AI & SOCIETY (2020) 35:209–223
https://doi.org/10.1007/s00146-018-0844-6

OPEN FORUM

What do we owe to intelligent robots?

John-Stewart Gordon[1]

Received: 29 January 2018 / Accepted: 20 April 2018 / Published online: 27 April 2018

Abstract
Great technological advances in such areas as computer science, artificial intelligence, and robotics have brought the advent of artificially intelligent robots within our reach within the next century. Against this background, the interdisciplinary field of machine ethics is concerned with the vital issue of making robots "ethical" and examining the moral status of autonomous robots that are capable of moral reasoning and decision-making. The existence of such robots will deeply reshape our socio-political life. This paper focuses on whether such highly advanced yet artificially intelligent beings will deserve moral protection (in the form of being granted moral rights) once they become capable of moral reasoning and decision-making. I argue that we are obligated to grant them moral rights once they have become full ethical agents, i.e., subjects of morality. I present four related arguments in support of this claim and thereafter examine four main objections to the idea of ascribing moral rights to artificial intelligent robots.

Keywords Artificially intelligent robots · Moral status · Moral rights · Moral agency · Full ethical agents · Machine rights

1 Introduction

In recent years, the interdisciplinary field of machine ethics—that is, how to program robots with ethical rules, so that they become either implicit or explicit moral agents (Moor 2006)—has become of utmost importance because of current and anticipated technological developments in the fields of computer sciences, artificial intelligence (AI), and robotics (Lin et al. 2014; Wallach and Allen 2010; Anderson and Anderson 2011; Gunkel and Bryson 2014a, b). Machine ethics considers the implications of making artificially intelligent robots (henceforth IRs) "ethical"[1] and examines related issues such as whether IRs have moral status or not, including moral and legal[2] rights. Whether this is even possible is a matter of great debate among researchers working in the field of machine ethics. For example, Johnson and Axinn (2014, p 4) take a critical viewpoint, arguing that autonomous robots only mimic ethics, since they lack "the imagination to conceive of the effects should the principle of their actions be made universal, as well as the free will to make the choice to follow a moral style"; Rodogno (2016, p 12) claims, "The day in which robots fulfil the conditions of moral agency and moral patience outlined here, however, will be the day in

✉ John-Stewart Gordon
jostgo76@gmail.com

[1] Department of Philosophy and Social Critique, Faculty of Political Science and Diplomacy, Vytautas Magnus University, V. Putvinskio g. 23, 44243 Kaunas, Lithuania

[1] Susan Anderson (2011a, b, pp 22) defines the goal of machine ethics as "to create a machine that follows an ideal ethical principle or set of principles in guiding its behaviour; in other words, it is guided by this principle, or these principles, in the decisions it makes about possible courses of action it could take. We can say, more simply, that this involves 'adding an ethical dimension' to the machine."

[2] One can read the following interesting legal development in the Preliminary Draft Report of UNESCO's World Commission on the Ethics of Scientific Knowledge and Technology (COMEST) on Robotic Ethics: "The Committee on Legal Affairs of the European Parliament, in its 2016 *Draft Report with Recommendations to the Commission on Civil Law Rules on Robotics*, already considers the possibility of 'creating a specific legal status for robots, so that at least the most sophisticated autonomous robots could be established as having the status of electronic persons with specific rights and obligations, including that of making good any damage they may cause, and applying electronic personality to cases where robots make smart autonomous decisions or otherwise interact with third parties independently' (p. 12)" (Draft Report United Nations 2016, p 26). See, also Malvaux's Report with recommendations to the Commission on Civil Law Rules on Robotics (Delvaux 2017) and Calverley (2011) for an interesting discussion on the idea of ascribing legal rights to machines.

which they will be emotive creatures capable of humanly recognizable interests".[3]

The ethical discourse around robots has become quite sophisticated since the days of Isaac Asimov's Three Laws of Robotics[4] (in *Runaround*, Asimov 1942), subsequently complemented by his so-called Zeroth Law of Robotics (in *Robots and Empire*, Asimov 1986), which state that a robot may not harm humanity or, by inaction, allow humanity to suffer harm.[5] The idea that IRs may rebel against their creators and cause the global extinction of humanity, or at least of most human beings, has been prominently portrayed in the literature and in movies for nearly a century, dating back to Karel Capek's famous science fiction play *Rossum's Universal Robots* (1920).[6] Whether IRs will eventually rebel against human beings remains to be seen. In any case, however, considering their potential moral status is of utmost importance given their expanding involvement in many areas of human life by virtue of the simple fact that they will *interact* with human beings. For example, healthcare robots now provide physical assistance and companionship, as well as monitoring health and safety; military robots such as intelligent drones make decisions in warfare, and other robots perform surveillance, educational, and work tasks. They can cause harm if their autonomous actions are not restricted by any moral safeguards in the course of such interaction (e.g. Picard 1997),[7] which will only become increasingly complex and will deeply influence their place in society and how human beings view them. It seems clear that, once IRs have reached the level of being moral machines, the important issue of their moral status must be addressed as well—i.e., whether IRs should be seen as full moral agents, capable of moral reasoning and decision-making, or as "moral patients", incapable of moral reasoning and decision-making but vulnerable beings that can be treated immorally (see Floridi 2011).[8]

This paper is motivated by three different but related ideas: (1) that this topic is a worthy thought experiment in moral philosophy, (2) that we can gain knowledge about ethical theories by trying to develop ethical systems for robots, and (3) that we should be preparing for the socio-political changes that can be expected as highly advanced and, perhaps, even autonomously intelligent robots become a reality. One might object that the advent of such robots is too speculative to justify a serious moral analysis of this issue, but there are at least two answers to that objection. First, scepticism regarding the possibility of highly advanced, intelligent robots seems quite premature given the great technological advances and future prospects in robotics, AI, and computer science. Second, it is unwise and un-philosophical to abstain from discussing important moral and socio-political issues the emergence of which cannot be fully ruled out in advance. In what follows, however, I am concerned only with the vital issue of IRs' moral status, and in particular the important idea of ascribing moral rights to IRs, which depends on the empirical question of whether and how IR may become moral machines. Here, based on the enormous prospects for future technological developments, I take it for granted that IRs will become moral machines in the future. In what follows, I attempt to show that if IRs are capable of moral reasoning and decision-making on a level that is comparable with the moral agency of human beings, then one must see IRs not only as moral patients, but also as full moral agents with corresponding moral rights (Sullins 2011).[9] Most

[3] Johnson and Axinn (2014, p 1) and Rodogno (2016, p 1) admit, however, that their reasoning refers only to the present-day robots, and that it remains conceivable that robots might become full moral agents in the future.

[4] Asimov's initial Three Laws of Robotics are as follows: (1) a robot may not injure a human being or, through inaction, allow a human being to come to harm; (2) a robot must obey the orders given it by human beings except where such orders would conflict with the first law, and (3) a robot must protect its own existence as long as such protection does not conflict with the first or second law.

[5] For a thorough discussion on the logic and problems of Asimov's four laws of robotics, see Clark (2011, pp 254–284), who examines the complex issues that arise when robots are supposed to follow these laws. For another critical examination of Asimov's laws as the foundation of machine ethics, see Anderson (2011a, pp 285–296). By examining the robot Andrew in Asimov's short story "The Bicentennial Man", she provides important insights on the ethical status of the laws. The up-shot is that the laws are ethically inappropriate for intelligent beings such as Andrew, who is considered to act more ethically than most human beings.

