Phenomenology and the Cognitive Sciences
https://doi.org/10.1007/s11097-022-09848-0

Understanding Sophia? On human interaction with artificial agents

Thomas Fuchs[1,2]

Accepted: 27 July 2022

Abstract
Advances in artificial intelligence (AI) create an increasing similarity between the performance of AI systems or AI-based robots and human communication. They raise the questions:

(1) whether it is possible to communicate with, understand, and even empathically perceive artificial agents;
(2) whether we should ascribe actual subjectivity and thus quasi-personal status to them beyond a certain level of simulation;
(3) what will be the impact of an increasing dissolution of the distinction between simulated and real encounters.

(1) To answer these questions, the paper argues that the precondition for actually understanding others consists in the implicit assumption of the subjectivity of our counterpart, which makes shared feelings and a „we-intentionality" possible. This assumption is ultimately based on the presupposition of a shared form of life, conceived here as „conviviality."

(2) The possibility that future artificial agents could meet these preconditions is refuted on the basis of embodied and enactive cognition, which links subjectivity and consciousness to the aliveness of an organism.

(3) Even if subjectivity is in principle impossible for artificial agents, the distinction between simulated and real subjectivity might nevertheless become increasingly blurred. Here, possible consequences are discussed, especially using the example of virtual psychotherapy. Finally, the paper makes case for a mindful appproach to the language we use to talk about artificial systems and pleads for preventing a systematic pretense of subjectivity.

Keywords Artificial intelligence · Artificial life · Robots · Embodiment · Consciousness · Aliveness · Empathy · We-intentionality · Communication · Understanding

Extended author information available on the last page of the article

Published online: 28 September 2022

1 Introduction

Since the beginning of modernity, people have tried to create artificial creatures and humanoid automata. Jacques de Vaucanson's flute player, his fluttering, quacking and drinking duck, or the „scribe" created by Pierre Jaquet-Droz in 1774 are well-known examples of the fascination that such lifelike products aroused in their contemporaries. Current robots have left their early predecessors far behind and are about to become common partners of humans as „social robots" – as care robots for the elderly, playmates for children, household helpers, or conversation partners for the lonely.

The problems we run into when interacting with androids are illustrated by „Sophia", a humanoid robot from the company Hanson Robotics (Parviainen & Coeckelbergh, 2021). Sophia has human-like facial expressions, displays about 60 different emotional signals, has a reasonably modulated tone of voice and makes eye contact with people she encounters. She (or „it"? „She" is typically reserved for people, but let's accept anthropomorphism for the moment) answers relatively complex questions, can recognize people and jokes about the English weather on a London talk show.[1] Even if it's just a bluff, her effect is astounding. Sophia is already on the edge of the „uncanny valley" (Mori et al., 2012), as it is called in robotics, the threshold where an android's human resemblance creates in us a feeling of both uncanniness and fascination.

In a different way than in robotics, this threshold is crossed in „Her", a science fiction film by Spike Jonze from 2013: Theodore, a shy but empathetic man, falls in love with a computer program named Samantha, who has no corporeality except for an erotic voice (spoken by Scarlett Johannson). However, as a „learning system", she seems to increasingly develop human sensations and empathy. The more Theodore feels understood by Samantha and finally falls in love with her, the more indifferent he becomes to the question of whether she is a real person or just a simulation – the delightful relationship is enough, and he loses his critical distance.

It seems timely that we account for our interactions with agents that simulate subjectivity and aliveness. After all, artificial systems such as Alexa or Siri and robots like Asimo (Sakagami et al. 2002), iCub (Gaudiello et al., 2016) or Sophia are designed to interact with us as convincingly as possible. The humanoid robot Pepper is able to analyze the facial expressions, gestures, and tone of voice of its human counterparts in order to calculate their emotional state (Pandey & Gelin, 2018).[2] Similarly, so-called „empathic chatbots" are said to exhibit emotional intelligence in order to help people with mental health problems (Devaram, 2020). Moreover, it is now frequently claimed that we can „understand" robots (Hegel et al., 2009; Ziemke, 2020), attribute to them rightly not only „states of mind", but also „desires, knowledge, beliefs, emotions, perceptions" (Hellström & Bensch, 2018), empathize with them (Schmetkamp, 2020), and accept them as partners (Breazal et al. 2004). Robots are thus regarded

[1] „Humanoid robot tells jokes on Good Morning Britain", https://www.youtube.com/watch?v=kWlL4KjIP4M.

[2] In general, „emotional robotics" is now a well-established field of research (Klein & Cook, 2012; Ficocelli et al., 2015).

as „intentional agents", whose „beliefs and desires" should be appropriately understood in order to interact with them (Thellman & Ziemke, 2020; Thellman, 2021). Conversely, robots should „understand others' actions, intentions, and emotions and show emotions themselves" (Brinck & Balkenius, 2020, 54), so that there could be „joint intention" (Breazal et al. 2004), even „mutual recognition" between humans and robots (Brinck & Balkenius, 2020). This development raises a number of interrelated questions:

(1) Is it really possible to *understand* AI systems or robots in the proper sense of the word, i.e., to regard them as agents with beliefs, intentions, and desires? And can there be mutual empathy or „shared goals and shared intentions" between a human and a robot (Herrmann & Melhuish, 2010)?
(2) If this assumption proves to be incorrect presently, could there be a stage in the future development of AI systems where we should actually attribute some kind of subjectivity and thus quasi-personal status to them?
(3) How will our attitudes toward AI systems change as we increasingly interact with them? Will the distinction between simulated and real encounters become increasingly blurred?

These questions will be explored in the following, with a focus on the question of a possible *understanding* of AI systems and robots. To this end, an initial conceptual clarification is needed. (a) „Understanding" here means not just „understanding how something works" – this functional meaning is obviously not meant in the above-mentioned contexts of artifical agents. (b) In what follows, understanding also means more than grasping the semantic meaning of words or other signs, as when one speaks of „understanding a text", for example. Of course, we can „understand" Alexa or Sophia in the sense that we can take their programmed output as information. What is meant in the following is *communicative understanding* in the proper sense, namely understanding the utterances of another as an expression of his or her intentions, beliefs and feelings – in short: understanding not *something*, but *someone*. The question then is whether this concept of understanding can also be applied to artificial systems so that there can be a communication with them in the proper sense.[3]

I will proceed in several steps. First, I will describe the conditions for mutual understanding on the empathic (2.1) and on the semantic level (2.2) and show in each case how talk of „understanding" current artificial systems represents a category mistake. According to my thesis, the basic condition for understanding turns out to be the sharing of a common form of life: sociality presupposes *conviviality.* I then show, by reference to an enactive concept of living beings as autopoietic systems, that artificial systems are unable in principle to fulfill this fundamental condition of understanding (3). In the final section, I examine the possible consequences of a creeping dissolu-

[3] For example, Hellström & Bensch (2018) define „understanding a robot" as „having sufficient knowledge of the robot's state of mind to successfully interact with it", where „state of mind" includes „the intentions, desires, knowledge, beliefs, emotions, perceptions, capabilities and limitations of the robot" (Hellström & Bensch, 2018, 120). Following my definition, this would be equivalent to understanding *someone*. Whether this use of language is justified or a category mistake, will be examined in the following.

tion of the categorial distinctions between genuine sociality and „we-intentionality" on the one hand, and simulated or feigned sociality on the other (4).

2 The preconditions for communicative understanding

Let us first examine in more detail whether we can speak of communicative understanding vis-à-vis an AI system or a robot. We can distinguish two forms of such understanding:

(a) *empathic understanding*, i.e., understanding the other's emotional expression, such as his or her joy or sadness;
(b) *semantic understanding*, i.e. understanding his or her verbal utterances.

In both respects, interaction with an artificial system can give the impression or illusion of *understanding someone*. Let us consider each of them separately.

