Minds and Machines (2021) 31:361–376
https://doi.org/10.1007/s11023-021-09558-7

ORIGINAL ARTICLE

What Does It Mean to Empathise with a Robot?

Joanna K. Malinowska[1]

Received: 29 September 2020 / Accepted: 9 March 2021 / Published online: 20 March 2021

Abstract
Given that empathy allows people to form and maintain satisfying social relationships with other subjects, it is no surprise that this is one of the most studied phenomena in the area of human–robot interaction (HRI). But the fact that the term 'empathy' has strong social connotations raises a question: can it be applied to robots? Can we actually use social terms and explanations in relation to these inanimate machines? In this article, I analyse the range of uses of the term empathy in the field of HRI studies and social robotics, and consider the substantial, functional and relational positions on this issue. I focus on the relational (cooperational) perspective presented by Luisa Damiano and Paul Dumouchel, who interpret emotions (together with empathy) as being the result of affective coordination. I also reflect on the criteria that should be used to determine when, in such relations, we are dealing with actual empathy.

Keywords Empathy · Social robots · Human–robot interactions · Neuroscience

In the interdisciplinary research field of human–robot interaction (HRI), not only do the empirical sciences, the social sciences and the humanities come together, but so do their methodological perspectives, theoretical assumptions and the terms they use. The intersection of these points of view, although very constructive, requires consistent and meticulous terminological and conceptual verification. In this article, I hope to contribute to this task by analysing the use of the term 'empathy' to study relations between humans and robots. Articles on HRI in which the term 'empathy' appears have been rapidly accumulating in recent years (Malinowska, 2020, 2021). In most cases, researchers borrow the term from psychology (especially social psychology) (Riek, Rabinowitch, Chakrabarti, & Robinson, 2009a, 2009b; Niculescu, van Dijk, Nijholt, Li, & See, 2013; Rosenthal-von der Pütten, Krämer, Hoffmann, Sobieraj, & Eimler, 2013, Rosenthal-Von Der Pütten et al., 2014). HRI scholars also often use a highly reduced interpretation of empathy (Kozima, Nakagawa, & Yano,

✉ Joanna K. Malinowska
malinowska@amu.edu.pl

[1] Faculty of Philosophy, Adam Mickiewicz University, Poznan, Poland

2004; Leite et al. 2013). However, there are also propositions which highlight the cooperational, inter-individual character of this phenomenon. The literature on this subject not only uses this term to describe people empathising with robots (Coeckelbergh, 2010a, 2010b; Darling, Nandy, & Breazeal, 2015) but also robots empathising with people (Kozima et al. 2004; Leite et al., 2014; Riek & Robinson, 2008; Williams, 2012). Nevertheless, what does it mean to say that robots empathise with people? Given that these are machines, not animate beings, how can it even be possible for them to empathise?

I narrow my research to so-called 'social robots' because the concept of empathy appears most often in discussions of their interactions with people (Malinowska, 2020, 2021). Social robots (also known as 'companion robots' or 'artificial companions') can be defined as 'embodied, autonomous entities that can communicate and enter into social relations with people' (Darling, 2016, p. 2). Due to the fact that empathy is a phenomenon closely related to the social dimension of human activity (Redmond, 1989; Stephan & Finlay, 1999), scholars often assume that it also plays an important role in the formation of people's reactions to robots that are designed to perform social functions (Leite et al., 2014; Riek et al., 2009a, 2009b). Some researchers (Leite et al. 2013) even indicate that it is one of the key elements which enable and maintain social relations between people, as well as between people and robotic agents. To understand such opinions, let me start by briefly analysing the phenomenon of empathy itself.

1 Empathy as an Important Element in Building Social Relations Between People and Robots

There are many ways to understand the concept of empathy (Stueber, 2019). In this article, I do not delve into the term's origin and history. Rather, I begin by asking: how is the concept of empathy most often understood? As Amy Coplan (2011, p. 4) notes, empathy can be interpreted as a variety of mental and/or emotional processes. For this reason, in the fields of philosophy, psychology and neuroscience, use of the term 'empathy' to describe and explain all of the above phenomena in many cases leads to great terminological confusion.

Such processes include:

(A) Feeling what someone else feels
(B) Caring about someone else
(C) Being emotionally affected by someone else's emotions and experiences, though not necessarily experiencing the same emotions
(D) Imagining oneself in another's situation
(E) Imagining being another in that other's situation
(F) Making inferences about another's mental states
(G) Some combination of the processes described in (A)–(F) (Coplan, 2011, p. 4).

In HRI, too, this term is understood in various ways (Malinowska, 2020, 2021). Given that empathy allows people to form and maintain satisfying social interaction with another subject, it is no surprise that it is one of the most studied phenomena in the area of HRI. However, the social foundations of the term 'empathy' (Stueber, 2019; Vignemont & Singer, 2006) makes it difficult to apply this term to the study of robots (Malinowska, 2020). The most basic doubt is whether the use of social terms and explanations in relation to these inanimate machines is methodologically justified.