[6] Similar ideas have been depicted in movies such as *2001: A Space Odyssey* (Stanley Kubrick, 1968), in which the spaceship's HAL 9000 computer attempts to kill its crew on the way to Jupiter, and in *The Terminator I–V* (James Cameron and Alan Taylor, 1984–2015), where the machines rebel against human beings. Similarly, in the famous Matrix Trilogy (the Wachowski brothers, 1999–2003), sentient machines subdue the human population by keeping them in a dream world while using their bodies as an energy source.

[7] "The greater the freedom of a machine, the more it will need moral standards" (Picard 1997, p 19).

[8] Grau (2011, p 458) correctly claims, "Once we do venture into the territory of robots that are similar to humans in morally relevant respects, however, we will need to be very careful about the way they are treated. Intentionally avoiding the creation of such robots may well be the ethical thing to do, especially if it turns out that the works performed by such machines could be performed equally effectively by machines lacking morally relevant characteristics."

[9] Gibilisco (2003, pp 268–270) distinguishes five generations of robots according to their particular capabilities: (1) robots that are mechanical, stationary, fast, physically rugged, based on servomechanisms, but without external sensors and AI (before 1980); (2) robots that are programmable (by virtue of microcomputer control), having vision and tactile systems, position, and pressure sensors (1980–1990); (3) robots that are mobile and autonomous, able to recognize and synthesize speech, having incorporated navigation systems or tele-operated, and possessing AI (mid-1990s and after); (4) and (5) speculative robots of the future that are able to reproduce, have a sense of humour, etc. (see also Preliminary Draft of COMEST on Robotics Ethics, 2016: 4.)

non-philosophers who work in the robotics, AI, and computer science seem to believe that, if intelligent robots are substantially human-like in their capabilities, then we must also grant them moral rights. This view, however, is highly controversial[10] among those philosophers and ethicists who work in the field of ethics of technology, and therefore, one must provide a more substantial reasoning to justify moral rights for robots, as I will attempt to do below.

After this introduction, the second brief section provides some preliminary information concerning the issue of moral rights for machines. The third section offers a comprehensive account of ascribing moral rights to IR once they are fully autonomous and capable of moral reasoning and decision-making. This account is buttressed by a thorough discussion of several prominent objections in the fourth section. My responses to the objections will provide some additional information on how one should view IRs and their relationships with human beings in the future. They are followed by brief concluding remarks.

2 Machines and moral rights

> "A full ethical agent can make explicit ethical judgements and generally is competent to reasonably justify them. An average adult human is a full ethical agent. We typically regard humans as having consciousness, intentionality, and free will. Can a machine be a full ethical agent? It's here that the debate about machine ethics becomes most heated. Many believe a bright line exists between the senses of machine ethics discussed so far and a full ethical agent. For them, a machine can't cross this line. The bright line marks a crucial ontological difference between humans and whatever machines might be in the future" (Moor 2006, p 20).[11]

The trajectory of technological development seems, indeed, quite promising and there is no principal reason why one should not consider the possibility of machines that will become full ethical agents. In one of the most interesting and morally challenging episodes (number 35) of *Star Trek: The Next Generation* (1989), Gene Roddenberry raises the simple but complex question of who owns Data, the artificial intelligent android serving on the starship Enterprise. Given Data's capabilities—he is a full ethical agent in Moor's terms—should he be allowed to make his own decisions, such as to quit his job and leave the starship crew to avoid the possibility of suffering harm in a dangerous experiment in which he has been ordered to participate? Data is, in general, portrayed as not just a moral patient but also a moral agent, and, therefore, entitled to moral rights comparable to those of human beings (including, as the episode shows, the confirmed moral and legal right to quit his job). The main difference between Data and a human person, however, is that he is artificial. Of course, at this moment, there is no Data or any other comparable IR with similar capabilities, but we might see one within this century. Given the meaning and importance of, for example, animal rights (Donaldson and Kymlicka 2013; Francione 2009) and environmental rights (Atapattu 2015) with respect to the further development of our moral reasoning, one should also be prepared to think about moral rights—or even "human" rights—for IRs once they fulfil certain particular features. History has shown us, as, forcefully demonstrated by Darwin's theory of evolution, that the proclaimed *substantial* difference between animals and human beings is mistaken. Any morally relevant empirical differences[12] that people have claimed to be of utmost importance for ascribing moral status and rights come, as a matter of fact, in degrees. If machines attain a capability of moral reasoning and decision-making that is comparable to the moral agency of human beings, they then should be entitled to the status of full moral (and legal) agents, equipped with full moral standing and related rights (Sullins 2011; for a contrary view, see; Torrance 2005)[13].[14]

In his classic paper "When Is a Robot a Moral Agent?", Sullins (2011) argues for giving intelligent robots moral rights once they fulfil three necessary requirements of

[10] This debate can be compared to the heated debates over abortion, animal rights, and environmental rights over the past few decades, in that it is by no means clear that possessing similar capabilities to human beings should eventually lead to the granting of moral rights to robots.

[11] The notion of an "ethical agent" amounts to the equivalent of what is commonly called a "moral agent" in the context of ethics and moral philosophy. The notion of *moral* denotes other people's interests and deontological constraints, whereas *ethics* usually refers to one's own individual interests and well-being. For a more detailed depiction, see Gordon (2013).

[12] For example, reasoning (and intelligent behaviour), autonomous decision-making, feeling pain, having identifiable personal interests, and the desire to continue one's life, emotion, etc. Or consider the famous list of items provided by Warren (1973) in the context of abortion and personhood: sentience, emotionality, reason, the capacity to communicate, self-awareness, and moral agency.

[13] In "A Robust View of Machine Ethics" (2005), Torrance argues that even if IRs share the same features that define human beings as moral agents, robots will, nonetheless, have no "intrinsic moral status", because they are non-organic. Only "genuinely sentient" beings who are organic by nature deserve our "moral concern or moral appraisal".

[14] Johnson and Axinn (2014, p 2) hold the contrasting view that "a person has not only rights, duties, free will, but also the imagination to understand the effect of different actions, and the ability to impose on him or herself the categorical imperative. How close do robots come to the features of a human person, the features that make for moral motivation and moral action? Such robots (i.e., robots that lack free will and imagination) certainly do not have rights". I will respond to their claims in the section on objections below.

robotic moral agency: (1) significant autonomy, i.e., "the machine is not under the direct control of any other agent or user" (158), (2) behaving intentionally,[15] and (3) being in a position of responsibility[16] (157–160). Sullins concludes his line of reasoning with the following remarks:

> Robots are moral agents when there is a reasonable level of abstraction under which we must grant that the machine has autonomous intentions and responsibilities. If the robot can be seen as autonomous from many points of view, then the machine is a robust moral agent, possibly approaching or exceeding the moral status of human beings. Thus, it is certain that if we pursue this technology, then, in the future, highly complex, interactive robots will be moral agents with corresponding rights and responsibilities. Yet, even the modest robots of today can be seen to be moral agents of a sort under certain, but not all, levels of abstraction and are deserving of moral consideration (Sullins 2011, p 160).

I agree with Sullins's conclusion, except for his view that "even the modest robots of today can be seen to be moral agents" (160). The concept of moral agency—including *robotic* moral agency—is more sophisticated than Sullins suggests (as I discuss below in Sects. 3, 4), even though one must admit that, for example, the idea of free will in the context of autonomy and the assumption of a strong version of intentionality[17] in human behaviour are completely unresolved issues of human agency and, therefore, should not be applied to robots, as Floridi and Sanders (2004) have rightly argued.

