Published in "Review of Philosophy and Psychology 11, pages 881–897 (2020)" which should be cited to refer to this work.

Understanding A.I. — Can and Should we Empathize with Robots?

Susanne Schmetkamp [1]

Published online: 28 April 2020

Abstract
Expanding the debate about empathy with human beings, animals, or fictional characters to include human-robot relationships, this paper proposes two different perspectives from which to assess the scope and limits of empathy with robots: the first is epistemological, while the second is normative. The epistemological approach helps us to clarify whether we *can* empathize with artificial intelligence or, more precisely, with social robots. The main puzzle here concerns, among other things, exactly *what it is* that we empathize with if robots do not have emotions or beliefs, since they do not have a consciousness in an elaborate sense. However, by comparing robots with fictional characters, the paper shows that we can still empathize with robots and that many of the existing accounts of empathy and mindreading are compatible with such a view. By so doing, the paper focuses on the significance of perspective-taking and claims that we also ascribe to robots something like a perspectival experience. The normative approach examines the moral impact of empathizing with robots. In this regard, the paper critically discusses three possible responses: strategic, anti-barbarizational, and pragmatist. The latter position is defended by stressing that we are increasingly compelled to interact with robots in a shared world and that to take robots into our moral consideration should be seen as an integral part of our self- and other-understanding.

Keywords Empathy · Artificial intelligence · Humanoid robots · Interaction · Perspective-taking · Fictional characters · Ethics

1 Introduction

Debates about empathy or, more broadly, interpersonal understanding have been a mainstay of scholarship across a broad range of disciplines. However, while much has been written on the human capacity to empathize with real people or fictional characters

✉ Susanne Schmetkamp
Susanne.schmetkamp@icloud.com

[1] Department of Philosophy, University of Fribourg, Fribourg, Switzerland

(for recent overviews, see Coplan and Goldie 2011; Maibom 2017), until recently philosophers have somewhat neglected the role of empathy in human-robot interactions (HRI) (cf. Brinck and Balkenius 2018; Lin et al. 2017). Yet, in line with the growing number of studies on the emotions or other features of artificial intelligence systems,[1] there has been much philosophical interest in the possibility and necessity of interacting and empathizing with different forms of artificial intelligence, especially with so-called social robots.[2] This interest has also given rise to discussions on the value of empathy for society in general or for health care and therapy in particular (Coeckelbergh 2018; Darling 2016; Engelen 2018; Loh 2019; Misselhorn in press; Vallor 2011). It is becoming clear that, in the future, robots and androids – that is, robots that look like humans – will become more or less independent actors with social skills. As such, they are set to become important companions and increasingly capable of establishing relationships with human beings (Benford and Malartre 2007; Breazeal 2002; Dumouchel and Damiano 2017). In addition, *Deep Learning Systems* (Kasparov 2017) will be deployed in many (so far) human professions, which will not only improve or facilitate some tasks or challenges (in medical research, for instance); they might also force us to reconsider some key concepts such as intelligence, agency, consciousness, autonomy, emotions, or perspectives (Schneider in press).

As studies have shown (Leite et al. 2013), the form and success of human-robot relationships often depend on humanlike features – such as the robots' capacity to express emotions, to interact, and to execute (more or less) autonomous decisions. These capacities are also important for reciprocal empathic understanding.[3] While human beings also recognize and ascribe emotions in relation to abstract virtual forms or even in view of technical devices (the best examples being smartphones and computers), for our cooperative and collaborative interaction with robots – particularly in the medical or health care context – a strong human likeness might be crucial in these interactions succeeding. As the presence of humanoid social robots in society grows, so too does the need to examine and shape our interactions with them. The current generation of robots is already able to express a range of emotions – the humanoid A.I. "Sophia", for instance, knows 60 different facial expressions and even seems capable of communicating with a sense of humor and irony. However, robots do not have a consciousness in the sense of subjective experience,[4] and they do not possess humor or emotions in an elaborate sense (Boden 2016; MacLennan 2014; Scheutz 2011). Yet, they might have something that can be seen as analogous to human emotions and some mental processes. Moreover, in light of recent insights from the philosophy of embodied cognition, it might be the case that the humanlike body and behavior and the "extended" cognition are what help humans to recognize androids as partners and as similar to them in some respects, while remaining totally distinct from

[1] A group at the MIT Media Lab and the IEEE standards association argues for the concept of "extended" intelligence instead of "artificial". By means of such a new narrative of "extended", they want to guarantee that robots do not substitute but rather support human beings and cooperate with them. Together they established the Council on Extended Intelligence CXI, see https://globalcxi.org (last accessed 12.12.2019).

[2] An ERC-funded project located at the University of Glasgow and headed by Emily Cross examines particularly the socializing of human beings with artificial intelligence and the importance of interaction and relationships with robots for social cognition. One focus lies on the ability of robots to be companions, http://www.so-bots.com (last access 20.12.2019).

[3] Concerning the phenomenon of the "uncanny valley" see below.

[4] At least when we follow an anti-physicalist position.

them in others (Benford and Malartre 2007, 181; Hoffmann and Pfeifer 2018; Newen et al. 2018).[5]

Empathy is broadly seen as a crucial way of apprehending and re-experiencing others' mental states by mindreading, emotional sharing, and/or experiential co-experience (see e.g. Engelen and Röttger-Rössler 2012; Goldman 2006; Stueber 2018; Zahavi 2014).[6] In philosophy, empathy is usually distinguished from affective contagion and moral sympathy or compassion.[7] Whereas the latter aims at the well-being of others and wants to promote (or at least not to impede) it (Darwall 1998), empathy, in the first instance, leads to the comprehension of others' mental processes – such as emotions or beliefs. In contrast to mere emotional contagion, a self-other differentiation must be in place (De Vignemont and Jacob 2012). There continues to be significant debate on this point and a variety of definitions and approaches have been put forward which seek to address questions such as: How do we *perceive* and *access* others' states and experiences? How is the empathic *process* to be characterized? What is the *outcome* of this process? Broadly speaking, the predominant theories – coming from philosophy of mind or phenomenology – are the *Mirror Neurons* or *Resonance Theory* (MNT) (Gallese 2001), the *Theory Theory* (TT) (Fodor 1987; Gopnik and Wellman 1994), *Simulation Theory* (ST) (De Vignemont and Jacob 2012; Goldman 2006, 2011; Stueber 2006), *Direct Perception Theory* (DPT) (Zahavi 2011) with its variations of *Interaction Theory* (IT) (Gallagher 2008, 2017) and *Narrativity Theory* (NT) (Gallagher and Hutto 2008). In addition, there are hybrid and pluralist theories which combine two or more approaches, such as direct perception and imagination[8] (Schmetkamp 2017, 2019; Dullstein 2013; for excellent overviews, see Newen 2015; Stueber 2018; Zahavi 2014; 2018). Given this diverse set of approaches, some make further distinctions between cognitive empathy (such as TT) or affective empathy (such as ST or MNT).[9]

By asking whether we *can* empathize with robots at all, section 2 will focus on the many epistemic dimensions of empathic interrelationships with robots: What do we perceive and understand if there are not *really* emotions, subjective experiences or

[5] The paper concentrates mainly on humanoid robots. One reason for this is that it helps to constrain the scope of the paper; another reason is the assumption that humanlike features indeed facilitate our social interaction with artificial intelligence and make it more plausible that we treat robots as social partners. However, we can also empathize with more abstract forms of A.I. by ascribing to them emotional states and motives (see Isik, Koldeewyn, Beeler and Kanwisher 2017). I am very thankful to one reviewer for this remark.