Empathising is often considered to be a process correlated with social cognition (Hoffman, 2001; Sparks, McDonald, Lino, O'Donnell, & Green, 2010) and social emotions (Singer & Lamm, 2009; Singer & Klimecki, 2014). In the case of robots, it is problematic because social relations are traditionally understood as the interaction of two conscious, intentional, rational entities. However, robots do not meet the criteria of rational and intentional subjectivity, and thus cannot be partakers of social relations understood this way. This is one of the reasons why when HRI studies employ the term 'empathy', it is also assumed that this is not a real or true empathy (Misselhorn, 2009; Redstone, 2014, 2017). However, there are theoretical frameworks that allow us to talk about robots as participants in social interaction (Coeckelbergh, 2018; Nickelsen, 2019; Seibt, 2017). One of them is the simulated social interactions framework (SISI) proposed by Johanna Seibt. She argues that 'there is social fictionality but no fictional sociality' (2017, p. 21), and thus something real is going on in the 'social' interactions between robots and humans—something which requires naming and detailed examination. To consider the question about the legitimacy of the use of the term 'empathy' in HRI to be pointless, only because robots are not standard members of the universally acknowledged social ontology, would be a reckless denial that this ontology is constantly changing.

2 Can Robots Empathise?

At first glance, talking about robots having empathy for people seems absurd. However, a great deal of research in the area of HRI indicates that people do indeed sympathise with—and even trust robots—that recognise their emotional states and react adequately to them. And some researchers often call this skill 'empathising'. Such are Hideki Kozima, Cocoro Nakagawa and Hiroyuki Yano (2004). In the article 'Can a robot empathise with people?' they postulate that it is possible to equip robotic agents with mechanisms that will allow them to develop empathy. The researchers define this term as 'the ability to understand another person's mental states (e.g., desire or pain) by putting oneself in the position of that person' (Kozima et al., 2004, p. 83) and suggest that this ability plays an indispensable role in the communication process. In their view, the key to building an empathetic robot is enabling it to perform two functions: maintaining eye contact and engaging in joint attention. They observed that people easily assigned mental states to robots which were able to perform these tasks. Finally, Kozima et al. state that the mechanism of empathy 'based on spatiotemporal coordination of attention and movement of self and another' (Kozima et al., 2004, p. 83) is not only adequate for people, but possibly also for robots. However, the definition of empathising presented above reduces this process to its cognitive aspect, that is, to making inferences about another's mental states, future behaviours and needs. Apart from the impression that the robot is focused on its user, those are the main functions of joint attention and eye contact. This perspective almost completely omits the issue of sharing affective states, which for many scientists is a crucial element of empathising (Coplan, 2011; Goldman, 2011; Vignemont & Singer, 2006). Empathy thus defined does not differ greatly from the cognitive skills related to mindreading (the ability to attribute mental states to ourselves and others). In this case, one should consider whether the use of the term 'empathy' is even necessary and whether it should not rather be reduced to other cognitive abilities.

A different approach to this problem is proposed by Iolanda Leite and her team in the article 'Empathic robots for long-term interactions' (2014). They base their considerations on the definition of empathy formulated by Martin L. Hoffman, who conceptualised this phenomenon as 'not requiring, though often including, a close match' between the affects of empathisers and the people with whom they are empathising (Hoffman 2001, p. 5). Leite et al. argue, that:

> […] artificial companions capable of behaving in an empathic manner will be more successful at establishing and maintaining a positive relationship with users in the long term. Hoffman defines empathy as 'an affective response more appropriate to someone else's situation than to one's own'. To behave empathically, social robots need to understand some of the user's affective states and respond appropriately. However, empathic responses can go beyond facial expressions (e.g., mimicking the other's expression): they can also foster actions taken to reduce the other's distress, such as social supportive behaviours. In fact, the perception of social support has been linked to positive outcomes in children's mental health and coping with traumatic events. Thus, our aim is to study the effects of a computational model of empathy in the long-term relationship established between the robot and the user' (Leite et al., 2014, p. 1).

Hoffman's definition quoted above is usually interpreted in a way which assumes that an 'affective response' to the other person's situation means being in a certain emotional state caused by the observation or imagination of that person. In the approach proposed by Leite et al., it seems to be understood more as taking certain action in response to others' affective state. They do not distinguish between the process of empathising and its behavioural consequences, treating them as an inseparable whole or reducing empathy to 'empathetic behaviour'. In the first case, the empathic response is a product of the processes associated with adopting the cognitive and affective perspective of another person. As Leite et al. are focused on the empathic behaviour and not empathising as such, they essentially omit most of the doubts that the use of this term usually raises in relation to robots (e.g. that they have no feelings, so they simply cannot empathise) (Malinowska, 2020). Their approach is thus very pragmatic or 'functional' (Damiano & Dumouchel, 2020, p. 187), in that it allows HRI to work on pursuing its main objective, which is to develop robots that are capable of giving their users ‘the sense of being with another’ (Damiano & Dumouchel, 2020; Nowak & Biocca, 2003). This perspective abandons the goal of creating socially intelligent robots with a genuine ability to empathise. Instead, it ‘seeks to make robotic agents whose physical appearance and behaviour trigger anthropomorphic projections’ (Damiano & Dumouchel, 2020, p. 187).

However, while this reductionist, behaviourist understanding of empathy is the easiest one to apply to robots, it completely ignores the phenomenal states of the empathiser. This allows Leite et al. to talk about robots having empathy as something genuine without additional theoretical concerns. That would also explain why they use the concepts of empathy and empathic behaviour interchangeably. At the end of their article, however, Leite et al. point out that although the robot can display empathic behaviour, its ‘concern’ for people ‘is not real’ (Leite et al. 2014, p. 11). By ‘concern’ they probably mean the robot’s affective state (the phenomenal experience of concern), which for humans would usually be the reason for taking empathic action. This takes us back to the question of the definition of empathy—it seems that the phenomenal experience of having emotions that occurs in the empathising process is in the end somehow important for these researchers. This is where they indicate that what they mean by robots having empathy for people is just a simulation. But what if we wanted to create truly sentient robotic agents?

