AI & SOCIETY (2022) 37:535–548
https://doi.org/10.1007/s00146-021-01211-2

ORIGINAL ARTICLE

# Empathic responses and moral status for social robots: an argument in favor of robot patienthood based on K. E. Løgstrup

**Simon N. Balle**[1]

Received: 16 April 2020 / Accepted: 25 March 2021 / Published online: 20 April 2021


**Abstract**
Empirical research on human–robot interaction (HRI) has demonstrated how humans tend to react to social robots with empathic responses and moral behavior. How should we ethically evaluate such responses to robots? Are people wrong to treat non-sentient artefacts as moral patients since this rests on anthropomorphism and 'over-identification' (Bryson and Kime, Proc Twenty-Second Int Jt Conf Artif Intell Barc Catalonia Spain 16–22:1641–1646, 2011)—or correct since spontaneous moral intuition and behavior toward nonhumans is indicative for moral patienthood, such that social robots become our 'Others' (Gunkel, Robot rights, MIT Press, London, 2018; Coeckelbergh, Kairos J Philos Sci 20:141–158, 2018)?. In this research paper, I weave extant HRI studies that demonstrate empathic responses toward robots with the recent debate on moral status for robots, on which the ethical evaluation of moral behavior toward them is dependent. Patienthood for robots has standardly been thought to obtain on some intrinsic ground, such as being sentient, conscious, or having interest. But since these attempts neglect moral experience and are curbed by epistemic difficulties, I take inspiration from Coeckelbergh and Gunkel's 'relational approach' to explore an alternative way of accounting for robot patienthood based on extrinsic premises. Based on the ethics of Danish theologian K. E. Løgstrup (1905–1981) I argue that empathic responses can be interpreted as sovereign expressions of life and that these expressions benefit human subjects—even if they emerge from social interaction afforded by robots we have anthropomorphized. I ultimately develop an argument in defense of treating robots as moral patients.

**Keywords** Social robots · Empathy · Ethics · K. E. Løgstrup · Moral status · Patienthood

## 1 Introduction

People tend to have empathic responses toward social robots. This has been established by a number studies in human–robot interaction (HRI) (e.g. Rosenthal-von der Pütten et al. 2014). And it is generally believed that people empathize with social robots because they attribute human properties to them (Crowell et al. 2019); a process popularly known as mental anthropomorphism (Epley et al. 2007; Airenti 2015; Damiano and Dumouchel 2018). Humans thus perceive nonhuman entities such as robots to have motives, intentions, emotions, and varying kinds of mental states (Krach et al. 2008). Moreover, sociable robots mimicking humans (or animals) has been found to easily invite people's moral intuitions; indeed, simple morphological similarities are often enough to invoke identification and prosocial behavior (Riek et al. 2009). By considering these observations we should not be too surprised that people have such moral intuitions and regulate their behavior accordingly around socially responsive robots.

But, the essential question here for moral philosophy is whether people are right or wrong in treating robots with moral consideration. Usually, this debate is had in terms of moral status: only if some robot qualify as a moral patient will other agents on the moral domain (humans, for instance) have obligations to treat it with moral regard (Gunkel and Bryson 2014).[1] The conditions on which some entity could be considered a moral patient have sparked quite some

✉ Simon N. Balle
snb@cas.au.dk

[1] Department of Theology, Aarhus University, Aarhus, Denmark

[1] What is at issue here is thus a separate question from moral agency for robots, even if some researchers treat them together (Rodogno 2016), consider them as subsets for "full moral status" (Gamez et al. 2020), or find that "moral rights" should be granted to robots once they are competent moral agents (Gordon 2020).

debate among scholars. Many theories share the basic assumption that patienthood rest on subjective experience of mental states, whether the particular position is formulated in terms of sentience or consciousness (Torrance 2008; Singer 2011; Donath 2020), or of having interests (Neely 2014; Rodogno 2016; Basl and Bowen 2020). In this vein, Bryson and Kime (2011) has argued that extending moral consideration to robots is a category mistake caused by mental anthropomorphism; it is an epistemic error in which we 'over-identify' with humanoid robots and mistakenly attribute mental properties on basically insentient and non-conscious things.

But dissatisfaction has mounted against the reasoning that underline these arguments, sometimes categorized as 'property' or 'standard' approaches (Chappell 2011; Gunkel 2017; Coeckelbergh 2018; Danaher 2019). On this thinking, moral status is credited entities that can reasonably be said to have some defined properties, often mental or psychological, that are counted as warrants of moral status in the first place. It is often argued that not only are such approaches inherently anthropocentric (they tend to start from human properties to establish patienthood for non-human entities), they also run into epistemic difficulties as mental properties are not discoverable. These problems have motivated several alternatives to ground moral status by other means. For example, virtue ethical approaches have suggested that rather than being about properties, internal states, or interests, the debate about moral consideration for robots should be concerned with how human behavior around robots reflect and shape the virtuous subject (Cappuccio et al. 2019; Sparrow 2020). Others have taken inspiration from environmental ethics that enables one to argue that moral status could obtain for teleologically organized systems such as nonsentient organisms and, perhaps, artificial objects such as robots (Basl 2019). A specific alternative is the 'relational approach' developed by Coeckelbergh and Gunkel (Coeckelbergh and Gunkel 2014; Gunkel 2017). They protest that standard approaches eschew moral experience and suggest with Levinas that "ethics is first philosophy". Indeed, they hold that ethics precedes ontology and call for deeper engagement with the relational nature of moral experience.

I share Coeckelbergh's (2018) worry in relation to standard approaches, "that empathic responses to robots are bound to remain unexplained and unjustified", and take it as a central motivation in this paper. We cannot simply bypass emotional and pre-conscious aspects of social cognition in our conceptual analyses, as Seibt and Rodogno (2019) formulate it. In this vein, I propose a reappraisal of empathic responses and suggest from an ethical framework that this kind of spontaneous moral reactions should inform our moral status ascription for robots. I do this by exploring and developing on the ethics of K.E. Løgstrup (1905–1981).[2]

I take Løgstrup's ethical analysis to be a relevant fit, for the reason that it deals exactly with the kind of pre-reflexive responses to the Other we witness in HRI studies; where our moral intuitions oftentimes drive us to be more cordial than we would on second thought. In fact, he took empathy, trust, compassion, and similar phenomena as inherently good and valued these spontaneous responses to the Other above acts of moral deliberation. And, interestingly, Løgstrup emphasized that our moral responses do not merely work toward the wellbeing of the Other, but also promote it in what I will call a 'first-person' perspective. Sympathetic acts of charity toward a fellow's plight help resolve one's encircling self-absorption, the human predicament that Augustine of Hippo and later Martin Luther dubbed incurvatus in se (Rabjerg and Stern 2018). This then reveals an ethical aim native to theology—perhaps a surprising or unexpected source for dealing with the ethical challenges of human–robot interaction.

Building up to the ethical debate, I will first make some observations of the phenomenon in question by sketching extant empirical studies on empathic responses, mind-attribution, and moral intuition in HRI. In Sect. 3 I briefly discuss essential definitions and develop commensurability between the explanandum (empathic responses toward robots) and the *explanans* I propose (Løgstrup's ethics). In Sect. 4 I then survey and engage with the recent debate on moral patienthood for robots, which regulates the acceptability of moral behavior toward social robots. Since existing solutions are unsatisfying for reasons I shall explore, I take steps similar to the relational approach before exploring and developing the Løgstrupian alternative in Sects. 5 and 6. Lastly, in Sect. 7, I offer a summation of the argument before reflecting on some implications from a broader societal perspective.

## 2 Sympathies for the synthetic

A growing body of research on HRI have proliferated in recent years, and a number of these have set out to pick up on the effects of anthropomorphism on emotional and moral reactions toward robots (e.g. Gazzola et al. 2007; Krach et al. 2008; Riek et al. 2009; Young et al. 2011; Rosenthal-von der Pütten et al. 2013, 2014; Rosenthal-von der Pütten and Krämer 2015; Krämer et al. 2015; Wang and Quadflieg 2015; Suzuki et al. 2015; Graaf and Allouch 2016; Crowell

[2] Several commentators have regarded Løgstrup as natural comparator to Levinas (e.g. Thornton 2020), and Gunkel (2017) similarly suggest exposing his Levinasian position to that of Løgstrup.

et al. 2019). In some case studies, test subjects have self-reported emotional responses from interacting with robots, while others supplement this by measuring neural activity (e.g. Rosenthal-von der Pütten et al. 2014; Wang and Quadflieg 2015) or electroencephalography (Suzuki et al. 2015).