2.1 Empathic understanding

Social understanding is primarily based on grasping the other's feelings and intentions through intercorporeal empathy (Zahavi, 2015; Fuchs, 2017). It is directed at the emotional expression of others, manifested in their facial and gestural movements, be it in face-to-face encounters or also in watching people in movies or on television. However, this primary empathy is by no means limited to living beings. It can also be directed towards inanimate objects, if they seem – e.g. by their movements – to show expressive or intentional behavior. One example is Heider and Simmel's (1944) famous experiment on simple geometric shapes such as circles or triangles moving around each other, which led people to interprete them in terms of intentional and emotional behavior. Similarly, a robotic lawnmower, „searching" in vain for a charging station for its expiring battery, can easily elicit sympathy. Numerous studies have shown that people treat robots or avatars as if they were living beings endowed with mental states, and cite intentions or desires rather than causes as explanations for their actions (Duffy, 2003; Waytz et al., 2010; Özdem et al., 2017; Harth, 2017).

At the same time, this anthropomorphism is usually accompanied by an „as-if-consciousness", i.e., by the implicit knowledge that what is involved is only an apparent intentionality (Fuchs 2014). We take the „intentional stance" (Dennett, 1987) even towards non-living agents, but without necessarily believing that they actually have genuine intentionality (Thellman et al., 2017).[4] This as-if-consciousness, however, dwindles with the increasing lifelikeness of the objects. We easily perceive human-

[4] I understand the concept of intentionality as necessarily tied to phenomenal consciousness. From Dennett's point of view, on the other hand, computers, robots, and humans alike can be considered under the intentional stance, because it serves only to predict their behavior appropriately; whether they are actually conscious is irrelevant. For a critical evaluation of Dennett's behavioristic position, see also Papagni & Koeszegi (2021).

like voices in particular as an expression of an „inside".[5] Something which listens and responds to us like Siri and Alexa, or advises us and performs services for us, we easily perceive as alive and animated. And when Sophia says in a tender voice, „That makes me happy," it takes some active distancing to realize that there is no one there to feel happy, that it is indeed not an utterance at all. In other words, we should not be deceived by the involuntary empathy to which we tend when objects are sufficiently expressive and life-like; it certainly does not correspond to a real sharing of feelings.

The increasingly perfected simulation of subjectivity and communication thus requires that we reject the pretense of an utterance and take Sophia's talk for what it actually is: hollow words, like those of a parrot. Otherwise, we abandon ourselves to appearances and, like Theodore in *Her*, simply give up the „as-if", the distinction between simulation and reality – in a move that Lombard & Ditton (1997) have termed „willingness to suspend disbelief". In which case the impression of an utterance is no longer rejected but passes over into the *illusion* of empathy or understanding of feelings.

Of the positions that see here not an illusionary but a justified empathy, I pick out only one. By comparing robots to fictional characters, Schmetkamp (2020) has argued that we can indeed empathize with robots „… by either inferring, feeling, interacting, or imagining how they perceive and move in their world" - just as we imagine how a character in a novel or movie perceives, acts, and feels. In this way, we might also „attribute something like a perspectival experience to robots" (Schmetkamp, 2020: 881). Now, there is undoubtedly empathy with fictional persons, such as Anna Karenina, even if we remain aware that they are not real (Fuchs 2014). However, if they *were* real, then our empathy would have an actual counterpart, precisely in the experience of these persons; they would be people like ourselves. In the case of humanoid robots, however, it is the other way around: they *are* quite real, but our empathy with them is only an unjustified anthropomorphism, since it does not correspond to any subjective experience.

Thus, either Schmetkamp's argument again boils down to the indisputable fact that humans easily attribute intentions and feelings to robots (for this, no reference to fictional characters is needed). Or she wrongly transfers the case of fictional characters to humanoid robots, as if they had something in common with Anna Karenina's feelings, with which we could sympathize. The latter seems more likely, because Schmetkamp also ascribes a perceptual experience to robots („a robot literally (e.g. visually) perceives the world in a certain way", p. 890). However, this means a category mistake: Robots can *simulate* perception (as the robotic lawnmawer), but they cannot actually perceive because they do not have subjectivity. Thus, our empathy with them remains without an adequate object.

[5] This corresponds to the neurobiolological findings: Heard voices, of whatever type, are typically processed in the anterior superior temporal sulcus, a brain region that is important for numerous aspects of social cognition including cognitive empathy and perspective-taking (Kriegstein & Giraud, 2004; McGettigan, 2015).

2.2 Semantic understanding

We can also understand utterances in a semantic sense, provided that it is a matter of linguistic communication. This is not necessarily tied to bodily presence, but can also be transmitted as a letter, email or chat. In such cases we still understand the utterance as utterance, i.e., we read it as an expression of the other's intentions, not just as factual information as in a newspaper. But in such communication the possibility of simulation and thus of feigning subjectivity is naturally increased. It is already possible that the friendly online partner or the empathetic online therapist is in fact just a chatbot. Let us assume that the simulation of intentional utterances is so successful that we can no longer recognize it as such and have the compelling impression of a real „counterpart". From this point on, would the attribution of intentionality and thus of subjectivity be justified?

This is the situation underlying Alan Turing's well-known test: a group of test subjects were to communicate in writing with a human and with a computer without having any optical or acoustic contact with either (Turing 1950). If the test subjects were subsequently unable to distinguish between human and computer, then, according to Turing, nothing prevents us from recognizing the latter as a „thinking machine." Critics have rightly pointed out that the Turing test defines thinking and its intentional expression in purely behavioristic terms, namely as the output of a computational system, be it the brain or the computer. Yet to the objection that thinking presupposes subjectivity or consciousness, Turing would reply that we can as little be sure of other humans actually thinking as we can be of machines:

> According to the most extreme form of this view the only way by which one could be sure that a machine thinks is to be the machine and to feel oneself thinking. One could then describe these feelings to the world, but of course no one would be justified in taking any notice. Likewise, according to this view the only way to know that a man thinks is to be that particular man. It is in fact the solipsist point of view. (Turing 1950: 446)

Subjectivity and indeed consciousness as such are, according to Turing, inaccessible and therefore unverifiable. Mere verbal output is sufficient for the attribution of „thinking" – embodied interaction is excluded by the scenario from the outset.

Now, the Turing test has not yet been passed by any AI system. The Loebner Prize, established in 1991 to reward any machine that could, has never had to be paid out. It is not on complex logical questions where AI systems fail, but rather on questions that require common sense and contextual understanding (Moor, 2001), such as: „Where is Peter's nose when Peter is in New York? What does the letter M look like when you turn it upside down? Does my budgie have ancestors who were alive in 1750? How many grains of sand do you call a heap?" Supposedly intelligent systems fail here, especially when it comes to understanding metaphors, irony, or sarcasm. They only know unambiguous individual elements, 0 or 1 – for everything that is ambiguous, enigmatic, vague, or has an atmospheric impression, they lack the sense. The relationship between foreground and background, object and context, that

helps us make sense of such questions, does not exist for them, nor does the shared background of commonsensical knowledge (Dreyfus, 1992; Fuchs, 2021).

But let us assume that future machine learning systems will be able to pass the Turing test – with sufficient training based on myriads of situations, context understanding might eventually be simulated. This is even more likely when the systems are implemented in robots that interact with their environment. Such a system with abilities that equal or even surpass those of the human mind is referred to in the research as „strong AI." So once a future Alexa can carry on any conversation, remember past situations and refer to itself – would we then also have to attribute subjectivity to it and concede that it can „understand" us in a genuine sense?