[6] It is very controversial whether empathy presupposes or implies affective mirroring, theoretical mindreading, simulative perspective-taking, emotional understanding and/or experiential comprehension, and there is currently no end to this debate in sight (see e.g. Zahavi 2018). Many philosophers stress that mindreading is something distinct from empathy, and that empathy is "something extra". Here, however, I have tried to apply all the different approaches. My own position is a phenomenological one, though.

[7] One problem of the whole debate is, though, that there is no conceptual consensus what empathy is and implies. The ERC-funded project on social robots, for instance, defines empathy as involving both emotional matching and prosocial behavior. In philosophy, though, empathy is usually not seen as a moral emotion or attitude (see Cross et al. 2018; Zahavi 2018).

[8] For instance, by referring to the classical positions of Stein or Dilthey and combining direct perception with imaginative re-presentation ("Vergegenwärtigung") (see also Gallagher 2019).

[9] Kanske (2018) distinguishes between affective empathy proper and cognitive theory of mind. Whereas the first capacity would enable us to feel what others feel, the other would help us to understand what others think or believe. Although I recognize the differences, I will not distinguish mentalizing from empathizing here, but will examine different forms of understanding other minds under the umbrella term of empathy since this is the central term in the current philosophical debate.

perspectives in a rich sense? Or do robots have something similar to emotions, beliefs, and experiences? Do they have an individual view on the world (Schmetkamp 2017) or a narrative (Gallagher 2012), since at least they are embodied and contextualized? By comparing robots with fictional characters, the answer will be affirmative: Yes, to a certain degree we can empathize with robots in a cognitive, affective, and even experiential way, by either inferring, feeling, interacting or imagining how they perceive and move through their world, just as we comprehend in plural ways (Vaage 2010) how a fictional character (e.g. in a movie) perceives her world, acts, and feels. The crucial aspect here will be that we ascribe an individual perspective to the other. We comprehend it independently of whether this perspective is only narrated, projected or programmed.[10]

The second question, which will be discussed in section 3, asks whether we *should* empathize with robots. There are two sides to this question: we can either ask whether empathy with robots has a mere *strategic* function with regard to the enhancement of the reciprocal understanding within the human-robot interrelationship, or we can ask whether empathy has an *ethical* impact such that we have a duty to empathize with robots (for an overview on the topic of ethics and A.I., see Boddington et al. 2017). If, for instance, we can epistemically understand what robots perceive, intend, or might even "feel", we are also able to predict what they will do next. In general, this might be helpful in terms of our interactions with them.[11] Obviously, this refers to a strategic or rational "should". The second meaning of the question gives rise to a normative answer: Do we *owe* empathy to others in a moral sense? And what, from a moral standpoint, do they or we – as empathizers – gain from this? Considering this question, at first glance a Kantian answer might be obvious, which follows the precedent set by Kant's view on animals and which can be modified to apply to artificial intelligence: namely we should empathize, so the argument goes, in order to avoid "moral barbarization". In the end, the paper will take neither the strategic nor the Kantian path, and instead propose a pragmatist and relational answer. This answer is related to the other two. However, it stresses the impact of interaction and of the self-other-understanding.

2 Can we Empathize with Robots?

For reasons of space, I will concentrate on robots that have both a face and a body, show humanlike expressions and behavior, are intended to interact with human beings, and are therefore embodied and embedded in our everyday life and as such are subject to social appraisal by humans. A second reason for this focus is the assumption that robots with humanlike features and expressions are probably even more capable of building confidence and eliciting emotional responses similar to those of real humans (Brinck and Balkenius 2018; Mori 2005) and are in this regard more likely to be recognized and accepted as partners in social interaction. Although studies in cognitive psychology have shown that we can also empathize or mindread with systems which have little physical resemblance (Bretan et al. 2015), a humanlike appearance is

[10] The paper focuses on the epistemological question. It will not answer the metaphysical question whether robots or A.I. *have* a consciousness.

[11] Concerning the deployment of *Deep Learning Systems* in medicine, among other features, it is necessary to trust the intelligent machine and to understand what it is going to do, for instance in a medical robot-patient interaction.

important for the use of robots as caregivers or colleagues in healthcare (Vallor 2011).[12] But what kind of empathy is at stake here? Do we mirror the robots' expressions? Do we interpret and predict their behavior? Or do we empathize in a more phenomenological, interactive way?

Very minimally, empathy can be defined as the human capacity to comprehend others' mental states and to re-experience them in one way or another, although it remains a matter of debate whether the empathic subject needs to feel the same as the other. Some theories restrict the objects of empathy to persons' emotions and their expressions as indicating affective states. Others are broader and include other cognitive processes as objects of empathy – such as beliefs, desires, and their respective reasons (for overviews, see Batson 2009; Slote 2017). A prominent definition implies an isomorphism condition: empathizer and target are in the same or at least a similar affective state (De Vignemont and Singer 2006). However, as some critics have argued, empathy does not necessarily imply that we replicate others' mental states (Zahavi and Michael 2018). Nor do we have to care about the other in a more elaborate sense.