Although at present robots cannot experience emotions (as traditionally understood), it is not certain what the future will bring. In fact, robots are already being designed so that their decision-making process is based on so-called ‘artificial emotions’ like happiness, sadness or fear (Guzzi, Giusti, Gambardella, & Di Caro, 2018; Salichs & Malfaz, 2011), but this type of emotion is still a long way from contributing to the achievement of actual sentience. The position that HRI should be focused not on the simulation of social skills in robots but on its authentic reproduction can be called ‘substantial’. Its aim is ‘to endow robots with social abilities’ (Damiano & Dumouchel, 2020, p. 187).

At this point it is also worth recalling Glaskin’s (2012) conception, which uses the definition of emotions proposed by Damasio (1999) to analyse this issue (along with empathy) for HRI. Glaskin (2012, p. 73) refers to the distinction between emotions and feelings, according to which the former are public and external, while the latter are private and internal. Damasio (1999, p. 42) notes that emotions and feelings constitute a continuum in which emotions are (most often) observable reactions to experienced mental states, including feelings. Damasio recognises as emotions such reactions as facial expressions, changes in breath (accelerated or stopped breathing, etc.), changes in tone of voice, pulse, posture, and so on. Glaskin (after Picard, 1997) notes that using the above distinction, it can be assumed that social robots have (or may have in the future) specific emotional states, as evidenced by the following properties: (a) expressing emotions through behaviour; (b) having quick reactions to situations that are associated with experiencing basic emotions (i.e. fear, happiness, sadness, anger, disgust and surprise); (c) processing information about a given situation and drawing conclusions about adequate emotional reactions; (d) ‘emotional experience’, or cognitive access to their own emotional states and their history; and (e) interactions between the robot’s ‘body’ and its ‘mind’, consisting of an emulation of processes such as memory, learning, decision making, physical expression of emotions, and the like. According to the authors cited, meeting all of these conditions will enable us to create machines that will have emotions (Picard, 1997, pp. 135–137; Glaskin, 2012, pp. 75–76).

Furthermore, Glaskin also emphasises the fact that the development of an embodied approach to the mind leads to the thesis that emotions are an extremely important element of cognitive processes. Their application in social robots is thus very important if we want them to expand their cognitive capabilities. Interestingly, due to the fact that robots have bodies that are different to those of humans, it should be assumed that their emotions and related cognitive processes will be fundamentally different from those experienced by humans (Glaskin, 2012, p. 81). From the perspective of this article, however, what is most important is the fact that it is possible to approach the issue of emotions in a way which takes into account social robots, and thus allows them to be considered agents capable of empathy. At present, the question of whether social robots can have affective states analogous to human affective states is the source of considerable controversy. However, with an appropriate theoretical framework (such as Glaskin's) and assuming the further development of robotics, it is possible that robots will reproduce some human cognitive processes and emotions, which will also enable them to empathise.

3 The Relational Turn

Since the main objective of social robotics is to find an answer to the question of how to produce a robot whose behaviour invites its users to enter into and maintain social and emotional relationships, the development of a functional approach to robots within HRI and social robotics has contributed to the authenticity of a robot's affective states becoming a secondary issue. Damiano and Dumouchel (2020, p. 188) explain:

> Within social robotics, the growing attention towards the users' evaluation can be viewed as a general trend, to the extent that the literature characterizes it as a paradigm shift. Raya Jones (2017), for example, describes this orientation as an increasing focus on users' experience that leads specialists to assess robots' 'sociability' primarily—or even exclusively—in terms of the users' evaluation of their social interaction with these artificial agents. Jones interprets this as a 'subtle "paradigm shift"' in the field, which brings specialists to implicitly dismiss the diffused idea that sociability corresponds to a set of individual skills or competences that can be re-enacted, deeply or superficially, by machines. As Jones emphasizes, current procedures assessing robots' sociability do not consider that this ability is an ensemble of traits characterizing the artifacts as individual agents. Evaluations do not focus on the robots' social features, but on the users' perception of their artificial partners, suggesting that robots' artificial sociability arises in, and is a property of, human-robot interaction.

As Prescott (2017) has also noted, it is people's psychological attitudes to the ontological status of robots that essentially shapes their interaction. This is in line with

Gunkel's (2015, 2018, 2020) and Coeckelbergh's (2014) concept of the 'relational turn'. Here, the fundamental statement posits that it does not really matter what features robots actually have: it only matters how people experience them. In this perspective, the distinction between reality and appearance or simulation is blurred. What we have cognitive access to, and how we interpret this information, is our only reality. Therefore, what is experienced as real, regardless of the objective state, really affects people and is thus somehow genuine. If in our experience of a robot we feel and/or think that it is aware and has feelings, it influences our interactions with this agent as well as the way we experience it. This position thus emphasises the importance and reality of the individual phenomenal states of the participants in human–robot interaction. The emotions and empathy generated this way have a variety of consequences for individuals as well as for human society, and therefore cannot be ignored.