Searle countered Turing's argument with his well-known thought experiment of the „Chinese room" (Searle 1980). A man who does not understand a word of Chinese is locked in a room containing only a manual with all the rules for answering Chinese questions. The man now receives incomprehensible Chinese characters from a Chinese man through a slit in the room („input"), but with the help of the program is able to find appropriate answers, which he then passes on to the outside („output"). However, as Searle argues, even if the Chinese man outside does not notice the deception, one could certainly not claim of the man in the room that he *understands* Chinese. Searle's „Chinese Room" is, of course, an illustration of a computer which functions completely adequately and yet lacks the decisive prerequisite for understanding, namely intentional (and for that matter, phenomenal) consciousness. Consequently, human understanding cannot be reduced to functional algorithms: even „strong AI", should it be possible at all, would only simulate understanding.

Dennett and others have objected to Searle that while understanding or comprehension cannot be attributed to the person in the room, it can well be attributed to the system as a whole, provided it is equipped with sufficiently complex programs:

> The competence is in the software (…) The central processing unit in your laptop doesn't know anything about chess, but when it is running a chess program, it can beat you at chess, and so forth […] The way to reproduce human competence and hence comprehension (eventually) is to stack virtual machines on top of virtual machines on top of virtual machines – the power is in the system, not in the underlying hardware […] comprehension is an effect created (bubbling up) from a host of competences piled on competences (Dennett, 2013, 325).

However, the idea that AI could eventually reach the level of human, i.e. conscious, intelligence simply by increasing the complexity of the software is no more than an assumption. It is often justified with the principle of recursivity, i.e. the feedback of the state of a system into its further processes. But this principle is already realized in a thermostat, and no one would argue that, e.g. a refrigerator can „feel" too warm and „decide" to lower the temperature. A drone also has all the homing systems and feedback mechanisms that allow it to continuously self-adjust its trajectory, but we are unlikely to attribute to it an *understanding* of its search process or a sense of success when it reaches its target. Whatever properties of a system or whatever relations to its environment are fed into its information processing, there is nothing to suggest that this could at some point produce qualitative experience or understanding. Den-

nett does not even try to make this plausible, but simply defines comprehension in functionalist terms as the result of competences, i.e. of the appropriate performance of a system (e.g. the chess computer) – exactly in the sense of Turing. „Piling up" these competences does not change this.[6]

But what do we actually mean when we talk about someone understanding another's verbal utterance? Obviously not merely that he is able to give a suitable answer to it (even if this is normally a sufficient indication). In other words, it is not enough to link the verbal symbols to a fact represented in one's mind, so that this link becomes the trigger for further chains of symbols and an appropriate linguistic output. All this could also be reproduced by the algorithms of a program, or by Dennett's „virtual machines". Understanding means instead the embedding of the heard words in a context of what is known or pre-understood, so that a feeling of *recognition, congruence and familiarity* arises.

So, for example, when I hear my friend's request, „give me the hammer, please!", I need to match it with my prior understanding of a hammer, at the same time grasp my friend's intention, and finally have the bodily knowledge of how to grasp and hand over a hammer. This familiarity with the words and the meaning of the situation is necessary part of understanding the request. It manifests itself in an implicit feeling of „I understood," which then prompts me to take the appropriate action. Thus, a feeling of familiarity and congruence is the characteristic of understanding – the appropriate response or reaction is merely its consequence. Semantic understanding, too, is therefore by no means a purely functional or cognitive process, but also an affective one; it again presupposes a feeling and thus an experiencing subject. This is where a functionalist description that eliminates subjective, qualitative experience and reduces understanding to a suitable input-output relation fails.

This is even more true if we consider the entire situation of communication: understanding means not only grasping the meaning of another's utterance, but also being aware that he addressed me with his utterance, i.e., *that he intended an understanding*. His *communicative intention* is a necessary part of the utterance that I understand (Grice, 1957). The fact that I thus understand not only the other's words but also the other himself *as an intentional subject* ultimately enables the shared intentionality or „we-intentionality" of understanding. It implies both (a) that I perceive my interlocutor as an intentional agent like myself, and (b) that he in turn has an awareness of me as an intentional agent. This is the reciprocal relation of the *second-person perspective*: each partner in the interaction experiences himself or herself as the other's 'you', as the addressee of his communicative intention: „[T]he unique feature of relating to you as you is that you also have a second-person perspective on me, that is, you take me as your you" (Zahavi, 2015: 93). This, in turn, is the basis for a sense of „we" that connects us with the other person, a feeling of mutual understanding.[7]

[6] It is often assumed that the principle of recursivity offers an adequate explanation of consciousness, namely as a representation or monitoring of a (still unconscious) mental state by a higher system. However, every attempt to explain consciousness through higher-order concepts of reflection, recursivity, or self-modeling leads only to an infinite regress, as shown in detail by Henrich (1982), Frank (2002, 2007), Zahavi (1999, 2006, 2007) and other representatives of the „Heidelberg School".

[7] It should be pointed out that we are only concerned here with a fundamental commonality of intersubjective understanding, which is by no means linked to a positive relationship to the other. It applies in the

Thus, in order to *understand* Alexa or Samantha in the communicative sense, we would have to attribute to them not only an actual understanding of our words in the sense given above, but also a second-person perspective, namely an awareness of us as understanding subjects, along with a communicative intention, i.e. the will to convey something to us with their utterances. Even in a perfect simulation of communication, one which would let an AI system pass the Turing test, this would be lacking; there could be no question of a mutual understanding, let alone „mutual recognition" (Brinck & Balkenius, 2020).

3 Why robots can't experience

I have described the conditions for communicative understanding in the empathic and semantic sense – conditions that are clearly not fulfilled by current AI-based systems:

(1) The involuntary empathy we feel towards artificial agents is based merely on our tendency to anthropomorphism.
(2) The mere transmission of information between such agents and a human does not mean understanding *someone*. In other words, it implies no more understanding than reading an instruction manual, even if it proceeds via verbal interactions.

Now one could argue that the future development of humanoid robots will at some point cross the threshold beyond which we should ascribe subjectivity to them. Already the increasingly perfect simulation – as shown in „Her" – can give rise to doubts as to whether we are not dealing with subjects after all, with the possibility of mutual understanding in the proper sense.

I reject this possibility for the following reasons. (1) Our everyday mutual understanding is not only based on the attribution of intentional states, but more fundamentally on a common form of life: sociality presupposes *conviviality*. (2) AI systems and robots do not belong to this shared form of life, since they do not have a *vital* and thus *phenomenal embodiment*. (3) The approximation of robots to living beings (in the sense of so-called '*Artificial Life*') fails because living beings represent autopoietic systems with a developmental history, which are not accessible to biological engineering.

3.1 Conviviality as basis of social understanding

Already Turing argued that we had no reason to deny subjective states such as beliefs and desires to an AI system, provided its performance was equivalent to that of humans. The insistence on human subjectivity would be based only on our own experience, not on that of others, and was therefore „the solipsistic point of view" (see above, 2.2). However, our assumption that other humans (as well as other higher animals) are conscious is by no means based on solipsism and inference. Subjectivity is not something we first suspect in others and then attribute to them if there are

same way to a heated argument or a hierarchical communication, such as a military order.

sufficient signs for it, as the Theory of Mind assumes (Gallagher, 2001). Rather, we perceive others from the outset *as embodied participants in a common form of life*, in which we do not merely infer selfhood from signs but always already presuppose it.[8] This intercorporeal perception is bound up with our common aliveness, embodiment and life history. We share with others the existential facts of being born and growing, the need for air, food and warmth, waking and sleeping, last not least mortality; and this is the common background against which we also interpret all their verbal utterances. Whatever does not belong to this form of life – i.e. artifacts such as computers or robots – is not subject to the implicit assumption of subjectivity; mere similarities of performance are not sufficient for its attribution.