As is widely known, the current "hype" surrounding the topic of empathy can be largely attributed to the discovery of the so-called "mirror neurons" (Iacoboni et al. 1999; 2011). Broadly speaking, mirror neurons are those neurons located in an area of the brain that are discharged for both the observation and execution of similar actions. This imitation process has been applied to the understanding of human emotions: when observing another person's affective expression – such as a sad face – the same neurons would be activated as if we – as observers – had made a sad face and felt sadness ourselves. Whereas this theory has been widely criticized (Hickok 2014) and rejected as a theory of *empathy*, others have invoked it in their more elaborate approach to empathy. In his account of *Simulation Theory* (ST), Alvin Goldman, for instance, distinguishes makes a distinction between a low-level and a high-level form of mindreading or a "mirror route" and a "reconstructive route", though emotional "resonance" is implemented via both routes (Goldman 2006, 2011). Mirror neurons are the main part of the low-level processes through which we comprehend another person's mental states immediately and automatically. On a more complex, higher level, we simulate the other's state in our own mind and then arrive at the knowledge of how the other feels, not by deploying a theory, but rather by imitating the others' behavior in our mind and then projecting our own mental process onto the other. According to ST, we simulate, via a first-person perspective, being in the other's situation and utilize our own mental mechanisms to generate thoughts, beliefs, desires, and emotions. For the past couple of decades, ST – alongside its opponent the *Theory Theory* (TT) – has dominated the debate on mindreading. TT claims that our understanding of other minds essentially relies on folk psychology, which is either inborn or acquired during early childhood (Baron-Cohen 1995). TT assumes that we make theory-based inferences in order to understand others.[13] From a third-personal, observational standpoint, we deploy (implicitly or explicitly) law-like generalizations, which imply concepts of mental states such as perception, belief, and desire. TT has been criticized for being overly theoretical and too general (Zahavi 2014; but cf. Fodor 1987). Its detractors claim that TT does not take the concrete other into account, nor does it

[12] However, empirically it remains uncertain whether robots indeed must be humanlike in HRI (Brinck and Balkenius 2018).

[13] One problem, of course, is how we understand the term "understanding". Monika Dullstein (2012) has shown that *Theory of Mind* accounts use quite a different notion from phenomenological accounts.

recognize the embodiment and embeddedness of others. Furthermore, both TT and ST are seen to be taken in by a false Cartesian occlusionistic view of the mind, as if we cannot perceive what is going on in the mind of another (Zahavi 2011, 2014). By contrast, phenomenological accounts stress the embodiment and embeddedness of human beings and argue that we are able to see directly in the other's face and bodily expressions what she is experiencing: in this view, we do not have to infer or imagine what she is feeling; we only need to perceive it. Moreover, we do so in a shared situational context and through interaction. For this reason, such an approach is called *Direct Perception Theory* (DPT) (Zahavi 2011, 2014) or *Interaction Theory* (IT) (Gallagher 2001; 2012). In contrast to TT, DPT and IT argue that we do not adopt a third-person stance towards others and observe them. In addition, DPT and IT also argue that we do not have an imaginative indirect access to others. Instead, we socially interact in a second-personal way, whereby two "yous" recognize each other complementarily and reciprocally (Dullstein 2012; Engelen 2018; Zahavi and Michael 2018). DPT's limits obviously arise in situations whereby the other is not present to us: for instance, when someone tells us a story about someone else, or if we read a novel, watch a movie, or see a play where the experiences of others are in some way mediated by someone else (e.g. a narrator), we do not have direct encounters. Hence, all such instances are cases where the other is given by narration, sometimes even within a fictional framework. This is why some philosophers add that a *narrative* is essential in order to understand other minds or to elicit empathy in anything more than the most basic sense. Daniel D. Hutto (2008) formulated the *Narrative Practice Hypothesis* (NPH). According to this thesis, we understand others' reasons for acting, their beliefs and desires only when we also take into account individual circumstances, the subject's history, her current situation, her hopes and experiences, her character traits, and so on. In other words, according to the NPH, to grasp someone's situation, we have to rely on the person's "story" (Gallagher 2012). This view also allows for empathizing with "monsters or aliens from other planets, as portrayed in film" (Gallagher 2012). However, a kind of imagination is needed here: There are so many cases – not only but especially in our dealings with fiction – where we rely on our imagination as a way of making available something that is not present to us. Even one of the early pioneers of phenomenological approaches to empathy, namely Edith Stein (1989), claimed that imagination or "re-presentation"[14] plays a crucial role within a multi-stage process of empathic comprehension. This is why some theories of empathy combine a second-personal approach with a form of imaginative re-presentation of the concrete other's situation, narrative, and/or perspective (Schmetkamp 2019; Gallagher and Gallagher 2019).[15]

Regardless of whether we ought to consider all these different accounts as theories of empathy or more broadly as theories of interpersonal understanding, for every account we can ask the following from a descriptive and epistemological perspective: How do we

[14] It is difficult to give an exact translation of Stein's concept of "Vergegenwärtigung". The English translation (Stein 1989) uses "representation" or "representational act" (Stein 1989: 8) as a non-primordial represented "givenness" of others' or indirect experiences (analogous to memory, expectation, and fantasy) (ibid.). In the debate it is often overlooked that Stein proposes a step model of empathy, according to which the first level is direct perception of the other's experience, with the second level being a kind of reflection and perspective-taking (Stein 1989: 10).

[15] Gallagher recently defined empathy as follows: "Empathy might […] not only [count] as something that happens, but as a method; and that […] involves putting oneself into the other's perspective or situation" (2018). In so doing, Gallagher expanded his narrative approach into a perspectival approach (combining the narrative with the subjective perspective).

empathize with A.I., for instance humanlike robots, if the respective account were the most plausible one? For instance, if we observe a robot's expressions and/or actions, one might argue that we automatically resonate and imitate the expressive behavior. If we want to predict what the robot will do next, we might also deploy a folk-psychological theory and infer their reasons for acting. We might simulate what we would do if we were in their situation and then project our experience on them. Or, in direct encounters, we might be able to interactively perceive their actions. We might consider their embeddedness within a narrative context and comprehend their emotions' intentional structure without at the same time replicating their "qualitative" content. Empirically speaking, these interactive ways of understanding certainly occur.

However, some obvious metaphysical and epistemological objections can be raised. The main problem is that robots do not actually feel or experience anything. Nor do they really have mental states such as desires or beliefs, for they have no consciousness. That said, it also seems odd to speak of a robot's individual perspective or personal narrative. Insofar as empathy is directed towards mental states and someone's "being in the world", the answer would be: we *cannot* empathize with robots.