Unsurprisingly, robot morphology has been found to impact empathic attitudes. In a study from 2009, Laurel Riek et al. explored how people empathized with robots along the anthropomorphic spectrum. They presented test subjects with short video-clips of several robots of incrementing degrees of human likeness, including one actual human person. The videos showed different kinds of mistreatment to the robots and human in turn, and the researchers concluded, perhaps unsurprisingly, that higher degrees of human likeness in robot morphology incites more empathic user feedback. For the android 'Alicia' in the test, empathy levels were self-reported as almost on par with the human 'Anton' in the test (scoring 3.65 and 4.01 respectively on a Likert scale of 6) (Riek et al. 2009). Similar studies have found that sheer size has an impact on user attitudes as well, indicating that bigger is better in terms of perceived agency (an observation echoed in Crowell et al. 2019; Löffler et al. 2019).

Suzuki et al. (2015) measured empathic responses using electroencephalography (EEG) alongside self-reporting methods. They showed media-clips to test subjects portraying robots and humans in supposedly painful situations, such as scissors cutting into a human hand, and subsequently into a robotic one. They generally found that the EEG data followed self-reporting, and interestingly found that in the scissor-hand test, levels of empathy triggered from the perceived pain stimuli were very similar for the human to the robotic hand. Invoking and comparing this with the findings of an older study by Gazzola et al. (2007) who found mirror neurons activated equally in test subjects when observing a human or robot perform a given set of actions, it suggests that human brains are neurologically wired to empathize with mental states believed to be true of entities that are reminiscent of the self.

A similar study aiding and advancing this conjecture used functional magnetic resonance imaging (fMRI) to demonstrate that neurological activity, usually associated with mental state attribution and mental model-building, was activated in test subjects playing board games with robot interlocutors. The researchers concluded in this study that "the same cortical network contributing to mental state attribution in implicit human–human interactions […] was activated in the human–machine interactions" (Krach et al. 2008, 6). They also found that the activity in these networks increases linearly with the anthropomorphic design of the robot. That is, to put it in the vernacular, brain centers responsible for recognizing other minds simply 'light up' stronger the more that interlocutor looks human. Empathic responses make sense as a consequence of perceiving that entity as another mind.[3]

If true that we engage in mental model-building on the neurological level when interacting with humanoid robots, the findings of one particularly interesting study by Rosenthal-von der Pütten et al. (2014) makes sense. In this study, researchers had test subjects watching videos of affectionate and violent treatment of a robot and a human, and then comparing the emotional reaction of participants toward robots and humans in turn. The researchers tracked neural regions associated with emotional responses using fMRI scans in combination with self-reports, and both methods confirmed emotional reactions to observed pleasure or pain for both robot and human. And while they did find slight differences in neural activity when comparing only the videos showing abusive behavior, suggesting more emotional distress and concern for humans rather than robots, "no different neural activation patterns emerged for the affectionate interaction toward both, the robot and the human" (Rosenthal-von der Pütten et al. 2014, 201). If, as Krach et al. and others have demonstrated, test subjects perceive and interact with robots as if they have mental capabilities, it is perfectly reasonable for them to also empathize with them in the face of violent or affectionate treatment.[4]

At this point it is reasonable to ask if empathic behavior is dependent on participants' ignorance with regard to the non-sentient and non-conscious nature of the robots employed. Rosenthal-von der Pütten et al. (2013, 29) reflects on this question in another study where robotic pets were included: "since numerous studies show that people usually do not admit that they see robots or agents as social beings, it is still surprising that participants admitted to having negative feelings when an artificial animal is tortured. This and similar thoughts expressed by HRI-researchers suggests that even if humans are cognitively aware that what they are facing is nothing more than 'dumb' machinery, they cannot help but translate social responses from human–human interaction into encounters with these novel and responsive social actors. In fact, as Airenti comments in relation to the uncanny valley-effect, people are perfectly willing to interact socially with robots against the better of their knowledge,

[3] And it is likely that adopting what Dennett has called the intentional stance toward robots contribute in this process (Perez-Osorio and Wykowska 2020).

[4] In the light of these findings, it is interesting to note how the violent end of the hitchbot-project attracted so much empathic responses from people that was never even in contact with the robot (Vander-Maas 2015). The resulting #RIPHitchBot and #Vengebot outcries on social media when the HitchBot was eventually found in a ditch, dismembered and decapitated, shows remarkable empathy—the outcries even match the youtube-laments over the 'torture' of the Boston Dynamics canine-inspired robot 'Spot' (Coeckelbergh 2018).

as if that robot is another mind.[5] Indeed, people may actually prefer that robots not leave them in doubt of their non-conscious machine nature: "Humans may interact with machines", Airenti (2015, 125) concludes,"but they reserve to themselves the power to fill their mind, attributing both mental states and emotions".

It seems some people not only engage in this kind of suspension of disbelief, but actively engage in building relations with robots they know have limited capabilities. How do these observations square with the expectation that empathic responses toward robots will rescind as the 'novelty effect' wears off (Smedegaard 2019)? It is difficult to glean a tendency from the little research we have on long-term effects of social robots. But given what we know about human sentimentality one could equally well suspect empathic responses to increase. Some of the observations we do have on long-term effects suggest an attachment to robots does not fade after the novelty effect wears off. People arrange funerals for their 'dead' AIBO's (McCurry 2018) and one person has compared the heartbreaking loss of his JIBO to the pain of losing his mother to dementia (van Camp 2019)—suggesting levels of attachment way beyond naming a Roomba. In two studies on long-term interaction, researchers found that social interaction and attachment generally increased over time (Leite et al. 2013), especially when robots have autonomous movement and conversational abilities (Kertész and Turunen 2017). But observing that social interaction and attachment with robots in many cases continually increases is of course not synonymous with establishing that this translates into more substantial attitudes of moral responsibilities over time.

And while long-term perspectives on human relations to robots is an illuminating factor for the moral patiency debate, these concerns lie beyond the present scope. Why? Because in the position we develop here, building on the ethical analysis of K. E. Løgstrup, it is precisely the kind of intuitive and immediate moral response to the Other we are interested in. In contrast to most moral philosophies, Løgstrup valued pre-reflexive and spontaneous responses. Because moral intuition, deeply seated in our embodied co-existence, outperforms intellectualized moral reasoning in the interpersonal sphere.

## 3 Anthropomorphism, empathic responses, and robots—some definitions

At this point it is helpful to run through a couple of definitions of the rather interrelated, complex and not entirely undisputed terms employed throughout. Also, by spelling them out I hope it becomes clearer why the ethics of Løgstrup match the issue at hand.

Why are emotional attitudes such as empathic concern invoked toward social robots such as NAO, Pleo or Sophia?[6] Trivially, because we anthropomorphize them. A typical definition of anthropomorphism runs like"the tendency to imbue the real or imagined behavior of nonhuman agents with humanlike characteristics, motivations, intentions, or emotions" (Epley et al. 2007, 864). And since robots look and behave like entities we know to be alive, sentient, and conscious, we attribute to them such mental traits (Wang and Quadflieg 2015; Airenti 2015). The evolutionary origin of anthropomorphizing[7] is thought to be that human survival was more probable with a strong cognitive faculty for agency detection. It was better to believe a predator was approaching and take appropriate measures one time too many than too few. Early humans thus had a very high motivation to anthropomorphize in order to discover and understand the behavior of perceived agents in one's environment; a cognitive device still with us today. Hence, anthropomorphism has to do with agency recognition and prediction and sometimes considered to be part of a "Hyperactive Agency Detection Device" (HADD) (Damiano and Dumouchel 2018, 2). Anthropomorphizing a robot is thus epistemologically questionable; correct in that it picks out a social agent seemingly performing autonomous behavior and capable of manipulating the environment,[8] but wrong to infer the mental abilities that is usually true of such agents. Our cognitive faculties are simply not developed to categorize and deal with these new kinds of entities, as Nyholm (2020) also points out. It is due to this 'epistemic lapsus' that humans identify with robotic

[5] Another possibility not entertained in the empirical literature is that what people interact with when perceiving another mind in the robot, is something like a sum total of the mind of the designers and programmers who made the artefact. Similar to how one can feel connected to and in dialogue with the artist by engaging with their work whether it is a painting, a novel, theatre play, or a piece of music. Perhaps robots even derive moral status this way. To limit an already broad scope, I will leave it for future research to develop on this idea.