The above position shifts the weight of the discussion of the relationship between robots and humans to the emotional, behavioural and cognitive reactions of their users. It is still a relatively unidirectional approach to this issue. Damiano and Dumouchel (2017, 2020), together with Hagen Lehmann (Damiano, Dumouchel, & Lehmann, 2014), are trying to solve this problem quite differently. They propose that the classic definition of emotions should be replaced with a relational definition. They postulate that this would allow the dynamic affective processes arising in this interaction to be described. By the 'classical' definition of emotions, they understand the internalist and individualistic approach, assuming that they are private and internal phenomena, stimulated by environmental circumstances, which need not necessarily lead to external expressions that can be observed by others. They argue:

> […] a proper relational turn requires another step, which changes radically the angle on such a sociability. It demands that the social character of robots, rather than reduced to a users' property (and projection), be recognized as a distributed property. That is, one which emerges from the interactive dynamic taking place between users and robots. A property that can neither be implemented as a trait of individual robots, nor merely understood as their users' projection, because it is distributed in the mixed human-robot system that users and robotic agents together form through their interactions. (Damiano & Dumouchel, 2020, p. 189)

The researchers cite a number of arguments in favour of an embodied, enactive approach in which the body, facial expressions, gestures, and so on are an important part of the process of generating and shaping emotional and mental states. Also note that the development of social robotics increasingly focuses on the task of placing robots in a kind of interaction loop, i.e. giving them characteristics that enable them to involve people in a social situation that generates 'emotional dynamics that include, on both sides, affective expressions and related responses' (Damiano et al., 2014, p. 13). These researchers propose that emotions and empathy be understood in an intersubjective, processual and necessarily social way. They claim that emotions are phenomena shared and co-created by many actors participating in a given social situation. Its participants influence each other on many levels, which makes it difficult to clearly distinguish a single source of these emotions (the source is dispersed).

This perspective breaks with the view that robots merely mimic 'real' affects and that they trick their human interaction partners so that people get the impression that robots are experiencing feelings (Damiano et al., 2014, p. 16; Damiano & Dumouchel, 2017, 2018, 2020). On the contrary—they hold that the complex interactions that arise between robots and people in social interaction are real. According to Damiano et al. (2014, p. 17), this emergentist reinterpretation of emotions will push HRI towards investing in new research and technologies enabling the development of an affective co-evolution established between man and robot, which in time will lead to the emergence of new 'emotional and empathic agents capable of interactively populating human social ecologies'.

4 Is it Justified to Use the Term "Empathy" in the Study of the Relations Between People and Robots?

Damiano and Dumouchel postulate that it is affective coordination which drives the process of the supervision of emotions.

> It is a reciprocal process in which a first agent partially determines the intentions of a second towards her or him, while this second partially determines the first's intention towards the second. […] In this reciprocal process, affective expressions occupy the central place. The process is, however, generally unconscious. We are mostly unaware of the how our affective expression affects other, or even of what—gestures, postures, tone of voice—constitutes this expression. (Damiano & Dumouchel, 2020, pp. 191–192).

Thus, according to these authors, the slow, 'step-by-step' study and development of behavioural interactions between humans and robots should be at the centre of social robotics. But how to recognize the specific emotions and relations emerging on the interaction of two agents? If we are concerned with the methodological question of whether the term "empathy" can be used in the study of the relations between robots and humans, what are the criteria for recognizing such empathising? For Damiano and Dumouchel, probably the most important criterion is the behavioural one. This is also the area that is currently being developed most intensively by HRI and social robotics. So how does the problem of empathising with robots look in this light? Under what circumstances can the interactional loop between humans and robots be called empathy?

Over the past few years, a lot of papers have been published on the subject of humans empathising with robots (Darling, 2015, 2016; Leite et al., 2013, 2014; Niculescu et al., 2013; Seo, Geiskkovitch, Nakane, King, & Young, 2015). Although this is a rather 'hot' topic in the field of HRI, the use of the term 'empathy' in the context of human responses to robots raises some serious doubts. First of all, in most cases the authors of papers on this phenomenon have not questioned whether it is even possible to apply the concept of 'empathy' to HRI (Malinowska, 2020). But since robots are not conscious, sentient beings, when humans interact with such agents, with whom are they empathising? Due to this problem, some scientists have decided to consider human empathy with robots a form of illusion or fantasy

(Misselhorn, 2009; Redstone, 2014, 2017). Let us now consider: is this phenomenon in fact just a form of delusion, or is there more to it?

Observing robots (especially social robots with which an emotional bond has been established) that look as if they are suffering (for example, seeing robots being harassed) inspires human beings to empathise with them. Usually in this case, the folk definition of empathy is used, such as compassion for others, feeling as someone else feels, and so on. The term 'empathy' is often used to describe or study human–robot interactions on three levels: (1) on the level of declared beliefs, (2) on the behavioural level, and (3) on the level of neuronal activity (Darling, 2015, 2016; Darling et al., 2015; Malinowska, 2020, 2021; Scheutz & Arnold, 2016, 2017). As for the first point, there are many reports in both social media and empirical research that use the term 'empathy' to describe people's reactions to robots with which they interact (Rosenthal-von der Pütten et al., 2013, 2014; Darling, 2015). Regarding second point, people's behaviour towards robots (for example, when they refuse to turn robots off, lock them in a closed space, damage them; or when they speak to and about robots in terms that lend them subjectivity and sentience) is often described and analysed with the use of the category of 'empathy' (Riek et al., 2009a, 2009b; Niculescu et al. 2013; Rosenthal-von der Pütten et al. 2013, 2014). Finally, neuroscience research revealed that when people have observed robots being abused (being kicked or bitten), they have reacted similarly to how they reacted when seeing other people being abused in analogous ways. The brain activity of 'observers' associated with empathising was noticed regardless of whether they were looking at humans or robots (Gazzola, Rizzolatti, Wicker, & Keysers, 2007; Rosenthal-von der Pütten et al. 2013, 2014).

Robots are not only able to recognise verbal messages, but also to analyse the tone of a human voice and adapt the content and style of their formulated responses (the volume and pitch of voice varies in people depending on the emotions experienced and is one of the factors that enables robots to identify their affective states; see Niculescu et al. 2013). They can also learn[1] which personality (introverted or extroverted) a person has and adapt their behaviour to it.[2] They can recognise emotions (read facial expressions, recognise the tone of voice, gestures or words) and behave in such a way as to respond 'appropriately' to the affective state of their human companion.