Yet, two possible responses could be given: First, robots' "mental states" are often described as "computational states" which are considered to have a structure that is analogous to human mental states. So, if we assume that robots have something which is comparable to human mental states, do they also have something like emotions or experiences with which we empathize? According to some current philosophical accounts of emotions, emotional states or processes exhibit a complex structure consisting of cognitive and affective components (De Sousa 1987; Nussbaum 2001): when we feel anger, our anger is intended towards an object which we evaluate as being annoying. In psychology this is also called *appraisal theory*, which implies that we make judgments about objects in our environments with regards to their relevance to our goals. If emotions only consisted of this mere cognitive component, we could assume that robots have emotions in a minimal sense. Robots, we could argue, act upon a set of reasons which are based upon a set of beliefs about the world. However, emotions might comprise more than that: anger, for instance, is also *felt* on a sensational and bodily level; it feels, for instance, frustrating and narrowing. That said, anger also has negative connotations, which we become aware of proprioceptively (Colombetti 2013). Robots' bodies, though – if they are not purely virtual – are comprised of metal or plastic, and, more importantly, they are not related to a rich concept of consciousness: in the sense that it experiences itself as an emotional being. It cannot self-referentially feel what it is like to be in a plastic body. Furthermore, as narrative emotion accounts have argued, complex emotions are usually embedded within a narrative framework: we can tell a story about their arousal and development (Goldie 2000). And last but not least, human beings are able to creatively deal with their feelings and emotions: they can learn new emotions and they are able to modify some and cultivate others.

Yet, this might also be possible for and with robots. The crucial point here is that we, perfectly intuitively, also *ascribe* emotions to machines. When collaborating with robots, we might take on the "intentional stance". This concept, originating from the work of Daniel Dennett, implies that we treat an object whose behavior we want to predict as a rational agent; we ascribe beliefs and desires, and, on that basis, we predict its behavior (Dennett 1987). But still, this approach is based on a theory of mindreading and not on the theory of phenomenal interaction that phenomenologists have in mind. However, if we consider consciousness implying phenomenal experience, it seems difficult to apply other than

Theory of Mind accounts of empathy to the HRI. In other words, the problem concerning the compatibility of phenomenological theories for HRI seems to be the phenomenal aspect of mental states, particularly the *feeling* and *experiential* side of emotions. Whereas we could theorize (TT) about the cognitive components of, or simulate, a robot's decision situation (ST) and then infer or *project* from our conclusions to the robot's situation, it would be difficult to speak of an empathic comprehension of the robot's affective and sensational states in a non-projective way. If we expand the problem to the concept of "experience" – the central term of the phenomenological approach (DPT) – things become even more complicated. As described above, according to the DPT and its variations, in our social interaction with others we empathically perceive their experiences, and we do so from a reciprocal second-person standpoint. "Experience" is an elaborate phenomenological term and implies existential aspects and phenomenal qualities. We subjectively and consciously experience our world or what it is like to feel or to do something, for example perceiving a red table as red and what this redness feels like. DPT assumes that we experience others' phenomenal experiences directly and intersubjectively, though not by replicating the exact qualitative character of an experience, but rather by attending to the intentional structure of the other's perspective (Gallagher 2012; Zahavi and Michael 2018). In order for this process to function, intercorporeal and face-to-face interaction is important.[16] Now, whereas the latter is (at least very basically) guaranteed when we cooperate and collaborate with robots, some crucial criteria of this intersubjective relationship are missing: just as robots do not *feel* emotions, so they do not have a subjective experience with their phenomenal content and existential impact.[17] DPT presupposes, though, that by perceiving someone's affective state in their facial or bodily expressions, we thereby also *re-experience* what it is like for them. We do not have to employ theoretical inferences, imitations, or projections. We experience that the other has phenomenal experiences. That said, from a phenomenological perspective, it seems difficult to empathize with robots. Yet, by comparing artificial intelligence with fictional characters, I will propose a potential solution and also demonstrate that not only do we mindread or mirror robots' behavior, but that it is possible, at least to a certain degree, to apply a phenomenological approach, that is, to interactively empathize with the robots' *perspectival* "experience". And the argument even goes beyond this analogy: When we interact with robots in a shared environment, we develop a shared intentionality and even a joint history, and this is crucial for our relationship with robots (Coeckelberg 2018). However, similar to our empathic understanding of fictional characters, our capacity of imagination is crucial here.

Let's play out the analogy: It is commonly assumed that empathy plays an essential role in our dealings with fictional narratives and fictional characters – be it in a novel, movie or play. Since the 1990s, there has been considerable debate within the philosophy of literature and film as to whether "empathy" ought to be subsumed under the umbrella term of "emotional engagement" with fictional characters in general (e.g.

[16] The narrativistic version of phenomenological approaches, though, implies an imaginative component which enables us to comprehend the intentional structure by narrative imagination, e.g. if an intersubjective interaction is not given (Gallagher and Gallagher 2019).

[17] It is a similar question to that in the so-called "zombie thought experiment", which discusses whether we can assume or *ascribe* a consciousness in the case of zombies – which are like us in all physical respects but have no conscious experiences in a rich sense (Chalmers 1996; Dennett 1991).

Plantinga 2009; Smith 1995). Other forms of engagement include emotional contagion and emotional sharing – especially with respect to the moody effects of a fiction – moral sympathy or compassion, negative emotions such as antipathy, and synesthetic affects (Plantinga 2009; Schmetkamp 2017). As many film scholars have noted, empathy plays a crucial epistemic role in enabling the viewer to follow the narrative and remain attached to the characters (Smith 1995).[18] Leaving aside the other complex debate concerning the so-called "paradox of fiction" – which discusses whether we can feel real emotions towards fictional entities and whether these emotions are rational (Yanal 1999) – and presuming that we really feel and have to feel empathy towards fictional characters, we still have to explain how best to conceptualize empathy in the case of fiction. While I am generally convinced that we deploy different forms of empathizing, mindreading, and understanding – that is, the full spectrum of comprehension of others' mental states – when watching a movie or reading a novel, my assumption is that one aspect is particularly vital: Fictional characters express and represent certain individual perspectives on their (fictional) world. These perspectives are narrated in the diegetic world of the movie or novel; moreover, they are often additionally framed by an implicit or explicit narrator. They are embedded within a plausible narrative. Or, put differently: a narrative is a structured and shaped representation of events *from a certain perspective* (Goldie 2012: 8) and in fiction, the characters embody, express, and represent such embedded perspectives.

The importance of perspectives for fiction, and indeed for our empathic engagement with it, is in part due to the fact that a fiction usually (though not always) has different technical perspectives: a story is usually told from a first- or third-person perspective. But even more importantly, a perspective is a *worldview*. That said, a perspective means how a person is embedded in the world, how one perceives the world, how one experiences it. This perspective is shaped by and, in turn, shapes emotions, experiences, histories, memories; it is influenced by and itself influences character traits, judgments, and beliefs (Schmetkamp 2017). When, for instance, we are in a depressive mood, we see our world from a different – namely depressive or melancholic – point of view than if we are in a happy state.