[6] I take any physical robotic artifact with a social interface, autonomous movement/behavior, and with capacities to recognize and interact with other entities as a social robot. This is what I have in mind when I in the following when I simply write 'robot', unless explicitly stated otherwise. A minimal definition is sufficient for my purposes here, as I'm not interested in the robot per se, but in human responses.

[7] We should also be mindful of the cultural underpinnings of anthropomorphism; culture plays a role in determining which physical traits are associate with mind and agency. The cultural dimension has been testified at least since Xenophanes, who ironized that Greek gods were pale and blue-eyed while African deities were black-skinned and snub-nosed, and also remarked that if horses and lions had gods and the ability to paint them, they would probably look strikingly like horses and lions.

[8] Even if you could argue that a robot's agency is just an extension of its makers'.

artifacts that look and behave like humans (or other familiar social agents). And ultimately what paves the way for the effective attunement to the mental and emotional states we expect to find in such entities. While anthropomorphism as explanation raises a host of interesting questions (not least in relation to theory of mind), the significant one here is the ethical interpretation of the empathic responses it solicits humans to extend toward robots.

Defining empathy is no less tricky. At the very basic level I take empathy to be an 'affective resonance phenomenon' directed at the wellbeing of others. On a more narrow definition, having empathy with someone is the experience of feeling what one senses another person is feeling, a sort of copying of another's emotional state in a specific situation (Misselhorn 2009; Maibom 2014).[9] Like strings on guitars and pianos attune and resonate with each other in the same frequency range, people reverberate the emotion of others when empathizing. The difference to the next-door notion of sympathy is that while empathy is going through another's emotional state, sympathy is welfare directed (Clark 1987; Maibom 2014); it is an emotional reaction toward your fellow's plight without necessarily echoing their emotional state. While empathy is often understood as a protomoral feeling-with, sympathy, compassion, and empathic concern are variants of feeling-for and are thus closely linked to prosocial and moral behavior (Ugazio et al. 2014).

But getting too technical could prove counter-productive here, since the term is more loosely defined and employed in the empirical studies currently of interest. Observe that Rosenthal-von der Pütten et al. (2013, 2014) investigate "empathic concern" which they take to denote the basic emotional distress directed at the suffering of others. Suzuki et al. (2015) follow Decety's model for empathy as comprised of three components "affective arousal, emotional understanding, and emotion regulation". Leite et al. follow Hoffman in a quite loose definition of empathy as "an affective response more appropriate to someone else's situation than to one's own"(Leite et al. 2013, 303). Others (e.g. Riek et al. 2009) does not give a definition but links it to prosocial behavior, and leave it to participants to define if they experienced empathy toward some robot in the study. In short, empirical researchers keep their findings of empathy to robots loosely defined, a sort of empathy + that often include aspects of what should technically count as sympathy or compassion. In order to encompass these studies, we need to capture and employ a broader sense of the term.

Now, as we aim here to illuminate the moral patiency question by interrogating human empathic behavior toward robots, we also need to spell out the relationship between empathy and morality (Maibom 2014). As touched upon above, empathy as a feeling-with is the affective basis for moral interpersonal behavior. In this pre-reflexive attunement to the Other, deeply seated in our responsive bodies, we have access to significant knowledge about the state of the Other, that the conscious and reflective mind does not immediately have. Of course, how we choose to act morally from this knowledge is ultimately the result of other variables (Ugazio et al. 2014). But this goes to show that empathy is, at bottom, morally charged. To capture this pre-reflexive affective basis that generate spontaneous moral motions (that the subject is free to either reject or act upon) and to be in concert with empirical research, I will prefer the notion of 'empathic response' rather than sympathy or compassion.

Empathic responses have—along with most other pre-reflexive and unforced prosocial gestures—generally received little praise in western moral philosophy.[10] This has been pointed out by Stokes, who notes that neither Kantian nor utilitarian strands of ethics have valued pre-reflexive and spontaneous behavior (Stokes 2016). Morality proper is thought to rest on deliberation about maxims, principles, or utility and thus 'thoughtless' actions has received marginal attention. Løgstrup is one exception where the spontaneous responses to fellow men, especially the stranger, takes absolute center stage. But is his definition and phenomenological approach commensurable to the responses HRI research has classified as empathy? Since Løgstrup wrote in Danish, he did not use the term empathy but rather 'medfølelse', which translates literally to 'feeling-with' (Løgstrup 2015). But it is clear that Løgstrup thought of medfølelse as welfare directed, as he defines it as the preoccupation with the Other's plight to "remove hindrances to the freedom and flourishment of the distressed" (2015, 271, own translation). In this way he employs a broader definition of empathy, not too far from 'empathic concern' found in Rosenthal-von der Pütten et al. (2013, 2014).[11]

[9] Sometimes theorists divide empathy into an affective and a cognitive variant (Davis 1983; Maibom 2014). But the latter is defined very close to 'theory of mind', and I shall prefer this term when considering cognitive aspects of empathy (Tisseron et al. 2015; Redstone 2016).

[10] I do not employ a strict distinction between ethics and morality, but tend to use ethics as a meta-discourse of morality; as the philosophical and analytical dealing with the norms and manners of human social behavior. Behavior can thus be moral, while deliberating about morality is an ethical enquiry.

[11] Giving a phenomenological account of empathy is obviously a very different undertaking that measuring its neurological substrates as some of these studies does. In this sense, taking up Løgstrup is coming from a completely different interest, even if he did recognize that empathy was underpinned by "biological processes that cause ripples in our minds" (2015, 201, own translation). But I take it that keeping phenomenological definitions as descriptively close to the empirical observations as possible renders the analysis more probable.

In sum, empathic responses thus denote an emotional and pre-reflexive concern for the wellbeing of others that include extra-empathic motions—a definition that encompass its uses in empirical studies, captures the basic moral nature of the phenomenon and is in concert with the way Løgstrup employs the term.

## 4 Moral concern for robots and the patiency-debate

We are now in a position to move to the normative side of the issue at hand: whether empathic responses are morally blamed or praiseworthy. Answers to this are dependent on the status of robots as moral patients. We have seen that people tend to regulate their moral behavior toward social robots on account of empathic responses, to de-facto extending some degree of moral patiency; but do robots qualify as proper objects of human moral concern?

As introduced above, the debate about the moral status for robots has standardly been a discussion on properties and which ones permit moral status. On the property-approach, popular in the analytical tradition, moral patiency obtains for any entity that possesses relevant and sufficient properties. Consequently, moral concern and behavior toward entities devoid of the relevant properties is simply mistaken. But there seems to be very little agreement about which are the necessary or sufficient ones. One common denominator seems to be following Jeremy Bentham's break with Cartesian tradition—which for a very long period dictated rationality as the determining quality for moral standing—when he suggested we rather ask 'can it suffer' than 'can it think' when considering moral status. Entertaining this question led to preferring sentience over rationality, and the resultant increase in sensitivity to the suffering of other beings has led to a significant expanse of our moral circle, as the history of ethics testifies (cf. Singer 2011). At the very least then, the ability to feel pain seems to be a central property. But, as Dennett (1998) and others point out, we do not properly understand pain. Do we consider pain with corresponding mental qualia only, or do we take instinctive pain (in lower animals) or symbolic representations of pain (in robots) to be deserving of moral patiency?

In his model for moral patienthood, Torrance (2014) takes the conscious experience of suffering and satisfaction as constituents of moral patiency for artificial agents. Meaning that entities capable of deriving pleasure or pain from having one's goals achieved or frustrated deserve moral status. Others include in the same vein the ability to have significant interests[12] as criterion for moral patiency, arguing to the effect that only conscious beings are able to enjoy these (Rodogno 2016; Eberl 2017; Basl and Bowen 2020).

Some criticism has mounted against these approaches. A common objection is that mental properties are not discoverable, as Danaher (2019) among others has put forward. Mental states are not observable, as Turing also noted; we can only infer these on the grounds of observable behavior (Turing 1950). Passing the (in)famous Turing test is no proof of conscious thinking or experience, only of skillful manipulation of symbols. As Searle famously put forward in his 'Chinese room' argument, syntactic competence does not suffice for semantic understanding (Searle 1980).