So if future AI systems or robots are one day able to pass the Turing test, it is not their cognitive performance that should make us believe in a conscious being. Rather, our everyday sharing of emotions and intentions with others presupposes a *sharing of life*. Whatever can feel hunger, thirst, pleasure or pain, joy or suffering, so that we can empathize with these states, must be *of our kind* in the broadest sense, that is, a living being belonging to our species or descended from another species whose expressions of emotion and striving are sufficiently similar to ours. Whatever thinks and considers must also have an awareness of its thinking, thus again be a self-sensing, living being. And whatever speaks to us must be able to give expression to an inner experience, so that a „we-intentionality" emerges. In short, the perception of others as conscious beings is based on the presupposition of a common form of life that enables us to share our experience, or on our *"conviviality"*.[9]

Candidates for an attribution of subjectivity must therefore be of our kind: embodied, moving spontaneously and purposefully, expressive and alive. Could humanoid robots or androids fulfill this requirement? As yet, neither AI systems nor robots convincingly convey the impression of aliveness. However, the implicit presupposition of conviviality might change as we increasingly interact with AI systems. We might be persuaded that while we are not dealing with bodily beings whose life form we share, we are dealing with sentient and experiencing systems of a different kind. Empathy would then decouple from conviviality without succumbing to mere anthropomorphism. Is there, then, the prospect of some form of AI that could make the ontological claim to possess subjectivity and experience such that we can actually *understand* it empathically? Might future humanoid robots not only simulate life but actually come alive, so that we would rightly transfer our empathy to them without being subject to an illusion?

[8] As Merleau-Ponty wrote, „we must abandon the fundamental prejudice according to which the psyche is that which is accessible only to myself and cannot be seen from outside" (Merleau-Ponty, 1964, p. 116). Similarly, Wittgenstein rightly asked: „Do you look into yourself in order to recognize the fury in his face?" (Wittgenstein, 1967, § 220, p. 40).

[9] The term 'conviviality' was introduced by Ivan Illich in his book „Tools for Conviviality" (Illich, 1973) with a socio-critical meaning, namely to designate forms of living together in solidarity as opposed to the confinement of individuals to industrial productivity. Furthermore, the term today often refers to the idea of living together with differences, such as in immigrant or diversity societies. In contrast to these meanings, I use the term to refer to a primary and original kinship that we feel with living beings and other human beings because of common bodily structures, life processes, and life interests. In Peluchon (2019) we find related thoughts on a primary connectedness of the living by means of nourishment, breathing and other basic processes of life.

3.2 Robot functionalism versus vital embodiment

That robots are increasingly capable of simulating certain life functions is undeniable, including sensorimotor functions in particular. Operational mobility and interaction with the environment enable advanced robots to provide new forms of feedback and adaptation that go beyond the capabilities of stationary learning systems. Integrated self-models allow today's robots to localize themselves in space, register the results of their behavior in the environment and modify their own programs accordingly.[10] This suggests what Sharkey & Ziemke (2001) have termed „robot functionalism": a robot with bodily structures and interaction patterns similar to those of human beings could develop intrinsic intentionality or even self-awareness.

But the self-modeling of a robot is not, as is often assumed, a kind of self-awareness. The additional feedback loop, which comes about through an internally generated self-model, does not entail conscious self-reference; for this, the robot would have to *perceive* its self-model and recognize it *as itself*, as with a mirror image. This means, however, that it would have to have – beforehand – a *basal, pre-reflective self-consciousness* which for its part could not be generated by self-modeling – otherwise one would end up in an infinite regress.[11] Neither sensorimotor embodiment nor self-modeling are therefore sufficient for subjectivity. Instead, what is crucial is *vital embodiment*, which, from an enactivist perspective, is the basis of primary self-awareness, and thus, of the continuity of life and mind (Jonas, 1966; Thompson, 2007; Fuchs, 2018, 2020).

Conscious experience, from this point of view, is neither a model of the world nor a model of the self located inside the brain,[12] but primarily an activity of the whole organism in which its current homeostasis manifests itself. The emergence of experiencing is tied to the requirement of living beings to maintain themselves in a precarious equilibrium in exchange with their environment, which is made possible by metabolism (Jonas, 1966). Deviations from homeostasis must be registered and responded to by appropriate adaptive behavior toward the environment if the living being is not to perish (Di Paolo, 2009; Di Paolo, 2018). In higher animals, this happens by *feeling values* that integrally reflect the state of homeostasis in its ups and downs. „The source of feeling is life on the wire, balancing its act between flourishing and death" (Damasio, 2018, 20). Thus, the maintenance of homeostasis, i.e., the internal milieu and with it the viability of the organism, is the primary function of consciousness; this manifests itself in the phenomena of drive, hunger, thirst, displeasure, or satisfaction and pleasure. Consciousness, therefore, does not arise first in the cortex, but results from ongoing vital regulatory processes involving the whole organism, which are already integrated in the brainstem and midbrain centers (Panksepp, 1998; Damasio, 2010; Fuchs, 2018). In this way, a bodily-affective self-

[10] The roboticist Josh Bongard was the first to demonstrate the adaptability of robots on the basis of self-generated body models: a four-legged, walkable robot that has one leg amputated is able to reconfigure its own movement pattern by means of self-modeling, calculation of possible movement variants and repeated tests in such a way that it can walk again even with three legs (Bongard et al. 2006).

[11] This cannot be elaborated in all details here; see also Fuchs 2018, pp. 32 ff. For the aporias of higher-order theories of consciousness as self-representation or self-modeling, I refer again to footnote 6 above.

[12] This „self-model" theory of consciousness has been advocated most notably by Metzinger (2003).

experience emerges, namely the *feeling of life* with its various states of pleasure and displeasure, which, as basic subjectivity, underlies all higher mental functions. One can also express it as follows: all experiencing is a form of life; without life there is no subjectivity (Fuchs, 2018: 78, 94).[13]

In the same way, the *emotions* are also tied to the constant interaction of brain and body. Moods and feelings always involve the entire organism: brain, autonomous nervous system, heart, circulation, respiration, intestines, muscles, facial expressions, gestures, and posture. Every emotional experience is inseparably linked to changes in this body landscape (Fuchs & Koch, 2014).[14] An AI system, however, does not have a biological body and thus cannot have feelings. And of course, every cognition, perception and action is also mediated by the living body, realized through the interactions of brain, organism, and environment – through functional circuits in which our senses and limbs as well as things and other people are involved (Chiel & Beer 1997, Sharkey & Ziemke 2001).

The brain is capable of integrating all these organismic functions – but only within a continuous resonant loop, or a „functional fusion" of brain and body (Damasio, 2010, 273).[15] It is not a control center that receives information and issues commands, but part of the functional whole of body and environment. All these living processes and integrating functions are of a biological and biochemical nature and therefore cannot be simulated even by highly complex computers or AI-based robots. Robotic sensors, actors and digital self-models represent only a „mechanistic embodiment" (Sharkey & Ziemke, 2001) superficially similar to the human body and its functions. Without a biological body in metabolic exchange with the environment, the prerequisite for basal self-awareness and thus also higher-order consciousness is missing.

3.3 Autopoietic versus Artificial Life

Now, in robotics we are not only dealing with the simulation of expressions of life but increasingly also with the mimicking of adaptation, learning and development, as it characterizes the ontogeny and life course of higher organisms. Robots equipped with Artificial Neural Networks are able to „learn" from interactions with their environment, for example by reinforcement learning or evolutionary adaptation techniques (generation of new behavioral variants, selection and implementation of successful variants). Their behaviors are no longer determined solely by pre-programmed rules but by a „memory" of their interactions. Thus, one also speaks of „evolutionary

[13] The counterargument is that all these life processes need only be *represented* in the brain to be experienced, in which case they would not be constitutive for consciousness. But the integration that the brain undoubtedly provides is based on a continuous circular feedback between central and peripheral processes or between basal areas of the brain and the body as a whole; this interaction does not allow „representations" to be separated from what is represented. The integration, which corresponds to conscious experience, is therefore not an „image in the brain," but includes at every moment the organism itself. For a detailed critique of representationalism in brain research, see Fuchs (2018: 38 ff., 118 ff.) as well as Di Paolo et al. (2017, 11–40).