We can now speak of fictional characters as "having" (or rather expressing and representing) a perspective insofar as they are focalized and narrated by a narrator which constructs and directs their worldview. As readers or viewers, we attend to them *as if* they have a perspective and we can imagine what it might be like to have such a perspective. Empathy with fictional characters involves a kind of other-centered perspective-taking without reducing this process to one of mere egocentric simulation or projection.[19] What is more, it is an advantage of fictional narratives that they impart others' perspectives in a condensed manner. Fictions afford us the opportunity to become immersed in perspectives which might be similar to or totally different from our own, and they often do so in an intense, condensed, and comprehensive way.

[18] Empathy as perspective-taking is indeed a capacity which enables viewers to comprehend the characters' narratives and perspectives. However, as a form of sensitive understanding as to why the character is feeling, thinking, and acting as she does, it is also an outcome. Thus, that empathy is both a process and an outcome has been argued by Coplan (2011) and Goldie (2000).

[19] Misselhorn made a similar argument by noting that "in seeing the T-ing of an inanimate object we imagine perceiving a human T-ing" (2009: 353).

By comparing robots with fictional characters, one central congruent characteristic stands out: Both do not *really have* emotions or conscious beliefs, but they can express and represent them. And partly on this basis we, as recipients or empathizers, attribute humanlike mental states to them (Weber 2013). However, we also experience them as somehow embodied entities with which we interact. As the phenomenological film philosopher Vivian Sobchack has argued, film and its characters are not just projections; they have a body and a voice, and they allow quasi-intersubjective experiences between themselves and recipients (Sobchack 2004). They might even enable tactile impressions. This embodied characteristic is also true of robots, perhaps even more so.

Yet, there are some crucial differences. Firstly, in contrast to robots, fictional characters lack a capacity that is vital for every intersubjective account of empathy: namely reciprocal interaction. In our relationships with fictional characters, we must imagine the characters having the expressed emotions, experiences, and perspectives, but we do not reciprocally interact with them. In addition, fictional characters cannot veto whatever we attribute to them. By contrast, in our encounters with robots, there is at least an *existent* and *present* embodied and embedded, interacting entity with which we can develop a relationship. The robot is able to object to something – say, if I were a patient and unwilling to take my medicine, the robot could be charged with ensuring that I do so. Secondly, one might object that, unlike fictional characters, robots do not (yet) have an experiential perspective or individual narrative, as mentioned above. Fictions indeed offer a rich picture of how someone can perceive and evaluate her world; and through these narrative frames and practices we expand our horizon and learn new emotions or emotional nuances. However, fictional characters' emotions and experiences are also only narrated within a particular narrative frame; their development depends on both what a narrator has dramaturgically designed and how readers or viewers receive it against their own intellectual and experiential background. Fictional emotions and experiences have less flexibility and creativity than their human counterparts. That said, there is the question of whether fictional characters can still be contrasted with robots. Fictional characters do not really experience anything; similarly, robots do not have experiences in a rich, qualia-including sense. However, robots do at least perceive their environment, categorize, evaluate, and interact within it. They have a way of seeing and being in the world; they are embodied and contextualized. If we think of Thomas Nagel's famous anti-reductionistic example "What is it like to be a bat?" (Nagel 1974) we will never be able to comprehend other beings' experiential perspective entirely; a bat, or so his argument goes, has a totally different perceptive system which cannot be compared with human perception. Yet, scientists are permanently discovering new facts about non-human entities like fish or plants (Coeckelberg 2018: 148), and one argument here maintains that even if we might never know what it is like to be them, we can at least experience them and their perspective in our relation to them (ibid.; Gruen 2009).

If we try to compare the robot's perspective with our own, there are some similarities, and of course many differences too. But this is not a new phenomenon in our social cognition of other minds. Firstly, a robot literally (e.g. visually) perceives the world in a certain way (maybe humanlike, maybe not). Secondly, as an artificial *intelligence*, it also has a perspective in the sense that it perceives and evaluates the world around it, how it solves problems, etc. The robot's perspective is far from being a perspective in an elaborate sense, like that of human beings, but it is an epistemic and evaluative perspective: a robot knows something and makes judgments about the world. We can also state that it has a motivational perspective, for a robot acts on the

basis of its beliefs.[20] Even more importantly, robots are embedded in a context which we perceive or with which we interact. So, my answer to the question of whether we can empathize with robots is: yes. Moreover, all the existing accounts are more or less applicable to HRI. Of course, the next question we need to ask then is: *should* we?

3 Should we Empathize with Robots?

Given the preceding analysis, let us assume that we *can* empathize with humanoid robots in plural ways, that is, we can feel with, interact with or infer from their "beliefs", "emotions", "experiences", and "perspectives". But why *should* we empathize with them at all? In light of the growing use of robots in medicine, health and elderly care, for instance, it seems much more plausible for robots to empathize with patients than vice versa. They must somehow engage some sensibilities towards the patients' needs, while, in turn, human patients might need an empathic companion. That said, it seems as if the investigation so far has been primarily a theoretical test to reveal which of the different accounts of empathy are compatible with HRI. But is there also a reason that we, as humans, should also empathize with robots? This question is relevant, since the interrelationship between humans and robots is only successful and fruitful if *both* indeed interact with the other, and these interactions might presuppose – in one way or another – empathic engagement.

Three arguments could be given for this normative thesis:

1. A strategic argument;
2. An anti-barbarization argument;
3. A pragmatist, shared-community argument.

The first, strategic argument is not directly a morally relevant normative argument. It takes up the idea that in order to interact successfully, we must somehow be able to infer and understand what our interactive counterpart is up to. More precisely, we might want to empathize, take over a perspective or read another mind in order to better achieve our goals. Our interaction with robots and our empathy with them is in this sense only of *use* for something else; it is merely instrumental. The notion "should" refers to a hypothetical imperative. In this regard, robots are considered more as tools than as collaborators. In fact, they are not seen as moral agents or patients here, which have a moral status (Coeckelberg 2018).

More substantive and morally normative is the second argument of non-barbarization or cultivation. By not empathizing with others, so the argument goes, we risk becoming desensitized. In turn, empathy might cultivate prosocial behavior and improve our moral character. Before I explore the main problems with this thesis, I will explain two of its roots – namely a Kantian and an Aristotelian argument. The Kantian argument was originally made in respect of the human-animal relationship. It implies

[20] Again, similar arguments could be put forward for other A.I. forms of non-human agents, e.g. abstract virtual shapes. The focus of this paper is on humanoid robots with which human beings cooperate and collaborate. For this to be successful, human beings might ascribe to A.I. not only basic mental states, but also a perspective and a narrative. This might be important for collective intentionality and collective attention.

that we should not be cruel towards animals because this would damage or corrupt our moral character in general. According to this argument, animals are only indirect moral patients, without having a moral status of their own, since Kant ties one's moral status to the competence to act autonomously out of reasons and attributes this competence only to persons. The same argument would then hold for social robots which would not be per se moral addressees: this is because they might not have autonomy in an elaborate sense. However, by not empathizing with them, we would disrespect a crucial condition of humanity.[21] HRI-specialist Kate Darling is a contemporary proponent of this view: "The Kantian philosophical argument for preventing cruelty to animals is that our actions towards non-humans reflect our morality – if we treat animals in inhumane ways, we become inhumane persons. This logically extends to the treatment of robotic companions. […] It may also prevent desensitization towards actual living creatures and protect the empathy we have for each other" (Darling 2016: 19).