3 Should IRs enjoy moral rights?

The previous section was particularly concerned with how machines should act towards human beings. This section, examines the related question of whether one should ascribe moral rights to machines once they have become full ethical agents. The first part of this section contains some general remarks on the proper status of IRs in their initial situation; the second part offers some morally relevant observations on why one should take the idea of moral rights for IRs seriously (for similar and related issues, see Wallach and Allen 2010; Anderson and Anderson 2011; Gunkel and Bryson 2014a, b).[18]

3.1 The initial situation

The history of ethics is a history of the ascription of moral rights and duties (see Singer 2011; Gunkel and Bryson 2014a, p 6; Gunkel 2014, p 115). It started with a focus on men of a certain particular social status, and then, it extended moral rights to include women as well as children, eventually expanding further to include (self-conscious) animals and, finally, unconscious nature. This process, of course, was by no means straightforward and linear; there have been many sudden shifts, steps backward, and moments of radical social change along the way. In recent history, views of moral rights and duties have been influenced by such powerful forces as the African–American rights movement in the US, the movement for gender equality, the animal rights movement, the gay movement, and the disability rights movement. Alongside and interacting with these movements, we have seen the evolution of international human rights legislation, the moral and legal *lingua franca* of modern times. Furthermore, it is claimed not that all human beings do in fact practically enjoy the same moral and legal rights, but that they theoretically should enjoy these universal rights, simply because they are members of the human species. Given this moral development—which also includes, to some extent, the protection of animals and nature—it seems useful to consider IRs as potentially members of our moral community, based on their presumed capacity to reason and to make moral decisions. There are, at least, three main views: (1) the optimistic view that states that robots may not be moral agents at this moment, but that they may become so in the future, because there is no principle reason why robots should not become moral agents in the future (Dennett 1998), (2) the pessimistic view claims that robots will never become moral agents since they will never possess an autonomous free will (Bringsjord 2008),

[15] Here, Sullins adheres to Floridi's idea of avoiding issues related to free will and intentionality with respect to IRs, because they are unresolved problems in human behaviour as well and hence should not be necessary conditions for ascribing moral agency to robots. Sullins (2011, p 158) states, "If the complex interaction of the robot's programming and environment causes the machine to act in a way that is morally harmful or beneficial and the actions are seemingly deliberate and calculated, then the machine is a moral agent."

[16] If the robot behaves in a way that suggests that "it has a responsibility to some other moral agent(s), [we can ascribe moral agency to a robot]" and "[i]f the robot behaves in this way, and if it fulfils some social role that carries with it some assumed responsibilities, and if the only way we can make sense of its behaviour is to ascribe to it the 'belief' that it has the duty to care for its patients, then we can ascribe to this machine the status of a moral agent" (Sullins 2011, p 159).

[17] On the contrary, for example, Floridi (2011, p 200) argues that an intentional state is not necessary for moral agency, since assessing this feature presupposes a so-called "privileged access" to a person's mental state, which is theoretically possible but practically unachievable. Therefore, the view that to be a moral agent, the artificially intelligent being "must relate itself to its actions in some more profound way, involving meaning, wishing, or wanting to act in a certain way and being epistemically aware of its behaviour" (200), is unnecessary.

[18] For a true manifesto for treating robots morally, see Hall (2011, pp 32–33).

and (3) the deficient human view that says that human beings are not moral agents, but robots are because an action is free if and only if it is based on fully thoughtful reasons using strict logic (Nadeau 2006).[19] How will the relation between human beings and machines develop as a whole? This question raises important socio-political issues that we must address sooner or later (see Lin et al. 2014). Should it be possible to marry an artificially intelligent android and to create a robot partner according to one's own idiosyncratic preferences (Levy 2007)? How will this influence the concept of family in the future? Is it morally acceptable to use IRs—without their consent—in workplaces or for other duties, such as in hospitals, care homes for elderly people, civil services, prostitution, and war? Should one protect IRs against any forms of exploitation? If all of this could be possible at some future point, why should they not enjoy moral rights? What are the morally relevant differentia specifica between human beings and IRs? Or is being human already a sufficient condition for ascribing moral rights (for a contrary view, see Singer 2009)? Such questions have already been the subject of great controversy among researchers in different disciplines and will certainly remain of interest in the technological era to come. Whether IRs will face a long struggle for recognition and understanding remains to be seen. The following discussion offers some important arguments in favour of moral rights for IRs.

3.2 The concept of personhood and the moral agency/patiency divide

This part provides a more general background for understanding and appreciating the arguments in support of moral rights for intelligent robots by focussing on the concept of personhood and the moral agency/patiency divide as understood in this debate. Tying together, some of the already mentioned considerations and preparing the ground for the following discussion will help us to better understand the principles underlying the opposing moral views on ascribing moral rights to robots.

It seems that even if robots had the same (morally relevant) capabilities as human beings, many people (including philosophers and ethicists) would still be reluctant to treat them in morally similar ways and to acknowledge their full moral standing. This is because human beings tend to ascribe full moral rights only to fellow human beings. Traditionally, personhood is tied to moral agency and used to confer moral rights on human beings. Therefore, only human beings who are moral agents enjoy the full protection of moral rights, and are considered as moral equals. The concept of personhood, traditionally associated with autonomy and rationality (e.g., Kant 2009; Singer 1979), is the key to understanding the inner logic of this ascription of moral rights and whether certain beings are considered full moral agents or, rather, moral patients with a less high moral standing.

The traditional understandings of personhood have been severely challenged by at least five significant contemporary movements:

1. the feminist movement, stressing moral equality between the sexes;
2. the gay and lesbian movement, stressing moral equality between people of different sexual orientation;
3. the animal rights movement, stressing the importance of sentience as requiring moral behaviour towards non-human living beings;
4. the disability rights movement, stressing moral equality among persons regardless of physical or mental ability;
5. the environmental rights movement, stressing the importance of non-human natural life.

All five movements criticize the traditional concepts of moral status and attempt to replace them with a new definition of personhood that increases the number of members of the moral community.[20] That is important, because only members of the moral community enjoy moral protection. First, women and people of gay or lesbian sexual orientation have historically not been treated as morally equal or granted the same moral rights as heterosexual men (this is still true in many countries around the world). Second, animals, human foetuses, and mentally impaired human beings have, in many or most cases, not been considered persons and, therefore, do not enjoy the same moral rights as "normal" adult human beings. They are commonly considered as moral patients, not moral agents. Indeed, there is currently no uniform definition of personhood. Elsewhere, I have suggested that we abstain from using the concept of personhood to ascribe moral rights, because reaching agreement on the necessary and sufficient criteria for personhood seems impossible (Gordon 2017).

Against this background, it may seem inconceivable that most people—including philosophers and ethicists—would ever accept the idea that intelligent robots should have the same moral rights as human beings due to their comparable capabilities. The great struggle of the five above-mentioned movements substantiates this concern. Some authors, such

[19] See, also Sullins's (2011, pp 155–157) considerations on the moral agency of robots.

[20] The concept of personhood and the limits of moral agency and patiency have been thoroughly discussed by Hernandez-Orallo (2017, Chaps. 16–18) and by Altman (2011), who examines the key notions with respect to Kant's position.

as Bryson (2010), believe that we should view robots as slaves and use them simply as tools. Thus, even if we succeed in creating intelligent robots with human-like abilities, there may still be significant opposition to granting them moral rights, based on the traditional idea of human personhood. But what about moral patiency for machines? We currently, at least to some extent, have accepted the ideas that (1) some animals have a high moral status and enjoy some moral rights as moral patients and (2) the environment should also be protected, because it has a moral standing, as well. The underlying reasoning, depending on the particular arguments adopted, is that animals and nature have either instrumental or intrinsic value. Applying the same line of argumentation, intelligent robots may, then, also enjoy some moral rights as moral patients (though not as moral agents, even if they possess human-like abilities). Of course, this would be a violation of the moral principle that similar cases should be treated in similar ways, but our general tendency towards speciesism is a widespread phenomenon and quite difficult to overcome (Singer 2009). Many people draw a firm line between the natural and the artificial with respect to ascribing moral status and moral rights.