[14] According to Damasio, not only the feeling of being alive but feelings in general are „the subjective experience of the momentary state of homeostasis within a living body" (Damasio, 2018, 37).

[15] On the impossibility of a „brain in a vat" that models the body and the world without constitutive or strong embodiment, see Cosmelli & Thompson 2011.

robotics", or „Artificial Life" (Ziemke & Sharkey 2001, Kim & Cho, 2006, Bongard, 2013). Are we now dealing with the transition to technically generated living beings, to which we would have to ascribe something like self-preservation, self-development, and purposefulness, at least in principle?

The reasons for the principal distinction between living beings and machines have already been repeatedly pointed out (von Uexküll 1973, 1982, Maturana & Varela, 1980, Zlatev 2003, Sharkey & Ziemke 2001), and I will only mention the most important arguments here. The central difference is undoubtedly the *autopoietic organization* of living beings, which implies a special, reciprocal relationship between parts and whole (Varela, 1997). The organism as a whole makes possible the existence of the parts, cells and organs of which it is itself composed. It produces and reproduces the parts, which in turn, through their interaction, enable the persistence of the organism. Self-preservation therefore means self-reproduction: the living system separates itself from the environment by a semi-permeable membrane, which at the same time enables the metabolic exchange that the system requires for constant self-transformation, even down to the smallest parts. The living being thus exhibits a *fluid, dynamic process form*: it continuously incorporates and assimilates new matter, i.e. subjects it to its form and purpose.

In contrast to the autopoiesis of organisms, robots are *allopoietic* machines: they do not manufacture themselves, but are designed as an external synthesis of inanimate and rigid single elements (Maturana & Varela, 1980). As von Uexküll (1982) put it, they are built *centripetally* (the parts are first produced, then combined according to the designers' blueprint), whereas the construction of an animal is *centrifugal*, „from the inside out." Living beings develop from simple cells by self-differentiation and growth, in continuous metabolism, so that all parts form an indivisible unit (Sharkey & Ziemke, 2001; Ziemke, 2016). Artificial systems, on the other hand, may be able to incorporate available materials into their structures, but they do not assimilate and transform them because they have no metabolism – they only need to recharge their batteries from time to time. Likewise, their adaptation or „learning" processes relate only to their functional program, not to their structure and shape. Since artifacts do not undergo autonomous growth and development processes, they cannot die either, but only become defective (Fuchs, 2021).

Thus, the term *Artificial Life* ultimately proves to be a misnomer. There is no artificial life, because life is *per se* not something produced but autopoietic, self-effected and self-developing. Artificial life could therefore at best be life induced by humans: namely by providing all the conditions that must be fulfilled for life to spontaneously emerge and organize itself. But that would not be the production of living things themselves. Even „artificial life" would have to organize itself, develop by itself, and would thus no longer be artificial.[16]

Aliveness is also the prerequisite for *feeling and sensing*, which we presuppose in every empathic understanding of others, because it is through feelings that the living being attributes meaning to its environment for its homeostatic self-preservation. This meaningfulness manifests itself in the values – the attractive or aversive qualities – which the feeling animal discovers in the environment (Zlatev 2003). Mean-

[16] See also my detailed remarks in Fuchs (2021), pp. 35–40.

ingfulness or sense-making is thus originally tied to relevance for self-preservation, that is, to the living individuation of an autopoietic system.[17] An artificial system, on the other hand, has no inherent concern for its self-preservation, it is does not care for anything and so it cannot feel anything, neither pleasure nor suffering: „… the precariousness that grounds the concern inherent in living existence has no counterpart in a computer simulation whose entities are purely logical and hence essentially immortal" (Froese & Taguchi, 2019: 3).[18]

Finally, aliveness is also the basis for the development of differentiated human emotions such as shame, pride, guilt, compassion, etc., that are directed toward more complex, particularly social situations and their values (Barrett, 2005; Klimecki, 2015; Vaish, 2018). These emotions, while no longer aiming at mere survival, nevertheless stem from the biological and psychological history of the individual. Lived, embodied experiences are the basis for a person's emotional life. Moreover, socialization in early childhood also provides the implicit knowledge of *intercorporeality* as well as the *shared background* or commonsensical knowledge that AI-systems lack (Caminada, 2014; see above, 2.2). The history of robots is quite different: human designers have installed the functional states that underlie their behavior (Hofmann, 2018), and the adaptations they might undergo as „learning" systems are not based on any lived experience they might consciously remember.

Even if their programs are embodied in a weak sense, i.e. can perform sensorimotor interactions with the environment, robots lack the vital embodiment that characterizes living beings. And even if their programs can adapt to interactions and environments by means of artificial neural networks, they remain allopoietic machines that do not sustain themselves or evolve by themselves through metabolism and growth. Thus, they also lack the prerequisites for the experience of values and meaningfulness. No matter how perfectly they will simulate feeling, perceiving and thinking in the future – if we believe that we can understand them empathically, we are laboring under an illusion. There can be no „shared sense-making" with robots, because this presupposes shared living or conviviality.

[17] See also Gallagher (2011), Di Paolo (2009, 2018). Of course, the specific relevancies, meanings, and norms of the cultural world can no longer be explained in terms of biological self-preservation; the point is only that the precarious state of their organic life is the basis for humans to experience anything at all as valuable or harmful.

[18] The insight that mechanistic embodiment can at best simulate certain intelligence functions, but cannot produce a sense-making, let alone sentient subjectivity, has meanwhile led to projects that aim to realize an „organic embodiment" of robots (Man & Damasio, 2019, Damiano & Stan 2021). If feelings are ultimately expressions of a precarious homeostasis that living beings strive to maintain with their help (Man & Damasio, 2019), then machines would need to implement a process similar to homeostasis that would give them a „concern" for themselves. „This elementary concern would infuse meaning into its particular information processing" (ibid., 446). Soft robotics and synthetic biology should be used to implement such processes, including „soft tissues" equipped with sensors and actuators to create equivalents of intero- and proprioception. Admittedly, these are only projects so far, and the authors themselves express doubts about whether such „homeostatic robots" would not merely represent a simulacrum of subjectivity. The question remains whether „the 'wet' biochemistry of cellular tissue" is not „required for authentic homeostasis and for the mental experience we call feeling" (ibid., 451).

4 The perils of simulation

Even if there can be no AI endowed with subjectivity, sensation or intentionality, and if the simulation of life functions, however perfect, cannot generate consciousness – the advances in simulation technology will not fail to have an effect. The anthropomorphism inherent in our perception and thinking tempts us all too readily to attribute human intentions, actions, and even feelings to our machines. This „digital animism" is already beginning to spread today – either because the categorical difference between subjectivity and its simulation is no longer understood, or because it increasingly appears unimportant. The more frequent and varied the interactions with artificial agents become, the more likely it is that implicit attribution of intentions will emerge (Papagni & Koeszegi, 2021). The as-if-consciousness usually associated with anthropomorphism toward inanimate objects then gives way to illusory understanding. That AI systems supposedly already „think," „know," „plan," „predict," or „decide" paves the way for boundary dissolutions, of which Hans Jonas already warned:

> There is a strong and, it seems, almost irresistible tendency in the human mind to interpret human functions in terms of artifacts that take their place, and artifacts in terms of the replaced human functions. [. .] The use of an intentionally ambiguous and metaphorical terminology facilitates this transfer back and forth between the artifact and its maker. (Jonas, 1966: 110)

Such a dissolution of the categorical differences between subjectivity and its simulation could have far-reaching consequences. Engaging with artificial systems will then increasingly take the place of human relational experiences. If a cuddly robot called „Smart Toy Monkey" is supposed to serve as a friend to small children and thereby promote „social-emotional development;"[19] if friendly nursing robots replace the human care of dementia patients and supposedly listen to their stories (Maalouf et al., 2018); or if patients are prescribed programmed online psychotherapies that save them having to see a therapist (Stoll et al., 2020) – then machines become fake subjects or „relationship artefacts," as Turkle (2011) has put it. They cheat people out of real communication.