Taking knowledge from the fields of artificial intelligence, HCI and psychology (especially social psychology) and transferring it to HRI has initiated a series of experiments and research projects aimed at determining how robots can identify

[1] In this article, I deliberately use terms that indicate a robot's ability to learn. Although they do not have awareness, robots acquire knowledge and through its implementation (often by trial and error) they learn to use it. Thanks to appropriate algorithms, which usually to some extent reflect the functioning of neural networks, robots learn how to move (Nagata, Sekiguchi, & Asakawa, 1990), and to recognise and use new languages (Lyon, Nehaniv, & Saunders, 2012; Wermter et al., 2004).

[2] Tapus and Mataric's studies have shown that rehabilitation robots were more effective when they encouraged people to exercise by adapting their reactions to the users' personalities. Introverts responded better to calm, supportive comments, while extroverts responded better to loud comments, similar to the shouting of a trainer at a gym (Tapus & Mataric, 2007).

human affects, intentions and expectations, and then adapt their behaviour to these phenomena (Cañamero, 2005; Castellano et al., 2013; Fung, 2015; Kirby, Forlizzi, & Simmons, 2010). This task is quite difficult, because its implementation should take into account factors such as understanding verbal messages (recognising and interpreting their meaning) and recognising facial expressions and body language, in addition to the wider social context that makes the correct interpretation of a given behaviour possible. Thanks to the application of knowledge about the human psyche and the latest technologies (e.g. artificial neural networks), some social robots already have the ability to capture and learn human language and behaviour (including increasingly complex social situations) as well as how to most effectively react to them (Nehaniv & Dautenhahn, 2007).

This is a fundamental issue for the development of human and robot relations. Studies indicates that people are much more sympathetic to and willing to cooperate with robots that seem to accurately recognise their feelings (Riek et al., 2009a, 2009b; Fink, 2012; Leite et al., 2013, 2014). It also directly affects the effectiveness of companion robots. On the basis of experiments carried out by Tapus and Mataric (2007), it can be concluded that robots designed to encourage their users to train regularly (and as a result help them in activities such as rehabilitation or losing weight) were much more effective when they tailored their reactions to their needs (including emotional needs) and behaviour.

Thanks to advanced computational processes that enable robots to perceive their situation and what is happening in their surroundings, they are able to infer that they should express emotion and behave in a certain way, such as asking for help and curling up when they are beaten. Although in the above situation the robot itself does not actually experience any pain, it can still simulate being in a specific affective state through its behaviour. As in many cases robots are designed to induce feelings in their users, people's affective responses (including empathy) to these communications seem to be inevitable. Thus, using the term 'empathy' to study those responses seems justified.

Moreover, let's assume, that the computational processes performed by the robots (as well as the information that they obtain and assimilate in the course of their interactions with humans and according to which they plan their behaviour) are in fact specific cognitive states and a kind of robotic knowledge. Let's also suppose that in the near future robots will be able to experience some types of emotion. But even then, as noted by Mark Coeckelbergh (2010a, 2010b, pp. 4–5), we have no idea what it is to be a robot and we only project our fantasies, beliefs, experiences and expectations onto these entities. Such an argument can be refuted by the observation that such situations are similar to those in which we interact with other people: we draw conclusions about their emotional states based on our own experience. There is one key difference, however. Together with other human beings, we belong to a single species and although we do not really know exactly what it is like to be a person X, we know what it is like to be a human, and therefore, although our conclusions about the affective state of X may not be factual, they are justified to some extent. We have no insight, however, into what it is like to be a non-biological being like a robot. The behavioural criteria we use for other humans do not constitute a reliable justification for our inferences about robots. In other words, no clear conclusions about the affective states of robots should be drawn solely on the basis of their behaviour. A great example of such a situation is pain recognition. As Adamo (2016) notes, robots often behave identically to people in situations that in our case would involve pain. She lists, for example, robots designed to serve future dentists (they flinch and cry when students do their tasks improperly) and those that are equipped with 'artificial emotions' like fear or stress, thanks to which they learn how to get out of difficult situations (Adamo, 2016, p. 77). Adamo also mentions a robotic rat equipped with 32 million silicone neurons and 13 trillion artificial synapses, which begins to move when it identifies a discomfort or threat (when it falls into water, for example, it tries to get out of it as soon as possible). She argues that the behavioural criteria on the basis of which pain is usually inferred can be met by various beings without the accompanying phenomenal pain experience. This situation also has another side: if entities differ significantly from us, we may not recognise the symptoms indicating that they feel pain. Thus, is the adoption of a cognitive perspective completely inappropriate for robots or does its application have validity?

Despite the doubts mentioned above, the answer to this question is not unambiguous. Social robots are designed not only to take anthropomorphic forms but also to emulate anthropomorphic behaviour. The way they formulate their verbal messages, facial expressions and gestures is supposed to be easy for people to understand. Due to partially open learning algorithms, some robots continually develop and learn from people. As we are their closest social environment, their knowledge of social behaviour is largely based on the observation of interpersonal interaction. Thus, as long as they do not develop their own cultures and behaviours, their responses will largely reflect what is happening between people. On this basis, we can draw numerous conclusions about their condition. In fact, we continually do the same in our interactions with other people as well as with animals. In daily life, the behaviour of animals or people is very often the only criterion for our assessments of their situation. In such cases, we usually interpret these behaviours on the basis of our own experience.