As a way forward, Danaher advocates 'Ethical Behaviorism' which can be read as a modest version of the property approach (Danaher 2019). Given the "epistemic opacity of properties", we should simply take behavior to be sufficiently indicative of mental states. And we should affirm moral status of robots when they are "roughly performatively equivalent" to that of other entities enjoying moral patiency. He also argues that the performative threshold above which moral status obtain might not be that high. On this utilitarian outlook, this is no different from how other entities come to enjoy moral patiency on our moral circle—that we take their behavior as indicative of corresponding mental properties. No, we do not have any certainty of 'what's going on, on the inside', but it does not matter on this approach. According to Danaher, we should simply bite the bullet and concern ourselves only with the behavioral testimony to properties.[13]

Another alternative that shares the insight that behavior is a key concern is the virtue ethical approach. Virtue ethical accounts has received some traction in recent years (Vallor 2016; Cappuccio et al. 2019; Sparrow 2020; Coeckelbergh 2020), and they are generally concerned with delineating how interactions with robots both reflect and cultivate a virtuous character. Reversely, such approaches are concerned that cruel or inappropriate behavior towards robots—whether enacted unprovoked or solicited by the robot or the fantasy around it—reveals and exacerbates a vicious character. It is thus out of concern for human character formation and the shaping of appropriate habitual responses that moral consideration for robots should obtain. The price for

[12] Pursuing goals, exercising freedom, maintaining meaningful social relations, achieving pleasure, avoiding pain and so forth are often counted among interest (for humans at least). But which interest are more significant and which ones AI's and robots can be said to have are difficulties still debated.

[13] A similar argument is put forward by Gordon (2020) who essentially argue that autonomous deliberation and decision-making behavior warrant moral status. Behavior, in this case autonomous decision-making, not properties, is sufficient for being "full ethical agents". And since the AI of robots are already making autonomous decision, we can soon rightly consider them subjects of morality. Conferring moral patiency is then just around the corner, and Gordon provides a four-part cumulative argument in favor of that.

not observing common moral conduct around robots representing humans and animals, is the corruption of human morality and that undesirably conduct will ultimately translate back into relations with real humans or animals. Moral obligation owed in relation to robots as patients is then neither directed at them, nor derived from their inherent value based on some property, but owed to human morality or the humanity that robots represent.

Others take issue with the inherent anthropocentric bias of property approaches. (Coeckelbergh and Gunkel 2014; Gunkel 2017; Coeckelbergh 2018). That is, proponents of this approach often start with properties true for human beings, and then proceed to find those properties in other entities. It seems misguided, they argue, to establish the patiency for nonhumans, by looking for human properties. Coeckelbergh and Gunkel raise several other objections and think the cumulative case warrants another approach altogether (Coeckelbergh and Gunkel 2014). Whether they have amassed something insurmountable for property approaches is beyond the scope here, but I think their call for attention to moral experience motivates exploring alternatives. Property approaches simply lack engagement with and attention to the emotional and relational forces of social reality, and the worry is "that empathic responses to robots are bound to remain unexplained and unjustified" (Coeckelbergh 2018, 147).

For these compounded reasons, Gunkel (2017) thinks we need to 'think otherwise' and he advocates with Coeckelbergh a 'relational approach' (Coeckelbergh and Gunkel 2014). Crucially on this understanding is the priority between ontology and ethics. In their joint article on the moral standing of animals,[14] Coeckelbergh and Gunkel follow Emmanuel Levinas by contending that ethics precedes ontology. On this view, "morality is not a branch of philosophy, but first philosophy" (Gunkel 2014, 126), while the relational aspect of human nature is emphasized. Contrary to a property approach that begins by making ontological determinations about who or what is a legitimate moral subject, they propose with Levinas to see it the other way around: moral and social relations are given, and the Other always and already obligates me in advance of customary decisions and debates concerning who or what is (not) a moral subject (Coeckelbergh and Gunkel 2014). We first respond to the Other, and then, after having made the response, we identify and determine what we responded to (Gunkel 2017). Thus, when our moral intuition informs us to treat some encountered entity with moral consideration—when an entity 'supervenes' and 'faces' us and demands that we respond (Coeckelbergh and Gunkel 2014)—it truly becomes an Other for us in the Levinasian sense. Coeckelbergh (2018, 149) maintains that this approach "takes seriously the phenomenology and experience of other entities such as robots, and sees moral standing not as the starting point but rather as the outcome: moral standing is itself the outcome of the process of relation and interaction". Gunkel (2017, 10) further infer with environmental ethicist Callicott that "relations are prior to the things related" and can thus agree with Coeckelbergh that moral status does not obtain on intrinsic grounds, but extrinsically: "it is attributed to entities within social relations and within social context".

It is not entirely clear if this account amounts to metaphysical constructivism. But the idea that we identify and determine the Other subsequent to our response to them seems to suggest so. But on a weaker reading they might simply say that moral responses and intuition serve as heuristics for ontology. The difference lies in whether moral status is constructed as an ontological reality or rather discovered whenever an entity 'supervenes before us' and trigger our moral responses. In either case, moral patienthood for robots would ultimately obtain on the basis of human perception on this account; on how we intuit and respond to the Other and construct it in relations to us (thus ultimately not getting rid of the anthropocentric bias, one might add). But since we each have different moral intuitions, should applying the Levinasian idea not only mean that patienthood follow for specific robots in relation to specific people depending on their moral response to the robot? To my knowledge, Coeckelbergh and Gunkel have not in their published work contemplated if moral patiency should obtain only on the individual level, but rather seem to think in universal terms. And I suspect suggesting an inversion of ethics and ontology is by definition an across-the-board enterprise. Perhaps this is why they hold back on taking the Levinasian idea to its logical conclusion and never explicitly defend moral patienthood for robots; the implications are simply dauntingly vast. Rather they entertain and scrutinize the idea and formulate their approach as a new "relational and moral hermeneutics" (Coeckelbergh 2014) and a way of "thinking otherwise" (Gunkel 2017).

In any case, I agree that the epistemic uncertainty of mental properties and the negligence of social-relational experience are good reasons to motivate exploring alternatives to standard approaches.[15] I share Coeckelbergh's concern

[14] Though they are here focused on animal Others, they both apply the same idea for robotic Others (Coeckelbergh 2014, 2018; Gunkel 2017, 2018).

[15] I acknowledge there are more or perhaps better arguments (relative to ones worries and aims) such as the charge of anthropocentric bias, but I cannot consider them all here. Other arguments are explored in e.g. (Gunkel 2018; Coeckelbergh 2018; Danaher 2019). Another reason one could take issue with present property approaches is the implicit substance metaphysics they often build on. Conceiving of individual subjects as constituted by processes rather than substances (e.g. Eck and Levine 2017), allow for emergence of properties. Properties (e.g. those we base moral status on) would not be fixed to cer-

that empathic responses are left unexplained and unjustified. The view I develop here is in many ways compatible with the relational approach, and it is likely that comparing the Levinasian approach to a Løgstrupian one might contribute in honing it. Though the analysis I offer will bring out some details that direct it at a different and more limited conclusion with respect to moral patiency for robots. My view also shares with virtue ethics an emphasis on the effects of interpersonal behavior for human flourishment. But as we will see below, this is a superficial agreement that diverges at a deeper anthropological level. Yet, before being able to do so and develop my account in relation to similar ones, I first offer an exposition of the central tenets of Løgstrup's ethical thinking for our purposes.

## 5 The Løgstrupian alternative

The aim of exploring and developing Løgstrup's ethical apparatus to bear on the present issue is not merely to give a phenomenological interpretation of empathic responses to robots, but ultimately to provide a normative framework for moral status ascription. As teased above and developed in the following, Løgstrup found spontaneous empathic responses directed at the wellbeing of others to be instances of genuine good. This idea lies at the root of the argument advanced in this paper—that we can acceptably take robots as moral patients—so exploring it here is an essential task.

To get under the hood of Løgstrup's sometimes rather convoluted thinking[16] before we spell out the implications for moral patienthood for robots, we need to unpack a metaphysical dualism known in the Løgstrup reception as the Doctrine of two accounts[17] (Rabjerg 2017). To Løgstrup, the prime ethical question is not why there's so much evil in the world, but rather why there is good—such as our empathic responses to fellow man, animals and now apparently also social robots—given our selfish ways and everyone's struggle for themselves. The brilliance of Nietzsche, according to Løgstrup, was his brutal exposition of man's evil and hypocrisy (Løgstrup 2014).