Sharkey and Sharkey have argued „…that a deception can be said to have occurred in robotics if the appearance and the way that a robot is programmed to behave, creates, for example, the illusion that a robot is sentient, emotional, and caring or that it understands you or loves you" (Sharkey & Sharkey, 2021: 311). It should therefore be one of the basic ethical requirements for AI systems that they identify themselves as such and do not deceive people who are dealing with them in good faith. Nor should they use emotional language such as „I care", „I like you", „I'm sad", etc. This is particularly true in the areas of child rearing and care of the elderly, where those affected are not yet or no longer able to make the distinction between original and simulation (Epley et al., 2007).

[19] According to the advertisement of the manufacturer Fisher Price (https://www.fisher-price.com/en_CA/brands/smarttoy, last accessed 01.06.2021).

As one example, consider the possible consequences in the field of psychotherapy, where this distinction is certainly important for those affected. Here, mental health apps, virtual psychotherapists and chatbot therapies are increasingly taking the place of trained mental health professionals. Well over 10,000 mental health apps are already available for download on the market (Cabibihan et al., 2013). Particularly relevant for psychotherapy are „conversational chatbots" that conduct a speech-based dialog with humans via an interactive interface. They can imitate a therapeutic conversational style, simulate empathy, and thus create an interaction that sometimes cannot be distinguished from real interventions, even by experts (Fitzpatrick et al., 2017, Inkster et al., 2018, Bendig et al., 2019).

One might assume that users of virtual psychotherapies who are educated about the nature of the intervention maintain an „as-if" consciousness that avoids any illusion of being understood. However, this assumption is premature: users tend to quickly endow technical systems with human-like characteristics. This is called the „Eliza effect", after the computer program that Joseph Weizenbaum, as long ago as the 1960s, used in order to simulate a therapist (Weizenbaum, 1966, Cristea & Sucală 2013). The Eliza effect was confirmed in a recent study with the conversational agent *Woebot*, which supports patients in coping with bereavement or depression (Fitzpatrick et al., 2017). Based on learning networks, *Woebot* provides seemingly understanding responses, empathic affirmations and encouragements that are deceptively similar to a real interaction. The study showed that users (n = 36,070) established personal bonds with *Woebot* that were similar to those in face-to-face cognitive-behavioral therapies (Darcy et al., 2021). Though they were informed that *Woebot* was not a real person, patients endorsed phrases such as the following as frequently as with regard to real therapists:

> *I believe Woebot likes me. – Woebot and I respect each other. – I feel that Woebot appreciates me. – I feel Woebot cares about me even when I do things that it does not approve of.* (Darcy et al., 2021)

It becomes apparent that susceptibility to „digital animism" and the abandonment of the „as-if" is high among *Woebot*'s users. Their emotional distress and neediness can reinforce the general tendency toward anthropomorphism.[20]

The application of AI systems in psychiatry and psychotherapy is often justified with the prospect that they could help reach underserved populations in need of mental health services and promote patients' self-management skills (Blease et al., 2020). The evidence for perceived social support through chatbots is so far inconclusive, but many users seem to appreciate the availability and anonymity of contacts (Wezel et al., 2020). Yet it is obvious that these systems also blur the boundaries between reality, simulation, and fiction, with potentially problematic consequences. For example, the omission of face-to-face interaction in online communication generally favors the projection of feelings onto the virtual counterpart (Fuchs 2014). Thus, there is a risk of transferring emotions, expectations, and (often unfavorable) relationship patterns

[20] Evidence suggests that people who are lonely, lack social contact, or are otherwise vulnerable are more prone to anthropomorphism (Epley 2007).

to the chatbot (Fiske et al., 2019). Unlike the relationship with a real therapist, however, there is no person on the other side of this transference. The projections cannot be perceived by the counterpart, mirrored and resolved in a professional way.

A fortiori, the complex work of hermeneutic understanding cannot be done by an AI apparatus. No machine can see through the patient's behavior in its contrasts between speech and action or in its latent conflicts, recognize the meaning of symptoms on the basis of the patient's life situation and derive conclusions from it. The dialogue with the robot remains on the surface; it can be momentarily pleasant and supportive, but never insightful in the psychotherapeutic sense. Ultimately, the patient remains alone with himself; his need for a trusting relationship, as reflected in the statements quoted above, remains unfulfilled, because this is only feigned by the speech apparatus. He may feel understood, but there is no one who understands him.

5 Conclusion

Advances in simulations make it necessary to clarify the categorical differences between human and artificial intelligence, as well as between living beings and artificial systems. In this paper, I have explored whether we can meaningfully talk about communicating with, feeling empathy for, and understanding AI systems or robots. The result is clear: notions of communication, understanding, and empathy necessarily demand a counterpart endowed with subjectivity, an embodied person with whom we are connected in conviviality. The involuntary anthropomorphism that arises in our perception of AI systems should not deceive us, for it is typical of life-like and expressive objects that we know for certain do not possess subjectivity. Advances in simulation make it increasingly difficult for us to shake off the illusion of a subjective counterpart when dealing with AI; but that is no reason to abandon the distinction between subjectivity and its simulation as such. Rather, it is a reason to strive for a precise use of terms that avoids category errors whenever possible.

I have therefore examined the concept of understanding more closely and shown why it cannot be applied to our interaction with artificial systems and robots. In the *empathic* sense, we can only understand what has sensations and feelings – and robots have no feelings. Likewise, in the *semantic* sense, we can only understand what wants to communicate with us and in turn understands us, that is, what is able to enter into a shared or „we-intentionality". Understanding thus requires not only a transfer of information, or a suitable linking of symbols into syntax, but also an actual experience of meaningfulness and an intertwining of intentions – understanding *someone*, not just something. As I have shown further, this in turn presupposes belonging to a common form of life, or conviviality.

Against the assumption that future AI systems or robots could actually develop a kind of subjectivity, consciousness, or aliveness beyond their increasingly perfect simulation, I have outlined an embodied and enactive view of mind and life. Subjectivity, according to this view, is not a mere product of information processing in the brain, but is tied to the selfhood of an autopoietic organism that maintains itself in demarcation and exchange with the environment. *Vital embodiment* is the primary basis of experience, presupposing the biological processes of homeostasis, metabo-

lism, growth, and cell differentiation, among others. This basis cannot be replaced by an allopoietic machine, however complexly its programs and feedback loops are designed. This also applies to sensorimotor robots, which can model their own state and feed it into their programs, or adapt their behavior through artificial neural networks, but lack the vital embodiment required for subjectivity.

Despite these categorical differences, it can be assumed that the human tendency toward anthropomorphism will be difficult to curb in view of the increasing lifelikeness of AI agents. It is likely to produce a „digital animism" that increasingly blurs the distinction between subjectivity and its simulation. I have illustrated the associated dangers using the example of virtual psychotherapies. The dangers lie above all in the tendency toward projective empathy (Fuchs 2014), i.e., the transfer of feelings, expectations, and hopes onto quasi-subjects with whom there can be no real conviviality or we-intentionality. In this way, they suggest a trusting relationship and understanding, with the risk that their users lack beneficial human interactions.