So, should we use the term 'empathy' to study human-robots interactions? As I have argued, applying the term 'empathy' to studying human–robot interactions (empathic behaviours) on the three levels of: (1) declared beliefs, (2) behaviour, and (3) neuronal activity, can be methodologically useful. Although the empathy discussed in these situations is often understood in a folk way, through their behaviour and appearance in certain situations, robots are not only able to enter into relationships that evoke a sense of empathy in people, but can also maintain these relationships. On the other hand, people react with great sensitivity to the behaviour of robots, which evoke a feeling of being the object or subject of empathy. Thus, when we consider human–robot interaction from the functional and relational point of view, although many doubts and controversies remain to be analysed in the future, the use of the concept of empathy is justified. On the other hand, the substantial perspective, although it currently does not allow for talking about any 'real' occurrence of emotions or empathy between robots and humans, still leaves some 'hope' for their occurrence in the future.

5 Discussion

Depending on how we define empathy (e.g. functionally, substantially or relationally), our answers as to whether the use of this term is appropriate in studying human–robot interactions will be quite different. Depending on what these answers are, our approaches to robots and their development will also differ.

For example, the question of robots' ability to empathise (and their general sentience), whether understood functionally or substantially, partially shapes the current discussion on the problem of robot rights. There are arguments suggesting that even if robots somehow develop such properties, we may never become aware of it (Metzinger, 2009; Prescott, 2017). This situation results from, among other things, our lack of knowledge about the possibility of phenomenal states emerging and occurring in non-biological beings, and about what their characteristics may be. While the development of science tells us more and more about the inner universe of animals, we know absolutely nothing about what the experience or awareness of non-biological life forms may look like. Thus, it is very probable that the scientific community will under-attribute robots, while they will actually be developing some phenomenal states and an ability to empathise (although this is highly unlikely in the next few decades). A few researchers argue that humans' 'anthropectomy' (Andrews & Huss, 2014) could even lead to the creation of a race of 'robotic slaves' for whose suffering we would be guilty (Prescott, 2017, p. 145). Yet, on the other hand, the unjustified over-attribution of phenomenal states to robots may have a negative impact on the development of robotics and its benefits for human welfare.

The relational perspective on the above question would be fundamentally different from the substantial or functional ones. The relational stance emphasises one fact: by developing robotics, we also develop ourselves, our society. All of us and our entire culture are changing and transforming as robotics develops. Thus, social robots are becoming an integral part of our augmented, relational social reality. We have to consider what their role in this reality is, and how they will play this role, as well as how their presence affects and re-shapes the whole net of emotional and behavioural interdependencies between the different social agents functioning in it.

Moreover, with robots acquiring the status of social agents, many difficulties arise. For example, is this status just 'partial', only covering the practical consequences of robots maintaining relations with people (co-shaping the social reality)? Or does it also have a symbolic or legal dimension?

In this context, also, many ethical questions arise (Damiano & Dumouchel, 2020, pp. 200–202). They appear with the issue of empathy too. Should we empathise with robots? And if the answer is "yes", then should we empathise with all of them, or just with some special types/models? What kind of behaviour towards robots should we allow? How will empathising with these agents change us? In what situations should robots empathise with us, and what actions should robots perform in response to the empathy they 'feel'? How should such a value chain be determined?

The initial answers to these questions indicate that we not only need to protect people from the possibility of robots being used to manipulate them (e.g.

producers using the ability of robots to arouse empathy and induce shopping), but also to protect robots from being abused, due to the fact that it may negatively modulate interpersonal relations (for example, people may become desensitized to suffering if they are allowed to act aggressively towards robots) (Darling, 2015, 2016; Malinowska, 2021). Therefore, legal solutions (preceded by a scientific and ethical analysis) should not only consider the issue of providing people with legal protection from robots (e.g. by taking into account issues related to the potential of robots to manipulate human emotions by evoking empathy in their users). They should also consider providing legal protection for robots, even if only for our own sake. We need to regulate these issues as soon as possible because everything indicates that robots—robots that can coordinate their behaviour and characteristics with us to create interaction in which empathy emerges—are already here.

Finally, using the concept of empathy to study human and robot interactions can be cognitively valuable, not only for understanding these interactions but also in the context of subsequent attempts to capture, describe and explain the general phenomenon of empathy. Researching human responses to robotic behaviour, which we can at the moment control and design quite precisely, may advance our understanding of the subject or even wholly redefine it. The possibility of such a change is provided for instance by the relational approach to emotions, which not only changes our understanding of the social environment and assigns a more active status to the non-human beings co-shaping it but also completely transposes our understanding of how our effects emerge in this complex and processual network of mutual relations.

Acknowledgements This research was funded by the Ministry of Science and Higher Education in Poland, National Science Centre, No. DEC-2015/19/D/HS1/00965. The article is based on the conference presentation: "The growing need for reliable conceptual analysis in HRI studies: the example of the term 'empathy'" presented at *Robophilosophy* conference in 2020) which was printed in a form of conference proceedings (Malinowska, 2020). However, it is significantly developed and expanded version of the text. I am grateful to the reviewers for their insightful comments.

Funding This research was funded by the Ministry of Science and Higher Education in Poland, National Science Centre, No. DEC-2015/19/D/HS1/00965.

Data Availability Not applicable.

Code Availability Not applicable.

Declarations

Conflict of interest The author declare that she has no conflict of interest.

References

Adamo, S. A. (2016). Do insects feel pain? A question at the intersection of animal behaviour, philosophy and robotics. *Animal Behaviour, 118,* 75–79.