But if man really is just a selfish beast, how come we experience love, mercy, trust, empathy and so forth, Løgstrup asks. Conceding that human will is bound by selfishness was rather uncontroversial in Løgstrup's own Lutheran tradition,[18] and also fit well in a Darwinian scheme. Left to our own devices, human beings are 'curved in on themselves', incurvatus in se. And while Løgstrup did adopt this Lutheran tradition of describing humanity's sinful nature, he also maintained that we sometimes experience that the centripetal force of our incurvature is displaced (Rabjerg and Stern 2018). This would occasionally happen, observed Løgstrup, when responding to the plight of fellow man. Charitable acts directed at the welfare of others had the potential of breaking the subjects encircling self-absorption and calling forth instances of genuine good.

But Løgstrup found it implausible and inconsistent to credit these acts of kindness on the human self, if human volition was indeed radically corrupted and selfishly incurved. To avoid this, Løgstrup suggested that another source for goodness had to exists, outside of human volition—and that this is what Nietzsche, Kant, and Kierkegaard had overlooked (Løgstrup 2013). They are blind to this because they only keep one account, the anthropological, as Løgstrup terms it. Since that which breaks our incurvature never manifest in a social vacuum, but only appear in our interdependent lives, when responding to fellow man, the solution for Løgstrup was to open an 'ontological account' on which these phenomena can be credited (Løgstrup 2010). By denoting it 'ontological' he distinguishes it firmly from 'the anthropological account', suggesting the account is 'life itself'. By positing another account, Løgstrup avoided the problem he thought Kant and Kierkegaard struggled with, namely to explain how good and evil coexist in an inner human; where they battle for supremacy and causing humans to alternate between good or bad actions depending on how firmly the moral reins are held by human will, maxims or successful rational deliberation.

Genuine good is thus something that emerges between us and not something that originates from human deliberation or volition, as the latter was impossible on Løgstrup's reading of the negative Lutheran anthropology just mentioned. Consequently, he regarded the rationalized and coerced moral behavior as calculated appropriations of the genuine self-less neighbor love that emerge as a pre-reflexive response to the Other. In this way, spontaneity became a hallmark of the genuinely good act, barring moral deliberation about how to be a virtuous subject from tainting the pure and 'sovereign' act. He termed these responses

Footnote 15 (continued)

tain biological entities made of the right substance, but would be substrate indiscriminatory. At any rate, exploring this is beyond the present scope.

[16] For a fuller exposition of Løgstrups ethical thinking in English, see (Fink 2017; Rabjerg 2017; Wolf 2017; Niekerk 2017; Stern 2019).

[17] The 'Two Accounts' is mentioned in the Ethical Demand (2010) [1956] and later developed in Etiske begreber og problemer (2014) [1971]. "But there are two accounts to keep and to distinguish from each other. The account of our given life and the account of our ego" (2014, own translation). Note that the Danish 'konto' translated as 'account' does not mean 'explanation', but rather means 'a record', as in bank account.

[18] Cf. Luther's de servo arbitrio.

'sovereign expressions of life' (suveræne livsytringer), first in an article on 'Sartre's and Kierkegaard's account of the demonic enclosure' (Løgstrup 1966) and develops them later, most notably in his Controverting Kierkegaard from 1968 (Løgstrup 2013). Besides empathy, the examples Løgstrup often mentioned include trust, love, mercy and openness of speech.

Let us briefly look at what larger role sovereign expressions play in Løgstrup's thinking, before we can draw out the implications for the present issue. Løgstrup found them to correspond to the ethical demand, and he thus developed the concepts of sovereign expressions not just as a response to the anthropological problem, but also to the ethical one. But why think there is an ethical demand to care for others in the first place, and how is such a demand fulfilled?

Given our interdependent human lives and how we constitute each other's word, we always find ourselves in power relations with one another. These power relations become manifest in how we always have something of the life and welfare of the other within our power. Or, as Løgstrup metaphorically has it in The Ethical Demand (2010): we always have something of the other's life in our hand. No matter how small or great the amount, it is always up to the individual to decide whether to administer it to the destruction or flourishment of the other. The ethical demand is then nothing more or less than the demand to always take care of however much of the other's wellbeing is within our grasp.

But why heed this demand? Or: why think that other's wellbeing is my burden? Løgstrup does not point to God as a moral law giver as could be expected in his tradition, since his aim was to formulate the ethical demand in secular terms. Instead, he underpinned the ethical demand with the notion of life itself as a gift. Since we all receive life as an unmerited gift but have no benefactor to respond to,[19] we are left to direct our gratitude to the people who constitute our world. In fact, Løgstrup (2010) formulates the implication of being unmerited receivers of life as being 'in debt'. And to ward off protests to the demand in the name of reciprocity, Løgstrup was quick to qualify the demand as one-sided.

But the overwhelming problem for the moral subject, as Løgstrup sees it—and now slowly coming back to the sovereign expressions of life—is that the ethical demand is principally unsatisfiable. Why? Because as soon as we realize, or 'hear', the demand, it is already too late; hearing the demand marks the absence of love. If one has to be told to be merciful and do good (whether by reason or by a moral lawgiver) one has already failed. In other words, a sovereign expression failed to emerge, because the subject was busy deliberating about utility, motives, maxims, right and wrong etc. In response to hearing the demand, I appropriate what I should have been doing immediately, had I acted from spontaneous neighbor love. As Løgstrup (2013, 127, own translation) formulates it: "Morality is the supply of substitutive motives for substitutive acts". In this critique of Kantian ethics, Løgstrup points out that deliberating about maxims (or duties, utility, responsibilities, motives, etc.) is not just 'one thought too many' (as Bernard Williams had it): it is a testimony to moral failure.

This brings us full circle back to the sovereign expressions of life that correspond to the ethical demand: they begin to emerge in a pre-reflexive state before one hears the demand. Sovereign expressions are spontaneous or preconscious, something that happens 'behind our backs' as Løgstrup interchangeably has it. Only in this way are we able to satisfy the ethical demand, only we are strictly speaking not the subject of those acts. Our job is to surrender to or consummate the sovereign expressions (Løgstrup 2013) without corrupting these genuine goods with our selfish desire to be good.[20]

Now beginning to draw out consequences for our present purpose, it's important to note that Løgstrup's very central idea was that sovereign expressions manifest for the flourishment of the life of the Other. But he also argued that the acting subject is a beneficiary of the sovereign expression too.[21] In other words, these expressions fulfill a 'double task': they promote wellbeing for both agent and patient, or what I will call 'first-person' and 'third-person' benefits respectively. This is a critical distinction here, since even when there's no flourishment or interests per se to promote for the robot Other—there still is for the human counterpart. I do however suspect Løgstrup regarded the first-person benefit more as corollary than primary, though he definitely considered it a vital one.[22]

The first-person benefits of consummating the sovereign expressions of life are that they break the centripetal force of our self-encircling thoughts and feelings that characterize the human incurvatus in se. We are liberated from this "inturnedness" not by God's grace, but only through our

[19] Or, if you have the same religious background as Løgstrup, you have the Christian god. But since God in this tradition demands you to always love your neighbor, the ethical import is the same as not believing in a creator god.

[20] For precisely this reason, Løgstrup was very skeptical of virtue cultivation, as I shall return to below.

[21] Niekerk has brought out and analyzed the idea of the 'realization of self' that, according to Løgstrup's Controverting Kierkegaard, the sovereign expressions of life bring about (Niekerk 2017). Becoming a self is not a task for our reflection, is the charge Løgstrup levels against Kierkegaard (Løgstrup 2013). Consummating the sovereign expressions accomplishes this, as they not only lift me for a moment out of my self-encircling thoughts and feelings; I'm becoming a self as I surrender to them.

[22] "Even our very identity rests on them [the sovereign expressions of life]" (Løgstrup 2015, 112, own translation).

ethical encounter with other people, as Rabjerg and Stern (2018) argue. And for precisely this reason—that sovereign expressions of life only emerge in social encounters—we need other social beings to be 'unturned'. We are captives within ourselves but set free in social interaction, to rephrase a journal entry by Løgstrup.[23]

To illustrate: a friend of mine happened upon a lost tourist couple, looking for a meal in her small-town neighborhood after all stores and restaurants were closed. Overcome with empathy for them she invited the couple home for dinner before she could stop to think it over. Would it be okay with the family back home? Was this proper use of her economic and social capital? And why should the couples' poor planning be her burden? But going through with her empathic response—In Løgstrup's terminology: consummating the sovereign expression of life—she invested herself in trying to care for these people to the best of her imagination and ability. In taking the strangers home to cook for them, she was moved by the openness and kindness emerging between them; not just hers, nor solely theirs. She has often since described that night as one of her best memories, and how an openness toward the world lingered with her. To push the Løgstrupian vocabular: the spontaneous sovereign expression turned her outward from her inturnedness.