How can these tendencies be countered? – First of all, it seems necessary to reject the imprecise use of language that blurs the categorial and ontological differences between subjectivity and simulation, the animate and the inanimate, the artificially produced and the naturally developing. This would imply the preferential use of terms such as „simulated intentionality", „seemingly expressive behavior" or „simulated social interactions" for artifical systems (Papagni & Koeszegi, 2021). Second, it is to be demanded that AI systems remain transparent as such, i.e., that they must not systematically deceive humans about their simulation of subjectivity or aliveness. Otherwise, they create a pseudo-community that cheats subjects out of real interaction. Third, there is a need for a new awareness of what embodied interactions and empathic relationships mean to us as social beings. Valuing and nurturing these relationships, rather than increasingly replacing them with virtual quasi-encounters, is likely to become of particular importance in an increasingly digitalized lifeworld.

Acknowledgements I want to thank three anonymous reviewers for their helpful and stimulating comments on earlier versions of this paper.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest There are no conflicts of interest.

References

Barrett, K. C. (2005). The origins of social emotions and self-regulation in toddlerhood: New evidence. *Cognition & Emotion*, *19*, 953–979

Bendig, E., Erb, B., Schulze-Thuesing, L., & Baumeister, H. (2019). The next generation: chatbots in clinical psychology and psychotherapy to foster mental health–a scoping review. *Verhaltenstherapie*, 1–13. https://doi.org/10.1159/000501812. (published online)

Blease, C., Locher, C., Leon-Carlyle, M., & Doraiswamy, M. (2020). Artificial intelligence and the future of psychiatry: Qualitative findings from a global physician survey. *Digital Health*, *6*, 1–18

Bongard, J. C. (2013). Evolutionary robotics. *Communications of the ACM*, *56*, 74–83

Brinck, I., & Balkenius, C. (2020). Mutual recognition in human-robot interaction: A deflationary account. *Philosophy & Technology*, *33*, 53–70

Cabibihan, J. J., Javed, H., Ang, M. Jr., & Aljunied, S. M. (2013). Why robots? A survey on the roles and benefits of social robots in the therapy of children with autism. *International Journal of Social Robotics*, *5*, 593–618

Caminada, E. (2014). Joining the background: Habitual sentiments behind we-intentionality. In A. Konzelmann Ziv, & H. B. Schmid (Eds.), *Institutions, emotions, and group agents. Contributions to social ontology* (pp. 195–212). Dordrecht: Springer

Chiel, H. J., & Beer, R. D. (1997). The brain has a body: adaptive behavior emerges from interactions of nervous system, body and environment. *Trends in Neurosciences*, *20*, 553–557

Cosmelli, D., & Thompson, E. (2011). Embodiment or Envatment? Reflections on the Bodily Basis of Consciousness. In J. Stewart, O. Gapenne, & di E. Paolo (Eds.), *Enaction: Towards a New Paradigm for Cognitive Science* (pp. 361–385). Cambridge/MA: MIT Press

Cristea, I. A., & Sucală, M., David D (2013). Can you tell the difference? Comparing face-to-face versus computer-based interventions. The „Eliza" effect in psychotherapy. *Journal of Cognitive Behavioral Psychotherapy*, *13*(2), 291–298

Damasio, A. (2010). *Self comes to Mind. Constructing the Conscious Brain*. New York: Pantheon Books

Damasio, A. (2018). *The strange order of things: Life, feeling, and the making of cultures*. New York: Pantheon Books

Damiano, L., & Stano, P. (2021). A wetware embodied AI? Towards an autopoietic organizational approach grounded in synthetic biology.Frontiers in Bioengineering and Biotechnology, 9

Darcy, A., Daniels, J., Salinger, D., Wicks, P., & Robinson, A. (2021). Evidence of human-level bonds established with a digital conversational agent: cross-sectional, retrospective observational study. *JMIR Formative Research*, *5*(5), e27868

Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press

Dennett, D. (2013). *Intuition pumps and other tools for thinking*. New York: Norton & Co

Devaram, S. (2020). Empathic chatbot: Emotional intelligence for mental health well-being. *arXiv*: 2012.09130 (https://doi.org/10.48550/arXiv.2012.0913)

Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason*. Cambridge, MA: MIT Press

Di Paolo, E. A. (2009). Extended life. *Topoi*, *28*, 9–21

Di Paolo, E. A. (2018). The enactive conception of life. In: Rietveld, E., Denys, D., Van Westen, M., de Bruin, N. (2016). *The Oxford Handbook of Cognition: Embodied, Embedded, Enactive and Extended*, pp. 71–94. Oxford: Oxford University Press

Di Paolo, E., Buhrmann, T., & Barandiaran, X. (2017). *Sensorimotor life: An enactive proposal*. Oxford: Oxford University Press

Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, *42*, 177–190

Epley, N., Waytz, A., & Caciopo, T. (2007). On seeing human: A three factor theory of anthropomorphism. *Psychological Review*, *114*, 864–886

Ficocelli, M., Terao, J., & Nejat, G. (2015). Promoting interactions between humans and robots using robotic emotional behavior. *IEEE Transactions on Cybernetics*, *46*, 2911–2923

Fiske, A., Henningsen, P., & Buyx, A. (2019). Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *Journal of Medical Internet Research*, *21*(5), e13216

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering Cognitive Behavior Therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Ment Health*, *4*(2), e19

Frank, M. (2002). Self–consciousness and self–knowledge: On some difficulties with the reduction of subjectivity. *Constellations*, *9*, 390–408

Frank, M. (2007). Non-objectal subjectivity. *Journal of Consciousness Studies*, *14*, 152–173

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

Froese, T., & Taguchi, S. (2019). The problem of meaning in AI and robotics: still with us after all these years. *Philosophies*, *4*(2), 14

Fuchs, T. (2014). The virtual other. Empathy in the age of virtuality. *Journal of Consciousness Studies*, *21*, 152–173

Fuchs, T. (2017). Intercorporeality and interaffectivity. In C. Meyer, J. Streeck, & S. Jordan (Eds.), *Intercorporeality: Emerging Socialities in Interaction* (pp. 3–24). Oxford: Oxford University Press

Fuchs, T. (2018). *Ecology of the brain. The phenomenology and biology of the embodied mind*. Oxford: Oxford University Press

Fuchs, T. (2020). The circularity of the embodied mind. *Frontiers in Psychology*, *11*, 1707

Fuchs, T. (2021). Human and Artificial Intelligence: A Clarification. In T. Fuchs (Ed.), *In defense of the human being. Foundational questions of an embodied anthropology* (pp. 13–48). Oxford: Oxford University Press

Fuchs, T., & Koch, S. (2014). Embodied affectivity: on moving and being moved. *Frontiers in Psychology Psychology for Clinical Settings*, *5*, 508

Gallagher, S. (2011). Interpretations of Embodied Cognition. In W. Tschacher, & C. Bergomi (Eds.), *The implications of embodiment: Cognition and communication* (pp. 59–70). Exeter: Imprint Academic

Gaudiello, I., Zibetti, E., Lefort, S., Chetouani, M., & Ivaldi, S. (2016). Trust as indicator of robot functional and social acceptance. An experimental study on user conformation to iCub answers. *Computers in Human Behavior*, *61*, 633–655

Grice, H. P. (1957). Meaning. *Philosophical Review*, *64*, 377–388

Harth, J. (2017). Empathy with non-player characters? An empirical approach to the foundations of human/non-human relationships. *Journal For Virtual Worlds Research* 10 (2)

Hegel, F., Muhl, C., Wrede, B., Hielscher-Fastabend, M., & Sagerer, G. (2009). Understanding social robots. In *2009 Second International Conferences on Advances in Computer-Human Interactions* (pp. 169–174). IEEE