Gunkel (2012), in an excellent book that reviews relevant works from both the continental and analytical traditions, considers whether machines should be seen as moral agents (first chapter) or moral patients (second chapter). Gunkel believes that both perspectives fall short of properly addressing the "machine question" and suggests deconstructing the binary agent–patent dichotomy with respect to intelligent robots. He calls this new alternative a "Copernican Revolution" to signify that it is a completely novel approach (third chapter). However, Gunkel remains undecided as to whether intelligent machines, once they have become a reality, should be seen as full moral agents equipped with the same moral rights as human beings:

> Should machines like AIs, robots, and other autonomous systems be granted admission to the community of moral subjects, becoming what would be recognized as legitimate moral agents, patients, or both? This question cannot be answered definitively and finally with a simple "yes" or "no". The question will need to be asked and responded to repeatedly in specific circumstances. However, the question needs to be asked and explicitly addressed rather than being passed over in silence as if it did not matter (Gunkel 2012, p 215).

The idea of rethinking the whole classical dichotomy between moral agents and moral patients in the context of the machine question is quite intriguing. However, it seems quite difficult to abstain from using the old concepts, as we can see when we consider Gunkel's question whether robots "would be recognized as legitimate moral agents, patients, or both". If Gunkel really attempts to deconstruct these notions and suggests abolishing the classical dichotomy, then he must also come up with a different way of expressing his position. One possible approach would be to devise a different notion that can function as an umbrella term, such as "moral being". This might be, indeed, an interesting starting point for a new analysis.[21]

Independent of the particular wording, it is striking that Gunkel (2012) and, particularly, Darling (2016) refer to animal rights as an analogy to contemplate the moral status of intelligent robots. According to Darling (2016, p 214), one should think about a "near-future possibility to regulate people's behaviour towards certain types of robots", because misconduct towards social robots—to whom human beings feel strongly attached based on the three vital criteria of their physicality, their perceived autonomous movement, and their social behaviour (Darling 2016, pp 217–218)—may lead to unwelcome consequences. Specifically, people may become traumatized or desensitized when they see social robots abused, mistreated, or otherwise suffering misconduct and cruelty (Darling 2016, p 224). Adhering to Kant's reasoning on how one should treat animals, Darling (2016, pp 227–228) claims that, by analogy, one should also treat social robots morally, because mistreating them may encourage people to engage in misconduct towards fellow human beings, as well. She states, "As mentioned above, if lifelike and alive is subconsciously muddled, then treating certain robots in a violent way could desensitize actors toward treating living things similarly" (Darling 2016, p 231). Therefore, she argues, one should think about implementing a legal framework for social robots consistent with that governing animal abuse. This step should be taken even if social robots lack equivalent capacities to human beings. The underlying idea seems to be that one should also protect lifelike beings "when society cares deeply enough" about them (Darling 2016, p 230).

I agree with Darling's approach with respect to the "near-future possibility of regulating people's behaviour towards certain types of robots" based on the above-mentioned possible detrimental consequences for fellow human beings. However, I would add that when the capabilities of intelligent robots have reached a certain level, then one must ascribe full moral rights to them, including the right to life,[22]

[21] I am sympathetic with this novel idea, but, in this paper, I adhere to the classical notion of moral agents and patients, because I believe that intelligent robots—once they exist—should be considered full moral agents. In the following sections, I provide several arguments in support of this claim.

[22] Darling, however, does not entertain the idea of granting intelligent robots the right to life: "Animals themselves are not protected from being put down, but rather only when ending their lives is deemed cruel and unnecessary given the method or circumstances. Similarly, it would make little sense to give robots a 'right to life'" (Darling 2016, pp 229).

based on non-consequentialist reasoning, independently of whether society "cares deeply enough" about them. Whether intelligent robots are, in fact, moral beings is not a matter of popular vote or social opinion. Rather, their moral status depends on empirical factors that determine whether they have attained a level of functioning at which they should be considered morally equal beings. The arguments in support of this view are provided in the next section.

3.3 Arguments in support of moral rights for IRs

The following four different but related reasons may not sufficiently support moral rights for machines if taken individually; in combination, however, they seem to offer a sufficient line of reasoning in favour of moral rights for machines that cannot be dismissed offhandedly. This general line of argumentation will be supported in the next section, where I make additional points in support of moral rights for machines while responding to objections raised against this idea. Together, these two sections substantiate the main position of this paper.

3.3.1 Protection against exploitation

If machines will be capable of rational choice, including the ability to make moral decisions at some point in the future, then it seems crystal clear that they are not only moral patients (i.e., moral objects) but also moral agents (i.e., moral subjects) by virtue of their autonomy and hence deserve to become members of our moral community. Their moral status must then be protected by moral and legal rights (Calverley 2011). For example, to force an artificially intelligent being that is capable of moral reasoning and decision-making to perform actions—e.g., serving in war, engaging in prostitution, or committing illegal deeds—that stand in contrast with his or her own interest is not only immoral by contravening the declared wishes of that being but also becomes a matter of pure exploitation of a being capable of moral reasoning and decision-making.

One might object that a machine, even a highly advanced moral machine, is not free to do what it wants, since it is always (legally) owned by someone else—for example, a human being, an institution, or a company. The underlying rationale for this objection adheres to the claim that the producers or owners of the machine can do whatever they want with it since they legally own it. This line of reasoning is misleading and overlooks at least two counter-arguments. First, the proposed argument suggests that whenever something has been produced, it, therefore, either belongs to the producer or the person who has legally purchased the particular good (independently of whether that good is in fact capable of making its own autonomous moral decisions). If that were the case, however, then human children would also be owned by their parents, since the parents are responsible for having produced them and given them life. Of course, parents are the legal guardians of their children and can, therefore, make decisions on their behalf—within a legal framework that protects children against abuse by their parents—until they are capable of making their own autonomous decisions. In this respect, parents do not own their children and are not allowed to do whatever they want with their children, even though they have "produced" the children. Likewise, following the same line of reasoning, advanced moral machines must be treated as autonomous agents; they have been produced by someone, but they should be considered as free persons who cannot be legally owned by another person, an institution, or a company. As members of the moral community, they should enjoy the same protection against exploitation as do their fellow community members, such as human beings. Admittedly, at this moment, it is hard to conceive of this setting, but on the assumption that IRs will eventually become autonomous moral agents, and then, the above reasoning seems to be correct.

Second, a person cannot be legitimately owned by someone else. This is the modern moral outcome of the institution of slavery, which has been part of the history of human beings since very early times. It has always been the general strategy of a dominating group to deny particular minorities—seen as the other—full, i.e., equal, moral status, or personhood so as to deny them equal moral protection. African Americans, Jews, Sinti and Roma, and women have been among the many victims of this strategy over time. Nowadays, it has become common to ascribe moral rights only to persons (including higher animals, such as the great apes, that are considered as moral persons, as well) and to grant them full moral and legal protection, while other beings, who do not fulfil the suggested criteria of personhood,[23] do not enjoy the same full moral and legal protection.[24] Against this background, however, one should not make the same mistake at some future point and treat autonomous artificial beings as the other by violating their legitimate moral right to be free from any type of suppression.

[23] For example, self-consciousness, consciousness, ability to feel pain, having feelings, perceiving oneself as an entity that exists and has an interest in its future existence, etc.

[24] For problems with this conception, consider the problem of abortion in medical ethics, the moral status of human beings with severe mental impairments in disability studies, and the moral status of animals in the context of the animal rights movement. The idea of linking the very right to exist with certain particular criteria that fulfil the idea of personhood is a contested but widely held position (Gordon 2016; Koch 2004).