The Aristotelian argument moves in a similar direction. It implies that we can cultivate our emotions by perspective-taking, thereby defining perspective-taking as a distancing from one's own first-person standpoint, or by emotional sharing and thereby becoming acquainted with new emotions (Nussbaum 2011; Rorty 2001). Whereas the Kantian view stresses the problem of barbarization, the Aristotelian view stresses the ethical impact of cultivating something by empathizing: our emotions, moral perception, imagination, and power of judgment.

As I said, some problems arise here, which beset the Kantian view in particular: the first one is that the recognition of just an *indirect* status of non-persons or beings without "rationality" is unsatisfactory: it is counterintuitive, anthropocentric, and it excludes a lot more entities than non-humans (Gruen 2017). But does this also concern inanimate entities? Thus, the question remains: what is it that we harm when we use violence against robots that might not *feel* anything in an elaborate and subjective way? Do they have a concept of respect and dignity? Do they have moral claims? These complex questions will have to remain unanswered here, since they would require a dedicated investigation of their own. Another objection against the Kantian view is that the argument is based on a specific account of empathy as prosocial behavior. Not only does this imply an understanding of other minds, but it also involves the concern for the well-being of another entity. That is, the empathizer is not just interested in the other's experiences and "feels" into them; they are also motivated to alleviate the other's suffering or to promote her well-being. And if we were cruel to them and were to disrespect the robots' well-being – for instance, by beating or raping them (if we think of sex robots) – this would rebound on our behavior towards humans too. However, as noted previously, the ethical impact of concern or care is rather the impact of sympathy or compassion as a *sui generis* moral emotion, and as such it is distinct from empathy (Darwall 1998). As phenomenologists in particular have shown, empathy is not necessarily a positive attitude towards others, but can also lead to antisocial behavior. A sadistic person has to be empathic in this sense too, that is, they comprehend the other's suffering but do not want to alleviate it (Breithaupt 2019; Zahavi and Michael

[21] Kant writes: "If a man shoots his dog because the animal is no longer capable of service, he does not fail in his duty to the dog, for the dog cannot judge, but his act is inhuman and damages in himself that humanity which it is his duty to show towards mankind. If he is not to stifle his human feelings, he must practice kindness towards animals, for he who is cruel to animals becomes hard also in his dealings with men" (Kant 1997: 212).

2018).[22] In other words, a Kantian approach conflates some important conceptual differentiations, namely between empathy and compassion. Another objection could be raised here: empirically, it is not at all clear why someone who does not empathize with others becomes necessarily barbarized (Brinck and Balkenius 2018).

However, from a more optimistic angle, some argue that frequent empathic comprehension or perspective-taking can help us to learn how others might feel or think. The more we deploy empathy, the more we are able to get involved with others both in our everyday interactions and in more unusual encounters. Moreover, this might make us a more tolerant or more virtuous person. Again, this is argued with regard to fictional characters and narratives. Attending to others' perspectives and experiences is, as Richard Rorty famously claimed, of ethical value, since in so doing we abandon our egocentric perspective (Rorty 2001). But, of course, we could adopt this argument for HRI: Empathizing with robots would then improve our cooperative and collaborative interactions insofar as we would become more acquainted with them. This leads to the third argument which shares some features with both the strategic and the Kantian/Aristotelian approach, but stresses the interaction, relation, and social self-understanding of the empathizers.

This argument (which describes my own position) takes up Rorty's approach but modifies it to an even more pragmatist and relational thesis of social cognition and its preconditions. In contrast to the Kantian and Aristotelian approach, this view proceeds from an anti-anthropocentric standpoint and stresses the interactive relationship between human beings and robots. This position assumes that empathizing with others – in all its variations but especially in the phenomenological interactive tradition – can allow us to become acquainted with others' "being in the world" and thereby broaden our horizons, change our perspectives, and shape our social interactions and moral behavior towards non-human others.

In view of perspectives, my assumption here is that we can even speak of (future) robots and deep learning systems[23] as having a specific view on the world of their own. This view will be in some ways similar to and in other ways different from human perspectives. Science fiction movies such as Her (US 2013) have imagined what independent A.I. might become: superintelligent systems that far exceed the capabilities of human thinking. Empathizing with humanoid robots with which we increasingly interact – in the context of health care, for instance – might help us to prepare for future developments. For the time being, though, it is rather the case that insofar as we already share actions and environments with robots, and insofar as empathy and social cognition can improve our interactions with others, we can also assume that our interactions with robots will benefit from an empathetic standpoint – though not merely in an instrumental, strategic sense. This might also have a training effect, an argument that, as noted, has also been advanced in relation to fictional worlds. But the more important point is that such a view touches upon the question of how we want to understand ourselves: taking robots as social companions seriously should be implemented as a part of our self-understanding as both humans and members of democratic societies. How we interact with robots depends a lot on how we think of them: as tools with

[22] The phenomenon that empathizers can become even more cruel the more humanlike robots are is called the "uncanny valley" (see Misselhorn 2009; Mori 2005).

[23] Or as Susan Schneider calls them: "future minds" (in press).

which are supposed to interact from a mere instrumental point of view, or as partners which we should take seriously for their own sake. It is, thus, the relation and the shared *community* which come to the fore here. Such a position stresses the pragmatist and phenomenological impact of the interactions. This might also have implications for the status of the robots as moral agents and moral patients, as Mark Coeckelbergh argues: "The question of moral standing is always connected to the question who is part of the moral community and what moral games are already played" (Coeckelberg 2018: 149). Instead of a top-down implementation of morality, Coeckelbergh argues for a bottom-up perspective. By considering robots as companions in a relational context and by empathizing with their perspectival narrative, we develop a relationship with them which in turn has effects on how we see them morally (ibid.).[24] However, to discuss the moral status would go beyond the scope of this paper. As mentioned above, empathy is not in itself a moral emotion or attitude of caring. But it might sow the relevant seeds in this respect, since it provides the epistemological basis for an intersubjective morality. Moreover, it has a lot to do with our social and moral self-understanding: "[T]he way we deal with other entities, the way we experience them, what we say about them, the way we treat them, and so on, also says a lot about me and says a lot about us" (Coeckelberg 2018: 150). But instead of an anthropocentric view, this is rather a relational view that treats non-human entities as partners in interaction.