Andrews, K., & Huss, B. (2014). Anthropomorphism, anthropectomy, and the null hypothesis. *Biology & Philosophy, 29*(5), 711–729.

Cañamero, L. (2005). Emotion understanding from the perspective of autonomous robots research. *Neural Networks, 18*(4), 445–455.

Castellano, G., Paiva, A., Kappas, A., Aylett, R., Hastie, H., Barendregt, W., & Bull, S. (2013). Towards empathic virtual and robotic tutors. *International Conference on Artificial Intelligence in Education* (pp. 733–736). Springer.

Coeckelbergh, M. (2010a). Artificial companions: Empathy and vulnerability mirroring in human–robot relations. *Studies in Ethics, Law, and Technology*, *4*(3). https://doi.org/10.2202/1941-6008.1126.

Coeckelbergh, M. (2010b). Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology, 12*(3), 209–221.

Coeckelbergh, M. (2014). The moral standing of machines: Towards a relational and non-Cartesian moral hermeneutics. *Philosophy & Technology, 27*(1), 61–77.

Coeckelbergh, M. (2018). Technology and the good society: A polemical essay on social ontology, political principles, and responsibility for technology. *Technology in Society, 52,* 4–9.

Coplan, A. (2011). Understanding empathy: Its features and effects. In: A. Coplan & P. Goldie (Eds.), *Empathy: Philosophical and psychological perspectives* (pp. 5–18). Oxford University Press.

De Vignemont, F., & Singer, T. (2006). The empathetic brain: How, when, and why? *Trends in Cognitive Sciences, 10,* 435–441.

Damasio, A. R. (1999). *The feeling of what happens: Body and emotion in the making of consciousness*. Houghton Mifflin Harcourt.

Damiano, L., & Dumouchel, P. (2017). *Living with robots*. Harvard University Press.

Damiano, L., & Dumouchel, P. (2018). Anthropomorphism in human–robot co-evolution. *Frontiers in Psychology, 9,* 468.

Damiano, L., & Dumouchel, P. (2020). Emotions in relation. Epistemological and ethical scaffolding for mixed human–robot social ecologies. *HUMANA. MENTE Journal of Philosophical Studies, 13*(37), 181–206.

Damiano, L., Dumouchel, P., & Lehmann, H. (2014). Towards human–robot affective co-evolution. *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-014-0258-7.

Darling, K. (2015). 'Who's Johnny? 'Anthropomorphic framing in human–robot interaction, integration, and policy. *Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy. ROBOT ETHICS*, *2*.

Darling, K. (2016). Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior towards robotic objects. In R. Calo, A. M. Froomkin, & I. Kerr (Eds.), *Robot law*. Edward Elgar.

Darling K., Nandy, P., & Breazeal, C. (2015). Empathic concern and the effect of stories in human–robot interaction. *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN),* pp. 770–775.

Fink, J. (2012). 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.), *Social robotics. ICSR 2012. Lecture notes in computer science* (Vol. 7621). Springer.

Fung, P. (2015). Robots with heart. *Scientific American, 313*(5), 60–63.

Gazzola, V., Rizzolatti, G., Wicker, B., & Keysers, C. (2007). The anthropomorphic brain: The mirror neuron system responds to human and robotic actions. *NeuroImage, 35*(4), 1674–1684.

Glaskin, K. (2012). Empathy and the robot: A neuroanthropological analysis. *Annals of Anthropological Practice, 36*(1), 68–87.

Goldman, A. (2011). Two routes to empathy. In: A. Coplan & P. Goldie (Eds.), *Empathy: Philosophical and psychological perspectives* (pp. 31–34). Oxford University Press.

Gunkel, D. J. (2015). The rights of machines: Caring for robotic care-givers. In: S. van Rysewyk & M. Pontier (Eds.), *Machine medical ethics* (pp. 151–166). Cham: Springer.

Gunkel, D. J. (2018). The other question: Can and should robots have rights? *Ethics and Information Technology, 20*(2), 87–99.

Gunkel, D. J. (2020). Mind the gap: Responsible robotics and the problem of responsibility. *Ethics and Information Technology*, *22*, 307–320. https://doi.org/10.1007/s10676-017-9428-2.

Guzzi J., Giusti A., Gambardella, L. M., & Di Caro, G. A. (2018). A model of artificial emotions for behavior-modulation and implicit coordination in multi-robot systems. In *Proceedings of the Genetic and Evolutionary Computation Conference* (pp. 21–28).

Hoffman, M. L. (2001). *Empathy and moral development: Implications for caring and justice*. Cambridge University Press.

Kirby, R., Forlizzi, J., & Simmons, R. (2010). Affective social robots. *Robotics and Autonomous Systems, 58*(3), 322–332.

Kozima, H., Nakagawa, C., & Yano, H. (2004). Can a robot empathize with people? *Artificial Life and Robotics, 8*(1), 83–88.

Leite, I., Pereira, A., Mascarenhas, S., Martinho, C., Prada, R., & Paiva, A. (2013). The influence of empathy in human–robot relations. *International Journal of Human-Computer Studies, 71*(3), 250–260.

Leite, I., Castellano, G., Pereira, A., Martinho, C., & Paiva, A. (2014). Empathic robots for long-term interaction. *International Journal of Social Robotics, 6*(3), 329–341.

Lyon, C., Nehaniv, C. L., & Saunders, J. (2012). Interactive language learning by robots: The transition from babbling to word forms. *PLoS One, 7*(6), e38236.

Malinowska, J. K. (2020). The growing need for reliable conceptual analysis in HRI studies: The example of the term 'Empathy'. In *Frontiers in artificial intelligence and applications*, Volume 335: Culturally sustainable social robotics (pp. 96–104).