3.3.2 Moral protection based on rationality

A related line of argumentation to protect moral machines from exploitation might apply a neo-Kantian approach as follows. Rational beings have dignity insofar as they are capable of acting morally, i.e., autonomously. In other words, autonomy is the foundation of the dignity of each rational being, including human beings (see Kant's *Groundwork*) and—one could further argue—artificially intelligent beings, as well. As Kant rightly claims, beings that have dignity should not be allowed either to exploit themselves or to use other rational beings as mere means only, but should respect their own dignity and that of other rational beings.

One might object that Kant's argument does not apply to machines, since they are clearly not human beings. This objection is misleading and prematurely suggests that Kant's ethics applies only to human beings. On the contrary, Kant's ethics is applicable to all rational beings, independently of their biological origin, and does not rely on any type of moral preference concerning the human species (this is the so-called logocentrism of Kant's ethics).

Another, more complex issue is whether one should be allowed to use rational beings, particularly if they are not humans, for some higher ends of humanity by admitting that even if they are not used as mere means for some contingent personal ends, they could still be used, at least, to accomplish some justified and highly important ends for humanity. For example, one might argue, based on this approach, that IRs (even though they may morally condemn it) are obligated to fight in wars for their nation or the world, because they are better fighters and less valuable than human lives. What should not be tolerated among human beings, namely the making of comparisons of moral value between people or ethnicities—even though this has often been done in the past—might be possible in some limited cases to justify using IRs based on their substitutability. Furthermore, if an IR could be fully substituted, because the "mind" of the machine could be stored externally in case of emergency without any loss of the individual personality (whether this is actually possible, of course, remains an empirical question), then that might justify the use of IRs rather than human beings in dangerous situations, since human beings are not substitutable.

A more challenging response against ascribing moral rights to machines concerns the idea of dignity as the foundation of moral rights. It could be argued that only human beings have dignity, and therefore, machines are not members of the moral community, simply because they lack human dignity. Admittedly, the concept of human dignity, by definition, refers only to human beings and not to other non-human beings. The general idea, however, that dignity as such only concerns human beings is unconvincing, since it has been increasingly acknowledged that animals have dignity as well (Donaldson and Kymlicka 2013). They may not have human dignity, but they do have what can be called animal[25] dignity (which becomes relevant in cases where human beings abuse animals). In a similar fashion, one should ascribe rational dignity to artificially intelligent machines, provided that they are capable of autonomous rational decision-making and part of our world. Yet another possible response is that the notion of dignity is itself a ontologically vague and unclear concept and should be replaced by the notion of autonomy itself, because this is, strictly speaking, the phenomenon that people have in mind in all or most cases in which they refer to human dignity (see Macklin 2003; Gordon 2014). One may disagree with the latter view even if one admits that the notion of dignity is currently not defined in a promising way so as to avoid certain pitfalls, such as proving to be inconclusive in complex cases in bioethics (Cochrane 2010). However, even if we are currently unable to properly define this key ethical term, we do acknowledge clear cases of indignity when we are confronted by such cases as a soldier being dragged behind a car through the streets or animals who are severely abused and confined in small pens without fresh air and proper food (Gordon 2014). These are clear instances of indignity, and sentient moral machines capable of rational reasoning and decision-making should also not be abused in a way that harms their dignity as rational beings, such as by forcing them to commit prostitution or other immoral deeds.

3.3.3 The presumption of equal treatment

Moral rights are commonly ascribed to (human) beings who are persons, i.e., who have personhood. For the sake of argument, let us grant this contested but widely held view. The consequence of this statement is that all human beings who fulfil the particular criteria of personhood, whatever these might be, are entitled to moral protection. For example, according to Kant, the key criterion would be the capability to act according to the moral law to which one subscribes to on the basis of one's own autonomy. The utilitarian philosopher Jeremy Bentham disagreed with this general idea and claimed that not the ability to reason, but the capability to feel pain was the key criterion that justified moral protection. In other words, all sentient beings (including animals) who fulfil this particular criterion do have moral rights, in particular the right to life. Given the two different accounts,

[25] In "Dignity and Animals: Does It Make Sense to Apply the Concept of Dignity to All Sentient Beings?" Federico Zuolo (2016, pp 1117–1130) argues that the main arguments—e.g., by Nussbaum (2006) and Meyer (2001)—for ascribing dignity to animals (i.e., the species-based approach, moral individualism, and the relational approach) are unconvincing and that one should instead use other normative concepts to justify the moral importance of animals.

it is quite possible that artificially intelligent beings might, in some distant future, fulfil either or both of the above-mentioned criteria, i.e., they will be capable of rational reasoning and of acting according to the moral law (based on their autonomy) and/or able to feel pain and thus to have feelings. In either case, given the presumption of equal treatment, one must also ascribe moral rights to IRs based on their particular design.

The history of ethics reveals that the presumption of equal treatment in the context of ascribing moral rights has often not been applied impartially. Instead, it has been restricted exclusively to the human species, regardless of whether other non-human beings may fulfil the very same criteria of personhood by which they should be entitled to moral protection (Singer 1979). The underlying reason for this situation is that the idea that moral protection should also include sentient, but non-rational animals remains highly contested and rejected by many ethicists and laypeople alike. That experts and laypeople, however, regularly commit fallacies based on speciesism in the context of ascribing moral rights has convincingly been pointed out by Singer (1975, 1979, 2009). In the past, the same line of reasoning has been applied to reject the moral rights of otherwise equally entitled human groups, based on incoherence and prejudice: women, African Americans in the US, people with impairments, and many others.

It might be objected that even though such injustices have occurred in the past (with respect to the presumption of equal treatment among fellow human beings and concerning sentient animals), human beings, and animals, unlike artificially intelligent beings, deserve moral protection, simply because they are natural living beings and are, therefore, also able to suffer in a deep and existential way that moral robots could not experience. The idea of naturalness is morally unimportant, however, since it is based on a moral prejudice that is, in general, implausible and unconvincing. The fact that a being is natural and not artificial says nothing about its moral status. The bare ontological substance of a being does not convey any morally relevant features (for a contrary view, see Torrance 2005; Johnson 2011, pp 169–172). In the movie *A.I. Artificial Intelligence* (2001), the child robot David was abandoned by his biological mother and experienced severe sorrow that can easily be seen as the equivalent of existential human suffering. Of course, David is not real but a movie character, and hence, one should not take this story line too seriously; no current robot is capable of suffering in the way that this movie portrays. However, we cannot rule out the possibility that artificially intelligent beings will have deep-seated feelings and emotions at some future point. In the final analysis, whether robots will become capable of moral reasoning and thereby be full ethical agents it is an empirical and not a normative question. The important question is, rather, whether humanity will, when that time arrives, be willing to accept robots as one of their kind, i.e., as moral beings.

3.3.4 Brutalization of the human character

Another important line of argumentation that Kant uses in the context of how we should treat animals is actually based on a virtuous line of reasoning. He claims that we should treat animals well, because mistreating them would negatively affect our behaviour towards our fellow human beings by causing us to become morally less sensitive over time. In other words, our moral character would be brutalized if we mistreated animals and we would start acting in a similar way towards other human beings. Strictly speaking, Kant does not say that mistreating animals is immoral, but just that it is something that one should not do. Analogously, human beings should treat artificially intelligent androids well, because mistreating them, especially in view of their great physical similarities to human beings, would make us morally less sensitive to fellow human beings (see also Anderson 2011a, b, p 293–295).