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology*, *57*, 243–259

Hellström, T., & Bensch, S. (2018). Understandable robots – what, why, and how. *Journal of Behavioral Robotics*, *9*, 110–123

Henrich, D. (1982). Fichte's original insight. Trans. D. Lachterman. *Contemporary German Philosophy 1* (pp. 15–53). College Park, PA: Pennsylvania State University Press

Herrmann, G., & Melhuish, C. (2010). Towards safety in human robot interaction. *International Journal of Social Robotics*, *2*, 217–219

Hofmann, F. (2018). Could robots be phenomenally conscious? *Phenomenology and the Cognitive Sciences*, *17*(3), 579–590

Illich, I. (1973). *Tools for conviviality*. London: Calder and Boyars

Inkster, B., Sarda, S., & Subramanian, V. (2018). An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study.JMIR mHealth and uHealth6(11), e12106

Jonas, H. (1966). *The phenomenon of life: Toward a philosophical biology*. New York: Harper & Row

Kim, K. J., & Cho, S. B. (2006). A comprehensive overview of the applications of artificial life. *Artificial Life*, *12*, 153–182

Klein, B., & Cook, G. (2012). Emotional robotics in elder care – A comparison of findings in the UK and Germany. In *International Conference on Social Robotics* (pp. 108–117). Berlin, Heidelberg: Springer

Klimecki, O. M. (2015). The plasticity of social emotions. *Social Neuroscience*, *10*, 466–473

Kriegstein, K. V., & Giraud, A. L. (2004). Distinct functional substrates along the right superior temporal sulcus for the processing of voices. *Neuroimage*, *22*, 948–955

Lombard, M., & Ditton, T. (1997). At the heart of it all: the concept of presence. *Journal of Computer-Mediated Communication* 3 (2)

Maalouf, N., Sidaoui, A., Elhajj, I. H., & Asmar, D. (2018). Robotics in nursing: a scoping review. *Journal of Nursing Scholarship*, *50*, 590–600

Man, K., & Damasio, A. (2019). Homeostasis and soft robotics in the design of feeling machines. *Nature Machine Intelligence*, *1*, 446–452

Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and cognition – The realization of the living*. Dordrecht: D. Reidel

McGettigan, C. (2015). The social life of voices: studying the neural bases for the expression and perception of the self and others during spoken communication. *Frontiers in Human Neuroscience*, *9*, 129

Merleau-Ponty, M. (1964). The child's relations with others. In: *The primacy of perception.* Trans. W. Cobb, pp. 96–155. Evanston: Northwestern University Press

Metzinger, T. (2003). *Being No-one. The self-model theory of subjectivity*. Cambridge, MA: MIT Press

Moor, J. H. (2001). The status and future of the Turing test. *Minds and Machines*, *11*, 77–93

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

Özdem, C., Wiese, E., Wykowska, A., Müller, H., Brass, M., & Van Overwalle, F. (2017). Believing androids – fMRI activation in the right temporo-parietal junction is modulated by ascribing intentions to non-human agents. *Social Neuroscience*, *12*, 582–593

Pandey, A. K., & Gelin, R. (2018). A mass-produced sociable humanoid robot: Pepper: The first machine of its kind. *IEEE Robotics & Automation Magazine*, *25*, 40–48

Panksepp, J. (1998). *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford, New York: Oxford University Press

Papagni, G., & Koeszegi, S. (2021). A pragmatic approach to the intentional stance semantic, empirical and ethical considerations for the design of artificial agents. *Minds and Machines*, *31*, 505–534

Parviainen, J., & Coeckelbergh, M. (2021). The political choreography of the Sophia robot: beyond robot rights and citizenship to political performances for the social robotics market. *AI & Society*, *36*, 715–724

Peluchon, C. (2019). *Nourishment. A philosophy of the political body*. London, New York: Bloomsbury

Sakagami, Y., Watanabe, R., Aoyama, C., Matsunaga, S., Higaki, N., & Fujimura, K. (2002, September). The intelligent ASIMO: System overview and integration. In *IEEE/RSJ international conference on intelligent robots and systems* (Vol. 3, pp. 2478–2483). IEEE

Schmetkamp, S. (2020). Understanding AI—Can and Should we Empathize with Robots? *Review of Philosophy and Psychology*, *11*, 881–897

Sharkey, A., & Sharkey, N. (2021). We need to talk about deception in social robotics!. *Ethics and Information Technology*, *23*, 309–316

Sharkey, N. E., & Ziemke, T. (2001). Mechanistic versus phenomenal embodiment: Can robot embodiment lead to strong AI? *Cognitive Systems Research*, *2*, 251–262

Stoll, J., Müller, J., & Trachsel, A., M (2020). Ethical issues in online psychotherapy: A narrative review. *Frontiers in Psychiatry*, *10*, 993

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

Thellman, S., & Ziemke, T. (2020). Do you see what I see? Tracking the perceptual beliefs of robots. *Iscience*, *23*(10), 101625

Thellman, S. (2021). *Social Robots as Intentional Agents* (Doctoral dissertation, Linköping University Electronic Press)

Thompson, E. (2007). *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Cambridge, MA: Harvard University Press

Turkle, S. (2011). *Alone Together: Why We Expect More from Technology and Less from Each Other*. New York: Basic Books

Vaish, A. (2018). The prosocial functions of early social emotions: the case of guilt. *Current Opinion in Psychology*, *20*, 25–29

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

von Uexküll, J. (1973 [1928]). *Theoretische Biologie*. Berlin:Springer Verlag

von Uexküll, J. (1982). [1940]). The theory of meaning. *Semiotica*, *42*, 25–82

Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J. H., & Cacioppo, J. T. (2010). Making sense by making sentient: effectance motivation increases anthropomorphism. *Journal of Personality and Social Psychology*, *99*, 410–435

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*, 36–45

Wezel, M., Croes, E. A., & Antheunis, M. L. (2020). „I'm here for you": can social chatbots truly support their users? A literature review. In: *International Workshop on Chatbot Research and Design* (pp. 96–113). Cham: Springer

Wittgenstein, L. (1967). In von G. H. Wright (Ed.), *Zettel [Snippets] (G. E. M. Anscombe &*. Oxford: Blackwell

Zahavi, D. (1999). *Self-awareness and alterity: A phenomenological investigation*. Evanston, IL: Northwestern University Press

Zahavi, D. (2006). Thinking about (self-)consciousness: Phenomenological perspectives: Kriegel. In U. Williford, K. (Ed.), *Self-representational approaches to consciousness* (pp. 273–296). Cambridge, MA: MIT Press

Zahavi, D. (2007). „The Heidelberg School and the Limits of Reflection.". In S. Heinämaa, V. Lähteenmäki, & P. Remes (Eds.), *Consciousness: From Perception to Reflection in the History of Philosophy* (pp. 267–285). Berlin Heidelberg New York: Springer

Zahavi, D. (2015). You, Me, and We: The Sharing of Emotional Experiences. *Journal of Consciousness Studies*, *22*, 84–101

Ziemke, T., & Sharkey, N. E. (2001). A stroll through the worlds of robots and animals: Applying Jakob von Uexkülls theory of meaning to adaptive robots and artificial life. *Semiotica*, *134*, 701–746

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

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

Zlatev, J. (2003). Meaning = Life (+ Culture). An outline of a unified biocultural theory of meaning. *Evolution of Communication*, *4*, 253–296

Authors and Affiliations

Thomas Fuchs[1,2]

✉ Thomas Fuchs
thomas.fuchs@urz.uni-heidelberg.de

[1] Phenomenological Psychopathology and Psychotherapy, Psychiatric Clinic, University of Heidelberg, Heidelberg, Germany

[2] Psychiatric Clinic, University of Heidelberg, Voss-Str. 4, D-69115 Heidelberg, Germany