Suppose now that the couple had merely been actors or philosophical zombies. In that case, the couple was not really helped by my friend, but she might still have escaped the incurvature of selfishness through the experience. By the same token, this could simply have been a robot (albeit of a future and more sophisticated kind), and the encounter would in principle still have the potential to promote my friend's wellbeing. But, on a smaller scale, such interactions could in principle happen now in relation to responsive social robots we anthropomorphize and 'imaginatively perceive'[24] as others.

To recapitulate the point: Sovereign expressions contribute to first-person flourishment, separate from the needs they meet on the receiving end of the expression. Or, if you will, both subject and object are on the receiving end of sovereign expressions. And while singling out the positive first-person effects and disregarding the third-person benefits is probably beyond Løgstrup's intent,[25] it is nevertheless on this critical development my argument turns. It allows me to be agnostic about metaphysical and technical possibilities of robots acquiring mental states or intrinsic interests that might ground their moral status. But more importantly, it provides a normative reason for the permissibility of taking robots as moral patients.

On this approach then, we have an interpretive framework for the empathic responses toward robots that empirical research report. The ethical demand impinges on test-subjects witnessing robots being tortured, for example, and an empathic gesture might emerge as spontaneously response. But obviously, picking up on an empathic acting impulse is far from equal to a fully consummated sovereign expression. The lab-cases or robots themselves might be too limited in design or the interactional ability for someone to bring sovereign expressions to their fruition. Likewise, we might not be able to go through the motions in relation to many extant social robots in the wild. And this might in both cases cause awkwardness or discomfort on the human counterpart, not unlike the uncanny valley phenomenon. Moreover, this difficulty might even be accompanied by inappropriateness, if sovereign expressions toward robots divert attention away from true third-person beneficiaries, i.e. sentient beings like humans and animals. I shall discuss this more in closing, after considering how my Løgstrupian approach measures up against other approaches in the literature, specifically virtue ethics and the relational approach.

## 6 Differentiating the Løgstrupian approach

The position developed above shares with other approaches that patienthood for robots might be credited on extrinsic premises rather than intrinsic ones. Specifically, my proposal shares with virtue ethical approaches the prospects of what I have termed 'first-person benefits' from extending moral consideration to robots. In virtue ethical approaches, this arises as an opportunity to exercise or promote a virtuous character, as Cappuccio et al. (2019) have it. In their formulation, a virtue ethical account "recommends treating social robots in a morally considerate manner because this is what a humane and compassionate agent would habitually do in their social interactions and because the opposite behavior would not be compatible with a virtuous lifestyle and moral flourishing" (Cappuccio et al. 2019, 13). As a consequence, mistreating a robot would be detrimental to the subject's character by animating a vice, and this is why robots should be treated with moral considerations, even if such mistreatment does not cause harm to any sentient being. Along those lines, others invoke Kant's 'cruel habits argument' to describe this effect, that "he who is cruel to animals becomes hard also in his dealings with men" (e.g. Darling 2016, 1). More broadly, in the vocabulary of Vallor

[23] "We are captives within ourselves. We can only be set free by fellow man" (journal entry by Løgstrup quoted in Rabjerg 2017).

[24] The notion of 'imaginative perception', suggested by Misselhorn (2009) and developed by Redstone (2016), propose to make sense of empathy with sociable robots. The central idea is that empathy toward robots is triggered as humans imaginatively perceive emotions in them.

[25] I suspect one could argue from a Løgstrupian perspective that sovereign expressions require third-person benefits, that they only come as complete packages; that having 'half an expression' amounts to having nothing.

(2015), our interaction with robots carry the potential risk of moral 'deskilling' and, conversely, the opportunity for moral 'upskilling', in the same way employing new technologies has historically demoted certain practical skills while promoted others.

This way of framing how interaction with robots might contribute to human flourishing, however commendable and educational such accounts might be, is quite far from what we can take Løgstrup's ethical thinking to support. Løgstrup himself was very skeptical that we could edify our moral selves. His critique was not only born from his anthropological pessimism (cf. Rabjerg 2017; Rabjerg and Stern 2018), but he also regarded the project of exercising morally correct dispositions as a way for motives and outcomes to come apart (Løgstrup 2013, 128–129). In the self-reflection on aligning motivation with virtues, the individual loses sight of the Other and "thrown back onto itself" (2013, 128). For this reason, the motivation to extend moral consideration to a robot (or another human for that matter) as an opportunity to exercise one's virtues while protecting one's moral character from the corruption of bad behavior, essentially runs contra to his two-accounts thinking (2010, 158–162). Our moral characters are already corrupted, he would argue, and the self-congratulatory attempts being good is simply human incurvature in disguise: civil on the outside, but really just a way for the enclosed self to continue their self-encircling. The first-person benefits of taking robots as Others that I propose with Løgstrup, eventually lie elsewhere than where virtue ethicists suggest, namely in the capacity of sovereign expressions to displace our inturnedness.

Now, there is another tricky and overlooked issue related to moral status conferred extrinsically that applies equally to the relational approach as well as the one I have developed here: If moral status depends on extrinsic social premises—such as people's moral responses or intuition—what happens to moral status if those extrinsic premises change? Suppose some users no longer perceive the robot as a patient or if different users have differing views and intuitions regarding the same robot, will its moral status change so that it is a patient only sometimes? Exactly how volatile moral status is in relation to shifting moral responses among users is beyond the present scope. But bearing a sensitivity to this issue in mind, I think we can only stake a modest or 'weak' claim of patienthood for robots when working from this kind of extrinsic premises. With weak here meaning permissible rather than obligatory and individual rather than universal—provided the extrinsic premises hold. By contrast, on a 'strong' position all agents on the moral domain would have an obligation to respect the robot as a patient.

I think the Løgstrupian position I develop differ on this point from the relational approach based on Levinas. The phenomenological description of how robots "supervenes before us" (Gunkel 2017, 10) and "face us, take us out of our self-involvement, and demand from us that we respond" (Coeckelbergh and Gunkel 2014, 723) is remarkably close to the Løgstrupian vocabulary. Yet subordinating ontology to ethics and giving it priority in both temporal sequence and epistemological status suggests a fundamental disagreement with Løgstrup's ontological ethics (cf. Thornton 2020). And in the present context, it might be the Levinasian axiom of ethics as first philosophy that drives their relational approach towards a stronger conclusion than what I suggest the extrinsic premises can bear.

## 7 Conclusion and perspectives

A Løgstrupian approach as I have suggested and developed here provides an understanding of and appreciation for the moral engagement with robots that research in HRI has demonstrated to occur and that we might expect to increase as social robots continue to develop and proliferate. I have argued that empathic responses incited by engagements with robots, when anthropomorphizing and intuiting them as Others, can be read as the impulse of sovereign expressions of life. Such moral phenomena can promote good for the moral agent (the human), even if the patient (the robot) of this moral concern lacks the formal properties of being a moral patient. Namely in the way such forces displace the inturnedness of our human incurvature. By conceiving of empathic responses and similar moral phenomena as sovereign expressions, treating robots as moral patients has its merits, not because robots currently are true beneficiaries of these expressions but because we as human agents are. The implication is that the moral status of robots as patients is derivative of their relation to humans, and for this reason, we can only mount this argument to defend moral patienthood for robots in individual cases.

One obvious criticism is that the implication of my argument is quite exploitable by creators and interest-holders of robots. We should not be blind to scenarios where social robots are employed as friendly front-end interfaces of extensive data-harvesting systems. In the service of these or more malicious interests, manufacturers will likely design robots to be evocative objects that deliberately invoke empathic responses. As designed objects, robots materialize and mediate certain moral values and norms intended to guide human behavior (Verbeek 2017). Empathic responses might consequently be staged. Should we really count teased responses among genuine instances of good? I think the short answer is: sometimes. This bullet is easier to bite when we realize that this problem is already with us: I could just as well have ulterior motives in baiting your empathic response directly—we do not need to wait around for robots to embody shady motives to bring out the problem. But robots as proxies for this problem definitely magnifies and complexifies it. But I

do not think this detracts from the principally goodness of sovereign expressions, but rather show how we can be quite elaborate at hindering them. Granted, some scenarios might render empathic responses to robots inappropriate, mistaken, or even harmful.