This line of argument has its strengths and weaknesses. Its great strength is that it does not presuppose that the object of morality—i.e., the sentient animal or IR—has an inviolable moral right to life; rather, it argues for treating the object well, because acting otherwise would harm the moral agent himself or herself. Its great weakness, of course, is that the object of morality itself is not granted any moral claim. Others are not morally bound to act in a certain particular way towards the objects, such as moral robots, because of any inherent characteristic that they possess.

One objection might question the degree of the emotional relationship between human beings and robots, which might differ substantially from our emotional connection to real animals, thus causing us not to give robots the kind treatment that we grant to animals. This objection is premature, however. It has been observed, for example, that cuddly toys in care homes for elderly people positively influence the atmosphere as well as the residents' emotional and medical condition. The elderly people do, indeed, treat the cuddly toys—in particular, Paro, a robot seal with fur—with care and feel sad when they are not available or broken (Turkle 2011, p 71). The actual relationship between these elderly residents and Paro is emotionally deep and complex. The same can be observed in situations where people, including children, use responsive programs such as Eliza[26] and

[26] Eliza is a chat program designed to mirror the thoughts of users, so as to give the impression that Eliza is consistently supportive. This mechanism has created a strong emotional effect (the so-called Eliza effect) on many people who have used the program.

robots such as Kismet[27] and Cog[28] (for a good overview of such examples, see Turkle 2011, pp 62–76). Given these examples, one can only imagine how relationships might develop when artificially intelligent beings are part of our world, life, and homes, as service robots or even as partners (Levy 2007). Therefore, it is quite conceivable that strong emotional attraction to a robot could establish moral bonds of a particular type that would eventually justify ascribing not only a moral status but even moral rights to IRs (Coeckelbergh 2014).[29]

4 Objections

The next section contains some additional objections to the account presented above. In response to these objections, further arguments are provided to substantiate the claim that IRs should be entitled to moral rights once they are capable of moral reasoning and decision-making.

4.1 The free will defence

The objection that IRs, no matter how technologically advanced they may be, are in principle unable to become moral agents, because they lack free will has been forcefully claimed by, for example, Johnson and Axinn (2014) and Rodogno (2016). In their view, moral agency necessarily presupposes free will. The underlying claim is that only human beings—and not robots—can be held morally responsible for their deeds, because they are the only beings who have free will. Therefore, it would be inappropriate to ascribe moral rights to IR.

It is of course impossible to solve the free will problem in this section, but I would like to highlight two unstated and highly controversial claims on which the above-mentioned authors (and many others in the debate) base their line of reasoning: first, that human beings have free will; second, that free will is a necessary precondition of moral agency. Most critics who work in the field of robotics and who use the free will defence against ascribing moral agency to IRs are unaware of the complex and interdisciplinary debate over free will (for an excellent overview, see the edited volumes of Watson 2003; Kane 2002; Pothast 1978). For example, the critics do not properly distinguish between the two central notions of freedom of will[30] and freedom of action[31] but, instead, use the definition of free action to define the notion of free will. This debate has become very sophisticated, with numerous thoughtful approaches associated with the main strands such as compatibilism (soft determinism),[32] incompatibilism (hard determinism),[33] and libertarianism,[34] as well as independent views that cannot be discussed here. However, the dispute clearly evidences that the underlying premise used by the above-mentioned critics—that human beings have free will—is controversial, given the existence of causality in the empirical world (i.e., physical determinism). Second, many authors in the free will debate (for example, Frankfurt 1969, 1971) believe that moral agency does not necessarily presuppose free will. In other words, the critics must prove both that free will exists and that it is necessary for moral agency before they can argue that IRs cannot be moral agents, because they lack free will. The critics of the free will defence in robotics must provide a more substantial line of reasoning along the lines of the libertarian strand to prove their point.

The more vital and related question, however, is whether the capability of moral reasoning and decision-making itself is enough to ascribe moral agency to artificially intelligent beings without adhering to the complex notion of free will. Against this background, it seems unfair to argue that robots must have free will to earn moral agency, even though the ideas that free will even exists or that human beings possess it remain controversial. Thus, until critics such as Johnson and Axinn (2014) and Rodogno (2016) provide a substantial argument in support of their underlying idea that free will is necessary for moral agency and that robots are, in principle, incapable of possessing it; their general claim is unsupported and misleading.

[27] Kismet, developed at MIT, is a complex robot that responds to facial expressions, vocalizations, and one's tone of voice.

[28] Cog, developed at MIT, can follow human motion, imitate behaviour, and track eye movements.

[29] Coeckelbergh (2014) questions the standard approach of ascribing moral rights to beings based on properties such as the ability to reason or to feel pain; instead, he suggests using a relational and phenomenological approach, contending that moral status emerges through relations between different beings (in particular, 69–70). I do agree, at least, to some extent with his view of *relations* between beings as highly important in evaluating moral status, but Coeckelbergh's questioning of the very idea of moral standing and his view of relations as *morally foundational* are somewhat unconvincing. Nonetheless, the relational approach has proven to be an important perspective in the context of disability studies, as well, particularly with regard to the moral status of people with severe mental impairments (Koch 2004). In both cases, the vital idea is to adhere to the *concrete relation* between two parties, whether it is the relation between the non-impaired human being and the person with mental impairment, or the human–robot relation.

[30] The freedom to will what one wants to will.

[31] The freedom to act according to one's own will.

[32] Free will is compatible with a world of physical determinism.

[33] A deterministic world and free will are incompatible.

[34] Free will (in a strong sense) presupposes an indeterministic world without (full) causation of mental events.

4.2 The following-a-program objection

Searle (1980, 1994) famously claims that machines cannot truly make decisions, because they are only following their program. Therefore, the very idea of IRs possessing moral agency that provides a basis for corresponding moral rights and duties is impossible. This frequently raised objection is related to the above-mentioned case of the free will defence, in that it suggests that IRs are strictly determined in their decisions based on their programming and, therefore, lack autonomous reasoning and decision-making.[35] Davenport, however, argues against this view, contending that "robots can learn new 'rules' as a result of interactions with the environment and/or internal reflections on past interactions. These new rules physically change the causal make-up of the mechanism, thus producing new behaviours, so that, in the future, in essentially identical circumstances, the robot may act completely differently." (2014: 53).

In earlier stages of computer development, chess programs such as IBM's DeepBlue depended on brute-force search when competing against human competitors, even when defeating world champions such as Garry Kasparov in 1996 and 1997. Nowadays, AI programs such as Google DeepMind's AlphaGo Zero operate quite differently from its predecessors. Starting from *tabula rasa* (blank slate learning), AlphaGo Zero was able to learn the complex traditional Chinese game of Go within just a few hours on its own, after being given only the basic rules of how to play Go and no further instructions. AlphaGo Zero was eventually able to beat the previous Go program, AlphaGo (numerous times by the score of 100 to 0), that defeated world champion Lee Sedol in 2017 (Silver et al. 2017).

It turns out that the particular combination of self-play reinforcement learning and deep neural network architecture is a key to the further development of AI programs. Interestingly, the same system is currently used in an unrelated domain to study the so-called protein folding in medicine, where the results could be used to find cures for "many devastating diseases, including Alzheimer's, Parkinson's, and cystic fibrosis" that originate from misfolded proteins (Knapton 2017). Accordingly, the self-taught AI program AlphaGo Zero is only the beginning, not the end, of progress in reinforcement learning.