4 Conclusion

Artificial Intelligence in general and humanoid robots in particular will change our lives and maybe ourselves too. Philosophers have much to consider in terms of the epistemic, ethical, aesthetical, and political impacts of these new challenges. Empathy is just one of many topics which are being challenged by HRI. This paper has contributed to the necessary investigations that are already under way or those that are yet to come. I discussed the epistemic puzzle of whether we *can* empathize with robots, applying the dominant contemporary accounts of empathy to this domain. I then examined the normative question as to whether and why we should empathize with robots. The paper proposed a pragmatist viewpoint by demonstrating that a) indeed we *can* empathize with humanoid robots, not only on a basic level, but also, at least to a certain degree, on an imaginative perspective-taking level; moreover, it was shown that even from a phenomenological and intersubjective point of view, it is possible to speak of empathizing with robots which are embedded in our world, with which we interact and share a contextual narrative. The focus lied on empathy as a process of mutual interaction rather than as an outcome. However, the paper also argued that b) we *should* empathize with humanoid robots because in doing so we can acquire new knowledge of a very unfamiliar being-in-the-world, thereby broadening our horizons, training for future A.I. developments, and improving HRI in a shared social environment. This was deemed to be not only of instrumental value, but also valuable for our understanding of ourselves and our society in which robots and other forms of A.I. can be seen as companions.

[24] Coeckelbergh proposes a similar approach to mine but takes inspiration from Wittgenstein's concepts of a form of life and language-games. Yet, his paper lacks a clear definition of what he thinks empathy implies (e.g. whether empathy indeed involves caring for the other's well-being, as his paper seems to suggest).

References

Baron-Cohen, S. 1995. *Mindblindness. An essay on autism and theory of mind*. Cambridge, MA: MIT Press.

Batson, C.D. 2009. These things called empathy: Eight related but distinct phenomena. In *The social neuroscience of empathy*, ed. J. Decety and W. Ickes, 3–15. Cambridge, MA: MIT Press.

Benford, G., and E. Malartre. 2007. *Beyond human*. Tom Doherty Associates: *Living with robots and cyborgs.* New York.

Boddington, P., P. Millican, and M. Wooldridge. 2017. Minds and machines special issue: Ethics and artificial intelligence. *Minds and Machines* 27 (4): 569–574.

Boden, M.A. 2016. *AI. Its nature and future*. Oxford: Oxford University Press.

Breazeal, C.L. 2002. *Designing sociable robots*. Cambridge, MA: MIT Press.

Breithaupt, F. 2019. *The dark sides of empathy*. Ithaca: Cornell University Press.

Bretan, M., G. Hoffman, and G. Weinberg. 2015. Emotionally expressive dynamic physical behaviors in robots. *International Journal of Human-Computer Studies* 78: 1–16.

Brinck, I., and C. Balkenius. 2018. Mutual recognition in human-robot interaction: A deflationary account. *Philosophy and Technology*: 1–18. https://doi.org/10.1007/s13347-018-0339-x.

Chalmers, D.J. 1996. *The conscious mind*. Oxford: Oxford University Press.

Coeckelberg, M. 2018. Why care about robots? Empathy, moral standing, and the language of suffering. *Kairos. Journal of Philosophy & Science* 20: 141–158.

Colombetti, G. 2013. *The feeling body. Affective science meets the enactive mind*. Cambridge, MA: MIT Press.

Coplan, A. 2011. Understanding empathy, 3–18. Its features and effects. In Empathy. Philosophical and psychological perspectives. Oxford: Oxford University Press.

Coplan, A., and P. Goldie. 2011. *Empathy. Philosophical and psychological perspectives*. Oxford: Oxford University Press.

Cross, E.S, Riddoch, K.A., Pratts, J, Titone, S, Chaudhury, B, and Hortensius, R. 2018. A neurocognitive investigation of the impact of socialising with a robot on empathy for pain. Preprint. https://doi.org/10.1101/470534.

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

Darwall, S. 1998. Empathy, sympathy, care. *Philosophical Studies* 89: 261–282.

De Sousa, R. 1987. *The rationality of emotion*. Cambridge, MA: MIT Press.

De Vignemont, F., and P. Jacob. 2012. What is it like to feel another's pain? *Philosophy of Science* 79 (2): 295–316.

De Vignemont, F., and T. Singer. 2006. The empathic brain: How, when and why? Trends in cognitive sciences 10(10): 435–441.

Dennett, D. 1991. *Consciousness explained*. Boston: Little, Brown, and Co..

Dullstein, M. 2012. The second person in the theory of mind debate. *Review of Philosophy and Psychology* 3 (2): 231–248.

Dullstein, M. 2013. Direct perception and simulation: Stein's account of empathy. *Review of Philosophy and Psychology* 4: 333–350.

Dumouchel, P., and L. Damiano. 2017. *Living with robots*. Cambridge, MA: Harvard University Press.

Engelen, E.M. 2018. Can we share an us-feeling with a digital machine? Emotional sharing and the recognition of one as another. *Interdisciplinary Science Reviews* 43 (2): 125–135.

Engelen, E.M., and B. Röttger-Rössler. 2012. Current disciplinary and interdisciplinary debates on empathy. *Emotion Review* 4 (1): 3–8.

Fodor, J. 1987. *Psychosemantics. The problem of meaning in the philosophy of mind*. Cambridge, MA: MIT Press.

Gallagher, S. 2008. Direct perception in the interactive context. *Consciousness and Cognition* 17 (2): 535–543.

Gallagher, S. 2017. Empathy and theories of direct perception. In *The Routledge handbook of philosophy of empathy*, ed. H. Maibom, 158–168. New York: Routledge.

Gallagher, S., and J. Gallagher. 2019. Acting oneself as another: An actor's empathy for her character. *Topoi* (online first), https//doi.org/https://doi.org/10.1007/s11245-018-96247.