One reason for tempering our empathic responses to robots is if they divert emotional resources away from comparably more valuable human relationships. Turkle (2011) has been a frontrunner in championing the concern that if robots eventually provide an easier and more agreeable company, humans might increasingly opt for the fake appropriation only to end up lonely.[26] I share her sentiment: humans need humans. But I also think we need more research on the long-term effect of relationships with robots to glean whether Turkle might be right in this bleak assessment, as briefly discussed in Sect. 2. If accepting the robot as a moral patient-for-the-user has adverse effects on the user's wider human relations or otherwise disrupts human community and societies, we might want to regulate robot design to make it actively deter users from responding empathically—especially if true third-person beneficiaries are around.

Mentioning these examples is of course just scratching the surface of all the relevant aspects to consider in a broader societal application of moral patiency for robots. Beyond the present scope also lie legal conceptions of patienthood that are vital to consider in this eventuality. In this context some momentum has gathered in favor of extending some kind of personhood for robots (e.g. Gellers 2020). And while the two are distinct fields, we should like our legal framework and moral conviction to converge to some degree.

As a contribution to this wider discussion, I have in this paper developed an argument in favor of taking robots as moral patients on the basis of one particular ethical framework. Since the social affordances of robotic artefacts seem to allow for sovereign expressions of life to emerge for the benefit of the human agent, individual robots are acceptably moral patients in relation to individual humans.

**Acknowledgements** The author would like to thank Ulrik Nissen, Raffaele Rodogno and Jakob Donskov and the blind peer reviewers for helpful suggestions on earlier versions of the manuscript.

**Funding** Not applicable.

[26] Nowhere is this point illustrated as well as in the discussion on sex robots. It is often argued that erotic partners designed to always accommodate and mirror the users every fantasy is little more than self-gratification. If 'no one's home' and we simply stare into our own reflection, will very agreeable social robots after all contribute to our incurvature rather than displacing it and opening us up toward the world? If we really just respond to an echo of our own reflection when interacting with robots, such activity amount to nothing more than, in Løgstrup terminology, self-encircling feelings and motions rather than sovereign expressions.

## Declarations

**Conflict of interest** The author declares no known conflict of interests.

## References

Airenti G (2015) The cognitive bases of anthropomorphism: from relatedness to empathy. Int J Soc Robot 7(1):117–127. https://doi.org/10.1007/s12369-014-0263-x

Basl J (2019) The death of the ethic of life. Oxford University Press, Oxford

Basl J, Bowen J (2020) AI as a moral right-holder. In: Dubber MD, Pasquale F, Das S (eds) The Oxford handbook of ethics of AI. Oxford University Press, pp 288–306

Bryson JJ, Kime PP (2011) Just an artifact: why machines are perceived as moral agents. Proc Twenty-Second IntJtConfArtifIntellBarc Catalonia Spain 16–22:1641–1646. https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-276

Cappuccio ML, Peeters A, McDonald W (2019) Sympathy for dolores: moral consideration for robots based on virtue and recognition. PhilosTechnol 33(1):9–31. https://doi.org/10.1007/s13347-019-0341-y

Chappell T (2011) On the very idea of criteria for personhood: criteria for personhood. South J Philos 49(1):1–27. https://doi.org/10.1111/j.2041-6962.2010.00042.x

Clark C (1987) Sympathy biography and sympathy margin. Am J Sociol 93(2):290–321

Coeckelbergh M (2014) The moral standing of machines: towards a relational and non-Cartesian moral hermeneutics. PhilosTechnol. https://doi.org/10.1007/s13347-013-0133-8

Coeckelbergh M (2018) Why care about robots? empathy, moral standing, and the language of suffering. Kairos J PhilosSci 20(1):141–158. https://doi.org/10.2478/kjps-2018-0007

Coeckelbergh M (2020) How to use virtue ethics for thinking about the moral standing of social robots: a relational interpretation in terms of practices, habits, and performance. Int J Soc Robot. https://doi.org/10.1007/s12369-020-00707-z

Coeckelbergh M, Gunkel D (2014) Facing animals: a relational, other-oriented approach to moral standing. J Agric Environ Ethics 27:715–733. https://doi.org/10.1007/s10806-013-9486-3

Crowell CR, Deska JC, Villano M, Zenk J, Roddy JT Jr (2019) Anthropomorphism of robots: study of appearance and agency. JMIR Hum Factors 6:2. https://doi.org/10.2196/12629

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

Danaher J (2019) Welcoming robots into the moral circle: a defence of ethical behaviourism. SciEng Ethics. https://doi.org/10.1007/s11948-019-00119-x

Darling K (2016) Extending legal protection to social robots: the effects of anthropomorphism, empathy, and violent behavior towards robotic objects. In: Calo R, Froomkin A, Kerr I (eds) Robot Law. Edward Elgar Publishing, pp 213–232

Davis M (1983) Measuring individual differences in empathy: evidence for a multidimensional approach. J PersSocPsychol 44(1):113–126

Dennett D (1998) Brainstorms: philosophical essays on mind and psychology. MIT Press, Cambridge

Donath J (2020) Ethical issues in our relationship with artificial entities. In: Dubber MD, Pasquale F, Das S (eds) The oxford handbook of ethics of AI. Oxford University Press, pp 51–73

Eberl JT (2017) The ontological and moral significance of persons. Sci Fides 5(2):217. https://doi.org/10.12775/SetF.2017.016