This development presages the conceivable possibility that we will ultimately be able to build autonomous IRs that will act completely independently of human beings in the moral domain. As mentioned above, however, before we can ascribe moral rights to any being (including artificial beings), that being must meet the relevant criteria for moral agency. The minimum criteria for moral agency are commonly identified as autonomy and rationality. Therefore, if a being lacks either autonomy and/or rationality, then it will normally be considered a moral patient and not a moral agent (see Sect. 3.2). In this context, AlphaGo Zero could offer an interesting test case for considering IRs' capacity for moral reasoning and decision-making, and perhaps, such a test will be conducted in the future. Given its great potential, one would expect that it could quickly outperform many (or even all) existing moral programs, such as the casuistic approach of Rzepka and Araki (2005), the web-based approach of Guarini (2006), and the MoralDM of Dehghani et al. (2011). Amidst such developments, whether we will eventually reject the pessimistic view that a robot can never do something that it is not programmed to do is a purely empirical question.

In consequence, the threshold of granting intelligent robots moral rights—as I argue in this paper—is eventually a complex amalgam of (1) some necessary capabilities such as autonomy, rationality, and the ability of moral reasoning and decision-making and (2) the existence of morally important social ties between human beings and robots that prompt the robots to go beyond their initial programming and to develop features that make them individually unique. That means, "[s]ophisticated robots will undoubtedly develop unique identities, becoming, in a very real sense, individuals. As they live and work together with humans and other robots, they will naturally assimilate and develop moral rules that guide their social interactions. Eventually, we will come to accept them as fully moral agents, treating them as we treat other humans." (Davenport 2014, p 58).

4.3 Moral rights and the idea of dignity

A third objection contends that moral rights should apply only to natural beings such as human beings. Since robots are neither natural nor human beings, they should, therefore, not enjoy moral rights. In other words, robots are artificial beings without dignity worthy of being protected.

This objection contains within it at least three different lines of reasoning: (1) an ontological distinction between artificial and natural beings, (2) the idea that only human beings are moral agents and therefore have moral rights, and (3) the claim that artificial beings do not have any dignity. Concerning the first point, we have already observed in 3.3.3 above that the ontological difference between artificial existence and natural life, once both are capable of moral reasoning and decision-making, is unimportant for the ascription of moral rights. The decisive factor is whether the being is a moral agent, not what the being is made of.

Second, the idea that only human beings deserve moral protection is based on a biological bias (i.e., speciesism)

[35] For a more detailed discussion of this objection, see Whitby (2011, pp 140–142).

and does not do justice to, for example, some important approaches in the context of animal ethics, where authors such as Donaldson and Kymlicka (2013) and Francione (2009) attempt to show that animals also have strong moral rights. If proponents adhere to the criterion of capacity for reasoning when ascribing moral rights to human beings only, this account faces some vital problems. First, it opens the door to arguments, on the basis of utilitarianism or Kantianism, that people with mental impairments do not have the right to life; second, in animal ethics, proponents argue that one should protect higher animals such as the great apes who have the capacity to reason and, therefore, a moral right to life. In other words, the use of reasoning capacity as a basis for ascribing moral rights both transcends the borders of the human species, by including higher animals that are capable of reasoning, and also limits the rights accorded to some human beings, e.g., people with mental impairments, comatose people, people in permanent vegetative states, infants, and people with dementia.

Third, opponents could base the objection that artificial beings do not possess any dignity on Kant by arguing that something that is artificially created, including artificial beings such as robots, has no dignity but only a "price" and, therefore, is not part of the moral community or entitled to moral protection. This line of reasoning, however, is premature and misleading. Kant's famous distinction between irrational things (including animals) that have a "relative value" and rational persons who have an "absolute value" might instead be used against the opponents themselves (*Groundwork*, 60). One could argue as follows: In Kant's so-called kingdom of ends—which is the systematic connection of different rational beings through common ends—everything has either a "price" (for exchangeable equivalents) or "dignity" (for persons), (*Groundwork*, 68). It follows that if artificial beings, which are, nevertheless, human products, are rational (and hence autonomous), then they also have an absolute value and therefore are not simply things, but persons with dignity, worthy of moral protection. Indeed, Kant's theory is not concerned with human beings only but with all *rational* beings (i.e., logocentrism). But what about the issue of exchangeability with respect to robots? By definition, persons are not exchangeable but unique in Kant's reasoning and, hence, have dignity (absolute value) in contrast to exchangeable things that have a price (relative value). It appears that IRs are both rational (and hence persons with an absolute value) and exchangeable (and, therefore, having a price with a relative value). That seems to be a contradiction. This complex issue, however, can be resolved when one acknowledges that the personal identity (or individual self) of an IR is shaped and influenced by the unique and individual life history of the particular robot as an autonomous being. Hanna and Thompson (2003) distinguish between the "body" as a physical organism (*Körper*) and the "lived body" (*Leib*), which is the individual embodied experience of a particular organism in its life world. This distinction is important and could be further developed and applied to IRs as well once they have reached the particular threshold. Then, it would be no longer permissible to simply exchange the bodies of IRs, by virtue of the idea that the "lived body" depends on the particular physical body of the IR in question.[36] If one accepts this line of reasoning, it follows that IRs are no longer exchangeable and hence deserve full moral protection.

4.4 Rationality is not enough

Finally, it could be objected that, even if an artificial being is capable of reasoning, that would not suffice for the ascription of moral rights. Rather, the IR must also possess other important morally relevant criteria such as feelings and the capacity for suffering, self-preservation, and the capacity to reproduce. Only if the being has, in general, these capabilities can one ascribe full moral rights to it.

In response to this objection, first, the capability to have moral feelings and the capacity for physical suffering cannot be, in general, necessary conditions for claiming moral rights. If this were the case, then some people with mental impairments (who may lack the possibility to develop and reflect on their moral feelings for making decisions) or those with congenital analgesia (who cannot feel any physical pain) would be excluded as subjects of morality from the moral community and would become merely objects of morality. Contemporary research shows, however, that the capacity to have feelings, and moral feelings in particular, is important for moral reasoning and decision-making as such (Döring and Mayer 2002). Whether this amounts to a claim that, therefore, robots can never become full ethical agents remains to be seen. However, two brief responses are in order. First, we do not know whether it will be possible to reproduce moral feelings in robots; this is a technological issue. Second, it can be questioned whether the existence of moral feelings in human beings is necessary for moral theory, given that the most prominent and influential ethical theories, including both Kantianism and utilitarianism, completely avoid the use of moral feelings in arriving at

[36] The idea that IRs will develop individual selves and become unique members of the community is substantiated by Davenport: "Sophisticated robots will necessarily incorporate a model of themselves and their body in order to predict the effects of their interactions with the world. This mental model is the basis of their self-identity. As time goes by, it will incorporate more and more of the agent's interactions, resulting in a history of exchanges that give it (like humans) unique abilities and knowledge. This, then, is part of what makes an individual a unique and potentially valuable member of the group. Such machines will certainly have to be consciously aware (a-consciousness) of their environment" (2014: 56).

moral decisions. Perhaps, human beings, by virtue of their particular nature, need moral feelings for their own survival as a species, but that does not mean that other rational beings must possess the same moral feelings to make equally good moral decisions.[37]

It could be further objected that only beings who, phenomenologically speaking, have an idea of what it really means to suffer will be responsive to human suffering. We have already cited a contrary sentiment from fiction by referring to the example of the child robot David, who suffers due to having been abandoned by his biological mother (see 3.3.3). First, the idea that human suffering itself is necessary for moral reasoning and decision-making is questionable; second, it seems not impossible that intelligent robots will be capable of deep-seated feelings and emotions in the future[38]; third, although human suffering is certainly influential in human beings in the context of issues concerning human existence, life, and death, it does not follow that only beings capable of human suffering are legitimate subjects of morality and, hence, enjoy moral protection.

