlefield.

Sandry, E. (2015a). Re-evaluating the form and communication of social robots. The benefits of collaborating with machinelike robots. International Journal of Social Robotics, 7(3), 335–346. https://doi.org/10.1007/s12369-014-0278-3

Sandry, E. (2015b). Robots and communication. Springer.

Scassellati, B. (2002). Theory of mind for a humanoid robot. Autonomous Robots, 12(1), 13–24. https://doi.org/10.1023/A:1013298507114

Schreck, J. L., Newton, O. B., Song, J., & Fiore, S. M. (2019). Reading the mind in robots: How theory of mind ability alters mental state attributions during human-robot interactions. In: Proceedings of the human factors and ergonomics society annual meeting, 2019, 1550–1554. https://doi.org/10.1177/1071181319631414

Sciutti, A., Mara, M., Tagliasco, V., & Sandini, G. (2018). Humanizing human-robot interaction: On the importance of mutual understanding. IEEE Technology and Society Magazine, 37(1), 22–29. https://doi.org/10.1109/MTS.2018.2795095

Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: Bodies and minds moving together. Trends in Cognitive Sciences, 10(2), 70–76. https://doi.org/10.1016/j.tics.2005.12.009

Seibt, J. (2014). Varieties of ‘as-if’: Five ways to simulate an action. In J. Seibt, R. Hakli, & M. Nørgskov (Eds.), Sociable robots and the future of social relations. Proceedings of robo-philosophy 2014 (pp. 105–114). IOS Press.

Seibt, J. (2017). Towards an ontology of simulated social interaction: Varieties of the ‘as if’ for robots and humans. In R. Hakli, & J. Seibt (Eds.), Sociality and normativity for robots. Philosophical inquiries into human-robot interactions (pp. 11–39). Springer.

Shannon, C. E. (1948). A mathematical theory of communication. The Bell System Technical Journal, 27(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shimada, M., Minato, T., Itakura, S., & Ishiguro, H. (2006). Evaluation of android using unconscious recognition. In: 6th IEEE-RAS international conference on humanoid robots, 2006 (pp. 157–162). https://doi.org/10.1109/ICHR.2006.321378

Sidner, C. L., Lee, C., Kidd, C., Lesh, N., & Rich, C. (2005). Explorations in engagement for humans and robots. Artificial Intelligence, 166(1–2), 140–164. https://doi.org/10.1016/j.artint.2005.03.005

Solis, J., Chida, K., Suefuji, K., & Takanishi, A. (2006). The development of the anthropomorphic flutist robot at Waseda University. International Journal of Humanoid Robotics, 3(02), 127–151. https://doi.org/10.1142/S0219843606000709

Stegmaier, P. (2020). The stupidity of artificial and human intelligence—On limitations of knowledge and certainty. Paper presented at the 2nd philosophy of human technology relations conference, PHTR 2020, Enschede, NL

Strasser, A. (2020). From tools to social agents. Rivista Italiana Di Filosofia Del Linguaggio. https://doi.org/10.4396/AISB201907

Syrdal, D. S., Dautenhahn, K., Walters, M. L., & Koay, K. L. (2008). Sharing spaces with robots in a home scenario-anthropomorphic attributions and their effect on proxemic expectations and evaluations in a live HRI trial. AAAI Fall Symposium: AI in Eldercare: New Solutions to Old Problems, 2008, 116–123.
Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences, 28*(5), 675–691. https://doi.org/10.1017/S0140525X05000129

Torrey, C., Powers, A., Marge, M., Fussell, S. R., & Kiesler, S. (2006). Effects of adaptive robot dialogue on information exchange and social relations. *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction, 2006* (pp. 126–133). https://doi.org/10.1145/1121241.1121264

Turkle, S. (2005). Relational artifacts/children/elders: The complexities of cybercompanions. In *Proceedings of the CogSci workshop on android science* (pp. 62–73).

Vanderelst, D., & Winfield, A. (2018). An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research, 48*, 56–66. https://doi.org/10.1016/j.cogsys.2017.04.002

Vélez, C. (2021). Moral zombies: Why algorithms are not moral agents. *AI & SOCIETY, 36*, 487–497. https://doi.org/10.1007/s00146-021-01189-x

Von Braun, J., S Archer, M., Reichberg, G. M., & Sánchez Sorondo, M. (2021). *Robotics, AI, and humanity: Science, ethics, and policy*. Springer.

Wada, K., Shibata, T., Sakamoto, K., & Tanie, K. (2006). Long-term interaction between seal robots and elderly people—robot assisted activity at a health service facility for the aged. In: K. Murase, K. Sekiyama, T. Naniwa, N. Kubota, & J. Sitte (Eds), *Proceedings of the 3rd international symposium on autonomous minirobots for research and edutainment (AMIRe 2005)* (pp. 325–330). Springer.

Weaver, W. (1949). The mathematics of communication. *Scientific American, 181*(1), 11–15.

Whitey, B. (2012). Do you want a robot lover? The ethics of caring technologies. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot ethics: The ethical and social implications of robotics* (pp. 233–250). MIT Press.

Winfield, A. F. (2018). Experiments in artificial theory of mind: From safety to story-telling. *Frontiers in Robotics and AI, 5*, 75. https://doi.org/10.3389/frobt.2018.00075

Yee, N., Bailenson, J. N., & Rickertsen, K. (2007). A meta-analysis of the impact of the inclusion and realism of human-like faces on user experiences in interfaces. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2007*, 1–10. https://doi.org/10.1145/1240624.1240626

Zlotowski, J., Proudfoot, D., Yogeeswaran, K., & Bartneck, C. (2015). Anthropomorphism: Opportunities and challenges in human–robot interaction. *International Journal of Social Robotics, 7*(3), 347–360. https://doi.org/10.1007/s12369-014-0267-6

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.