Eck D, Levine A (2017) Prioritizing otherness: the line between vacuous individuality and hollow collectivism. In: Hakli R, Seibt J (eds) Sociality and normativity for robots. Springer International Publishing, Cham, pp 67–87
Epley N, Waytz A, Cacioppo J (2007) On seeing human: a three-factor theory of anthropomorphism. Psychol Rev 114:864–886. https://doi.org/10.1037/0033-295X.114.4.864
Fink H (2017) What is ethically demanded?: K. E. Løgstrup's philosophy of moral life. University of Notre Dame Press, Notre Dame
Gamez P, Shank DB, Arnold C, North M (2020) Artificial virtue: the machine question and perceptions of moral character in artificial moral agents. AI Soc 35(4):795–809. https://doi.org/10.1007/s00146-020-00977-1
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. https://doi.org/10.1016/j.neuroimage.2007.02.003
Gellers J (2020) Rights for robots: artificial intelligence, animal and environmental law, 1st edn. Routledge, Milton Park, Abingdon, Oxon, New York
Gordon J-S (2020) What do we owe to intelligent robots? AI Soc 35(1):209–223. https://doi.org/10.1007/s00146-018-0844-6
Graaf MMA de, Allouch SB (2016) Anticipating our future robot society: the evaluation of future robot applications from a user's perspective. In: 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). pp 755–762
Gunkel D (2014) A vindication of the rights of machines. PhilosTechnol 27(1):113–132. https://doi.org/10.1007/s13347-013-0121-z
Gunkel D (2017) The other question: can and should robots have rights? Ethics InfTechnol 20(2):87–99. https://doi.org/10.1007/s10676-017-9442-4
Gunkel D (2018) Robot rights. MIT Press, London
Gunkel DJ, Bryson J (2014) Introduction to the special issue on machine morality: the machine as moral agent and patient. PhilosTechnol 27(1):5–8. https://doi.org/10.1007/s13347-014-0151-1
Kertész C, Turunen M (2017) What can we learn from the long-term users of a social robot? Lect Notes ComputSciSubserLect Notes ArtifIntellLect Notes Bioinforma 10652:657–665. https://doi.org/10.1007/978-3-319-70022-9_65
Krach S, Hegel F, Wrede B, Sagerer G, Binkofski F, Kircher T (2008) Can machines think? Interaction and perspective taking with robots investigated via fMRI. PLoS ONE 3(7):e2597. https://doi.org/10.1371/journal.pone.0002597
Krämer NC, der Pütten AMR, Hoffmann L (2015) Social effects of virtual and robot companions. In: ShyamSundar S (ed) The handbook of the psychology of communication technology. John Wiley & Sons, Ltd, pp 137–159
Leite I, Martinho C, Paiva A (2013) Social robots for long-term interaction: a survey. Int J Soc Robot 5(2):291–308. https://doi.org/10.1007/s12369-013-0178-y
Löffler D, Hurtienne J, Nord I (2019) Blessing Robot BlessU2: a discursive design study to understand the implications of social robots in religious contexts. Int J Soc Robot. https://doi.org/10.1007/s12369-019-00558-3
Løgstrup KE (1966) SartresogKierkegaardsskildringaf den dæmoniskeindesluttethed. Vindrosen 13:28–42
Løgstrup KE (2010) Den etiskefordring. Klim, Aarhus
Løgstrup KE (2013) Opgør med Kierkegaard, 4th edn. Klim, Aarhus
Løgstrup KE (2014) Etiskebegreberogproblemer. Klim, Aarhus
Løgstrup KE (2015) SkabelseogTilintetgørelse. Metafysik IV: Religionsfilosofiskebetragtninger, 4th edn. ForlagetKlim, Aarhus
Maibom HL (2014) Introduction: (almost) everything you ever wanted to know about empathy. In: Maibom HL (ed) Empathy and morality. Oxford University Press, pp 1–40
McCurry J (2018) Japan: robot dogs get solemn Buddhist send-off at funerals. In: the Guardian. http://www.theguardian.com/world/2018/may/03/japan-robot-dogs-get-solemn-buddhist-send-off-at-funerals. Accessed 2 Sep 2020
Misselhorn C (2009) Empathy with inanimate objects and the uncanny valley. Minds Mach 19(3):345–359. https://doi.org/10.1007/s11023-009-9158-2
Neely E (2014) Machines and the moral community. PhilosTechnol 27(1):97–111. https://doi.org/10.1007/s13347-013-0114-y
Niekerk K (2017) Løgstrup's conception of the sovereign expressions of life. In: Fink H, Stern R (eds) What is ethically demanded?: K. E. Logstrup's philosophy of moral life. University of Notre Dame Press, pp 186–215
Nyholm S (2020) Humans and robots: ethics, agency, and anthropomorphism. Rowman and Littlefield International, London, New York
Perez-Osorio J, Wykowska A (2020) Adopting the intentional stance toward natural and artificial agents. PhilosPsychol 33(3):369–395. https://doi.org/10.1080/09515089.2019.1688778
Rabjerg B (2017) Løgstrup's ontological ethics. An analysis of human interdependent existence. Res Cogitans 12(1):93–110
Rabjerg B, Stern R (2018) Freedom from the Self: Luther and Løgstrup on Sin as "Incurvatus in Se." Open Theol 4(1):268–280. https://doi.org/10.1515/opth-2018-0020
Redstone J (2016) Making sense of empathy with sociable robots: a new look at the "imaginative perception of emotion." Social robots: boundaries, potential, challenges. Routledge, London, pp 19–39
Riek LD, Rabinowitch T-C, Chakrabarti B, Robinson P (2009) How anthropomorphism affects empathy toward robots. In: Proceedings of the 4th ACM/IEEE international conference on Human robot interaction. Association for Computing Machinery, La Jolla, California, USA, pp 245–246
Rodogno R (2016) Robots and the limits of morality. In: Nørskov M (ed) Social robots: boundaries, potential, challenges, 1st edn. Ashgate, Farnham, Surrey, UK, Burlington
Rosenthal-von der Pütten AM, Krämer NC (2015) Individuals' evaluations of and attitudes towards potentially uncanny robots. Int J Soc Robotics 7(5):799–824. https://doi.org/10.1007/s12369-015-0321-z
Rosenthal-von der Pütten AM, Krämer NC, Hoffmann L, Sobieraj S, Eimler SC (2013) An experimental study on emotional reactions towards a robot. Int J Soc Robot 5(1):17–34. https://doi.org/10.1007/s12369-012-0173-8
Rosenthal-von der Pütten AM, Schulte FP, Eimler SC, Sobieraj S, Hoffmann L, Maderwald S, Brand M, Krämer NC (2014) Investigations on empathy towards humans and robots using fMRI. Comput Hum Behav 33:201–212. https://doi.org/10.1016/j.chb.2014.01.004
Searle JR (1980) Minds, brains, and programs. Behav Brain Sci 3(3):417–424. https://doi.org/10.1017/S0140525X00005756
Seibt J, Rodogno R (2019) Understanding emotions and their significance through social robots, and vice versa. Techne Res PhilosTechnol 23:257–269. https://doi.org/10.5840/techne2019233104
Singer P (2011) The expanding circle: ethics, evolution, and moral progress, 1st Princeton University Press pbk. Princeton University Press, Princeton
Smedegaard CV (2019) Reframing the Role of Novelty within Social HRI: from Noise to Information. In: 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI). pp 411–420
Sparrow R (2020) Virtue and vice in our relationships with robots: is there an asymmetry and how might it be explained? Int J Soc Robot. https://doi.org/10.1007/s12369-020-00631-2
Stern R (2019) The radical demand in logstrup's ethics. Oxford University Press, New York

Stokes P (2016) The problem of spontaneous goodness: from Kierkegaard to Løgstrup (via Zhuangzi and Eckhart). ContPhilos Rev 49(2):139–159. https://doi.org/10.1007/s11007-016-9377-1

Suzuki Y, Galli L, Ikeda A, Itakura S, Kitazaki M (2015) Measuring empathy for human and robot hand pain using electroencephalography. Sci Rep 5(1):1–9. https://doi.org/10.1038/srep15924

Thornton S (2020) Ontology and ethics: Løgstrup between heidegger and levinas. Monist 103(1):117–134. https://doi.org/10.1093/monist/onz030

Tisseron S, Tordo F, Baddoura R (2015) Testing empathy with robots: a model in four dimensions and sixteen items. Int J Soc Robot 7(1):97–102. https://doi.org/10.1007/s12369-014-0268-5

Torrance S (2008) Ethics and consciousness in artificial agents. AI Soc 22:495–521. https://doi.org/10.1007/s00146-007-0091-8

Torrance S (2014) Artificial consciousness and artificial ethics: between realism and social relationism. PhilosTechnol 27(1):9–29. https://doi.org/10.1007/s13347-013-0136-5

Turing AM (1950) Computing machinery and intelligence. Mind LIX(236):433–460. https://doi.org/10.1093/mind/LIX.236.433

Turkle S (2011) Alone together: why we expect more from technology and less from each other. Basic Books, New York

Ugazio G, Majdandžić J, Lamm C (2014) Are empathy and morality Linked? In: Maibom HL (ed) Empathy and morality. Oxford University Press, pp 155–171

Vallor S (2015) Moral deskilling and upskilling in a new machine age: reflections on the ambiguous future of character. PhilosTechnol 28(1):107–124. https://doi.org/10.1007/s13347-014-0156-9

Vallor S (2016) Technology and the virtues: a philosophical guide to a future worth wanting. Oxford University Press, Oxford

van Camp J (2019) My jibo is dying and it's breaking my heart | WIRED. https://www.wired.com/story/jibo-is-dying-eulogy/. Accessed 2 Sep 2020

VanderMaas J (2015) hitchBOT USA tour comes to an early end in Philadelphia. http://cdn1.hitchbot.me/wp-content/uploads/2015/08/hitchBOT-USA-Trip-End-Press-Release-FINAL.pdf. Accessed 1 Sept 2020

Verbeek P-P (2017) Designing the morality of things: the ethics of behaviour-guiding technology. In: van den Hoven J, Miller S, Pogge T (eds) Designing in Ethics, 1st edn. Cambridge University Press, pp 78–94

Wang Y, Quadflieg S (2015) In our own image? Emotional and neural processing differences when observing human–human vs human–robot interactions. SocCogn Affect Neurosci 10(11):1515–1524. https://doi.org/10.1093/scan/nsv043

Wolf J (2017) Phenomenology in Løgstrup's Creation Theology. In: Gregersen NH, Uggla BK, Wyller T (eds) Reformation theology for a post-secular age løgstrup, prenter, wingren, and the future of scandinavian creation theology. Vandenhoeck & Ruprecht, Göttingen

Young JE, Sung J, Voida A, Sharlin E, Igarashi T, Christensen HI, Grinter RE (2011) Evaluating human-robot interaction. Int J Soc Robot 3(1):53–67. https://doi.org/10.1007/s12369-010-0081-8