Malinowska, J. K. (2021). Can I feel your pain? The biological and socio-cognitive factors shaping people's empathy with social robots. *Unpublished manuscript*.

Metzinger, T. (2009). *The Ego Tunnel: The science of the mind and the myth of the self*. Basic Books.

Misselhorn, C. (2009). Empathy with inanimate objects and the uncanny valley. *Minds and Machines, 19*(3), 345.

Nagata, S., Sekiguchi, M., & Asakawa, K. (1990). Mobile robot control by a structured hierarchical neural network. *IEEE Control Systems Magazine, 10*(3), 69–76.

Nehaniv, C. L., & Dautenhahn, K. E. (2007). *Imitation and social learning in robots, humans and animals: Behavioural, social and communicative dimensions*. Cambridge University Press.

Nickelsen, N. C. M. (2019). 'Active Citizenship' and feeding assistive robotics. In: C. Hasse & D. M. Søndergaard (Eds.), *Designing robots, designing humans* (p. 73). Routledge.

Niculescu, A., van Dijk, B., Nijholt, A., Li, H., & See, S. L. (2013). Making social robots more attractive: The effects of voice pitch, humor and empathy. *International Journal of Social Robotics, 5*(2), 171–191.

Nowak, K. L., & Biocca, F. (2003). The effect of the agency and anthropomorphism on users' sense of telepresence, copresence, and social presence in virtual environments. *Presence: Teleoperators & Virtual Environments, 12*(5), 481–494.

Picard, R. (1997). *Affective computing*. MIT Press.

Prescott, T. J. (2017). Robots are not just tools. *Connection Science, 29*(2), 142–149.

Redmond, M. V. (1989). The functions of empathy (decentering) in human relations. *Human Relations, 42*(7), 593–605.

Redstone, J. (2014). Making sense of empathy with social robots. In: J. Seibt, et al. (Eds.), *Robophilosophy* (pp. 171–177). IOS Press.

Redstone, J. (2017). Making sense of empathy with sociable robots: A new look at the "imaginative perception of emotion". In: M. Nørskov (Ed.), *Social robots* (pp. 19–38). Routledge.

Riek, L. D., & Robinson, P. (2008). Real-time empathy: Facial mimicry on a robot. In *Workshop on Affective Interaction in Natural Environments (AFFINE) at the International ACM Conference on Multimodal Interfaces (ICMI 08). ACM.*

Riek, L. D., Rabinowitch, T. C., Chakrabarti, B., & Robinson, P. (2009a). Empathizing with robots: Fellow feeling along the anthropomorphic spectrum. *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops* (pp. 1–6). IEEE.

Riek, L. D., Rabinowitch, T. C., Chakrabarti, B., & Robinson, P. (2009b). How anthropomorphism affects empathy toward robots. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction* (pp. 245–246). ACM.

Rosenthal-von der Pütten, A. M., Krämer, N. C., Hoffmann, L., Sobieraj, S., & Eimler, S. C. (2013). An experimental study on emotional reactions towards a robot. *International Journal of Social Robotics, 5*(1), 17–34.

Rosenthal-Von Der Pütten, A. M., Schulte, F. P., Eimler, S. C., Sobieraj, S., Hoffmann, L., Maderwald, S., Brand, M., & Krämer, N. C. (2014). Investigations on empathy towards humans and robots using fMRI. *Computers in Human Behavior, 33,* 201–212.

Salichs, M. A., & Malfaz, M. (2011). A new approach to modeling emotions and their use on a decision-making system for artificial agents. *IEEE Transactions on Affective Computing, 3*(1), 56–68.

Scheutz, M., & Arnold, T. (2016). Are we ready for sex robots?. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction* (pp. 351–358). IEEE Press.

Scheutz, M., & Arnold, T. (2017). Intimacy, bonding, and sex robots: Examining empirical results and exploring ethical ramifications. *Unpublished manuscript*.

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* (pp. 11–39). Springer.

Seo, S. H., Geiskkovitch, D., Nakane, M., King, C., & Young, J. E. (2015). Poor thing! Would you feel sorry for a simulated robot? A comparison of empathy toward a physical and a simulated robot. *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 125–132). IEEE.

Singer, T., & Lamm, C. (2009). The social neuroscience of empathy. *Annals of the New York Academy of Sciences, 1156*(1), 81–96.

Singer, T., & Klimecki, O. M. (2014). Empathy and compassion. *Current Biology, 24*(18), R875–R878.

Sparks, A., McDonald, S., Lino, B., O'Donnell, M., & Green, M. J. (2010). Social cognition, empathy and functional outcome in schizophrenia. *Schizophrenia Research, 122*(1–3), 172–178.

Stephan, W. G., & Finlay, K. (1999). The role of empathy in improving intergroup relations. *Journal of Social Issues, 55*(4), 729–743.

Stueber, K. (2019). Empathy. *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/fall2019/entries/empathy/. Dostęp dnia: 15.08.2019.

Tapus, A., & Mataric, M. J. (2007). Emulating empathy in socially assistive robotics. *AAAI spring symposium: Multidisciplinary collaboration for socially assistive robotics*.

Wermter, S., Weber, C., Elshaw, M., Panchev, C., Erwin, H., & Pulvermüller, F. (2004). Towards multimodal neural robot learning. *Robotics and Autonomous Systems, 47*(2–3), 171–175.

Williams, M. A. (2012). Robot social intelligence. In *International Conference on Social Robotics* (pp. 45–55). Springer.