Gallagher, S., and D. Hutto. 2008. Understanding others through primary interaction and narrative practice. In *The shared mind: Perspectives on intersubjectivity*, ed. J. Zlatev, T. Racine, C. Sinha, and E. Itkonen, 17–38. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Gallese, V. 2001. The 'shared manifold' hypothesis: From mirror neurons to empathy. *Journal of Consciousness Studies* 8: 33–50.
Goldie, P. 2000. *The emotions*. Oxford: Oxford University Press.
Goldie, P. 2012. *The mess inside. Narrative, emotion, and the mind*. Oxford: Oxford University Press.
Goldman, A. 2006. *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. Oxford: Oxford University Press.
Goldman, A. 2011. Two routes to empathy: Insights from cognitive neuroscience. In *Empathy: Philosophical and psychological perspectives*, ed. A. Coplan and P. Goldie, 31–44. Oxford: Oxford University Press.
Gopnik, A., and H.M. Wellman. 1994. The theory theory. In *Mapping the mind: Domain specificity in cognition and culture*, ed. L.A. Hirschfeld and S.A. Gelman, 257–293. Cambridge: Cambridge University Press.
Gruen, L. 2009. Attending to nature: Empathetic engagement with the more than human world. *Ethics and the Environment* 14 (2): 23–38.
Gruen, L. 2017. The moral status of animals. In *The Stanford encyclopedia of philosophy* (Fall 2017 edition), ed. E. N. Zalta, https://plato.stanford.edu/archives/fall2017/entries/moral-animal/.
Hickok, G. 2014. *The myth of mirror neurons: The real neuroscience of communication and cognition*. New York: W. W. Norton & Company.
Hoffmann, M., and R. Pfeifer. 2018. Robots as powerful allies for the study of embodied cognition from the bottom up. In *The Oxford handbook of 4E cognition*, ed. A. Newen, L. de Bruin, and S. Gallagher. Oxford: Oxford University Press.
Hutto, D.D. 2008. The narrative practice hypothesis: Clarifications and implications. *Philosophical Explorations* 11 (3): 175–192.
Iacoboni, M. 2011. Within each other: Neural mechanisms for empathy in the primate brain. In *Empathy: Philosophical and psychological perspectives*, ed. A. Coplan and P. Goldie, 45–57. Oxford: Oxford University Press.
Iacoboni, M., R.P. Woods, et al. 1999. Cortical mechanisms of human imitation. *Science* 286: 2526–2528.
Kanske, P. 2018. The social mind: Disentangling affective and cognitive routes to understanding others. *Interdisciplinary Science Reviews* 43 (2): 115–124.
Kant, I. 1997. *Lectures on Ethics*, ed. and trans. P. Heath and J. B. Schneewind. Cambridge: Cambridge University press.
Kasparov, G. 2017. *Deep thinking: Where machine intelligence ends and human creativity begins*. New York: Public Affairs.
Leite, A., A. Pereira, S. Mascarenhas, C. Martinho, R. Prada, and A. Paiva. 2013. The influence of empathy in human-robot relations. *International Journal of Human-Computer Studies* 71 (3): 250–260.
Lin, P., R. Jenkins, and K. Abney. 2017. *Robot ethics 2.0: From autonomous cars to artificial intelligence*. Oxford: Oxford University Press.
Loh, J. 2019. *Roboterethik. Eine Einführung*. Berlin: Suhrkamp.
MacLennan, B.J. 2014. Ethical treatment of robots and the hard problem of robot emotions. *International Journal of Synthetic Emotions* 5 (1): 9–16.
Maibom, H. 2017. *The Routledge handbook of philosophy of empathy*. London: Routledge.
Misselhorn, C. 2009. Empathy with inanimate objects and the uncanny valley. *Minds and Machines* 19 (3): 345–359.
Misselhorn, C. In press. Is empathy with robots morally relevant? In *Emotional machines: Perspectives from affective computing and emotional human-machine interaction*, ed. C. Misselhorn and M. Klein. Wiesbaden.
Mori, M. 2005. On the uncanny valley. In *Proceedings of the Humanoids-2005 workshop: Views of the uncanny valley*. Tsukuba: Japan.
Nagel, T. 1974. What is it like to be a bat? *The Philosophical Review* 83 (4): 435–450.
Newen, A. 2015. Understanding others: The person model theory. In *In Open MIND: 26(T), ed*. T. Metzinger and J. M. Windt. Frankfurt am Main: MIND Group.
Newen, A., L. De Bruin, and S. Gallagher. 2018. *The Oxford handbook of 4E cognition*. Oxford: Oxford University Press.
Nussbaum, M. 2011. *Upheavals of thought: The intelligence of emotions*. Cambridge: Cambridge University Press.
Plantinga, C. 2009. *Moving viewers: American film and the spectator's experience*. Berkeley: University of California Press.
Rorty, R. 2001. Redemption from egotism: James and Proust as spiritual exercises. *Telos* 3 (3): 243–263.
Scheutz, M. 2011. Architectural roles of affect and how to evaluate them in artificial agents. *International Journal of Synthetic Emotions* 2 (2): 48–65.

Schmetkamp, S. 2017. Gaining perspectives on our lives: moods and aesthetic experience. Philosophia 45(4): 1681–1695.
Schmetkamp, S. 2019. Theorien der Empathie - Ein Einführung. Hamburg: Junius Publisher.
Schneider, S. In press. *Future minds: Enhancing and transcending the brain.*
Slote, M. 2017. The many faces of empathy. *Philosophia* 45 (3): 843–855.
Smith, M. 1995. *Engaging characters: Fiction, emotion, and the cinema*. Oxford: Clarendon Press.
Sobchack, V. 2004. *Carnal thoughts: Embodiment and moving image culture*. Berkeley: University of California Press.
Stein, E. 1989. On the problem of empathy: The collected works of Edith Stein. Vol. 3 (3rd revised edition), trans. W. Stein. Washington, D.C.: ICS Publications.
Stueber, K. 2006. *Rediscovering empathy: Agency, folk psychology, and the human sciences*. Cambridge, MA: MIT Press.
Stueber, K. 2018. Empathy. In *The Stanford encyclopedia of philosophy* (Spring 2018 edition), ed. E. N. Zalta, https://plato.stanford.edu/archives/spr2018/entries/empathy/.
Vaage, M.B. 2010. Fiction film and the varieties of empathic engagement. *Midwest Studies in Philosophy* 34: 158–179.
Vallor, S. 2011. Carebots and caregivers: Sustaining the ethical ideal of care in the 21st century. *Philosophy and Technology* 24 (3): 251–268.
Weber, K. 2013. What is it like to encounter an autonomous artificial agent? *AI & SOCIETY* 28: 483–489.
Yanal, R.J. 1999. *Paradoxes of emotion and fiction*. Pennsylvania: Penn State University Press.
Zahavi, D. 2011. Empathy and direct social perception: A phenomenological proposal. *Review of Philosophy and Psychology* 2 (3): 541–558.
Zahavi, D. 2014. *Self and other: Exploring subjectivity, empathy, and shame*. Oxford: Oxford University Press.
Zahavi, D., and J. Michael. 2018. Beyond mirroring: 4E perspectives on empathy. In *The Oxford handbook of 4E cognition*, ed. A. Newen, L. de Bruin, and S. Gallagher, 589–606. Oxford: Oxford University Press.