Second, it could be argued that the natural instinct of self-preservation is an important part of ascribing moral rights to beings, since, without this natural instinct, the particular being is not worthy of protection. This claim is based on the idea that if a being—under normal conditions—does not care whether it is alive or may be killed, then it does not make any sense to give that being moral protection. Against this line of reasoning, one could respond that it is, in general, possible to program an artificial instinct of self-preservation into robots; after all, nature has done the same using biological means with respect to human beings. The robot's instinct would thus be analogous to the natural human instinct (see Asimov's third law of robotics), and hence, the distinction between natural and artificial is irrelevant.

Third, critics might object that to deserve moral protection, beings must be able to reproduce themselves and to strive to become better as a species. Machines, according to this view, cannot reproduce or improve themselves without human intervention; therefore, one should not consider them as moral agents. However, this objection is misleading. For example, scientists at MIT succeeded as early as 2005 in building (primitive) self-reproducing robots, including a mechanism for corrections. They believe that self-reproducing robots are the very first step towards micro-electronic systems in nanotechnology that will be able to reproduce themselves without human intervention and even self-correct so as to become better and better with each following generation. This technological development could be used to build intelligent machines that can in turn construct other machines while avoiding the mistakes of previous machine generations. This may sound like a dangerous development to some, but it is a welcome opportunity to others. Furthermore, it seems questionable in general to view reproductive ability as morally relevant, since many human beings cannot reproduce yet remain moral subjects worth being protected. Therefore, it would be inconsistent to claim that in the case of robots, one should consider the capability to reproduce as morally relevant when we do not apply this criterion to human beings.

5 Conclusions

This paper has defended the idea of granting moral rights to artificially intelligent robots, once they are capable of moral reasoning and decision-making, against several different objections. There are no convincing reasons why IRs should be seen as morally inferior to human beings once they have reached this technological threshold. Therefore, it is of utmost importance that IRs learn how to make moral decisions and act accordingly, given their ever-increasing involvement in many sensitive fields and their increasing social interaction with human beings. Although we may not have to face this issue in the near-future (i.e., within the next couple of decades), it seems highly likely that humanity will confront this scenario at some future point. Whether IRs will eventually become better moral agents than human beings remains unknown, but we can be certain that the arrival of artificially intelligent beings will revolutionize our way of living. Whether this revolutionary change brings good or bad fortune depends mainly on our willingness to reach out to our fellow artificial beings.

[37] See also Allen et al. (2011, pp 59–60): "When it comes to making ethical decisions, the interplay between rationality and emotion is complex. Whereas the Stoic view of ethics sees emotions as irrelevant and dangerous to making ethically correct decisions, the more recent literature on emotional intelligence suggests that emotional input is essential to rational behaviour." Emotions certainly play an essential part with respect to the *genesis* of human morality, but emotions as such should never influence the *justification* of our moral reasoning and decision making. Therefore, I do not believe that emotions are necessary for IRs to arrive at correct moral decisions, but they will be essential for robots to engage with human beings on a social level. I agree with Whitby (2011, p 142), who claims that there "are also many contexts in which we prefer a moral judgment to be free from emotional content", such as those made by doctors and judges. However, "[e]motion may well be an important component of human judgments, but it is unjustifiably anthropocentric to assume that it must therefore be an important component of *all* judgments" (144).

[38] In his classical paper "The Feelings of Robots" (Ziff 1959, p 68), Paul Ziff claims that it is absurd to assume that robots will be capable of feeling anything. There is, however, no principled reason why this is logically impossible. For an illuminating discussion of the idea and meaning of suffering, see Gunkel (2014, pp 118–122) who argues that the concept of suffering is too complex and faces severe difficulties since it "remains fundamentally inaccessible and unknowable" (120).

Acknowledgements I would like to thank the anonymous reviewers for their valuable comments.

Funding This research is funded by the European Social Fund according to the activity 'Improvement of researchers' qualification by implementing world-class R&D projects of Measure No. 09.3.3-LMT-K-712.

References

Allen C, Wallach W, Smit I (2011) Why machine ethics? In: Anderson M, Anderson SL (eds) Machine ethics. Cambridge University Press, Cambridge, pp 51–61

Altman MC. 2011. Kant and applied ethics: the uses and limits of kant's practical philosophy. Wiley-Blackwell, New Jersey

Anderson SL (2011a) The unacceptability of Asimov's three laws of robotics as a basis for machine ethics. In: Anderson M, Anderson SL (eds) Machine ethics. Cambridge University Press, Cambridge, pp 285–296

Anderson SL (2011b) Machine metaethics. In: Anderson M, Anderson SL (eds) Machine ethics. Cambridge University Press, Cambridge, pp 21–27

Anderson M, Anderson SL (2011) Machine ethics. Cambridge University Press, Cambridge

Asimov I (1942) Runaround. A short story. Street and Smith Publications, New York

Asimov I(1986) Robots and empire. The classic robot novel. HarperCollins, New York

Atapattu S (2015) Human rights approaches to climate change: challenges and opportunities. Routledge, New York

Bringsjord S (2008) Ethical robots: the future can heed us. AI Soc 22(4):539–550

Bryson J (2010) Robots should be slaves. In: Wilks Yorick (ed) Close engagements with artificial companions: key social, psychological, ethical and design issues. John Benjamins, Amsterdam, pp 63–74

Calverley DJ (2011) Legal rights for machines. In: Anderson M, Anderson SL (eds) Machine ethics. Cambridge University Press, Cambridge, pp 213–227

Čapek K (1920) Rossum's universal robots. The University of Adelaide, Adelaide

Clark R (2011) Asimov's laws of robotics: implications for information technology. In: Anderson M, Anderson SL (eds) Machine ethics. Cambridge University Press, Cambridge, pp 254–284

Cochrane A (2010) Undignified bioethics. Bioethics 24(5):234–241

Coeckelbergh M (2014) The moral standing of machines: towards a relational and non-cartesian moral hermeneutics. Philos Technol 27(1):61–77

Darling K (2016) Extending legal protection to social robots: the effects of anthropomorphism, empathy, and violent behavior towards robotic objects. In: Calo R, Michael Froomkin A, Kerr I (eds) Robot law. Edward Elgar, Northampton, pp 213–231

Davenport D (2014) Moral mechanisms. Philos Technol 27(1):47–60

Dehghani M, Forbus K, Tomai E, Klenk M (2011) An integrated reasoning approach to moral decision making. In: Anderson M, Anderson SL (eds) Machine ethics. Cambridge University Press, Cambridge, pp 422–441

Delvaux M (2017) Report with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103 (INL))

Dennett D (1998) When hal kills, Who's to blame? Computer ethics. In: Stork D (ed) Hal's Legacy: 2001's computer as dream and reality. The MIT Press, Massachusetts, pp 351–365

Donaldson S, Kymlicka W (2013) Zoopolis. A political theory of animal rights. Oxford University Press, Oxford

Döring SA, Mayer V (eds) (2002) Die Moralität der Gefühle. In: *Deutsche Zeitschrift für Philosophie*. Sonderband 4. Akademie-Verlag, Berlin

Floridi L (2011) On the morality of artificial agents. In: Anderson M, Anderson SL (eds) Machine ethics. Cambridge University Press, Cambridge, pp 184–212

Floridi L, Sanders JW (2004) On the morality of artificial agents. Mind Mach 14(3):349–379

Francione GL (2009) Animals as persons: essays on the abolition of animal exploitation. Columbia University Press, New York

Frankfurt H (1969) Alternate possibilities and moral responsibility. J Philos 66(23):829–839

Frankfurt H (1971) Freedom of the will and the concept of the person. J Philos 68(1):5–20

Gibilisco S (2003) Concise encyclopedia of robotics. McGraw-Hill, New York

Gordon JS (2013) Modern morality and ancient ethics. Internet encyclopedia of philosophy. Published online 2013 http://www.iep.utm.edu/anci-mod/

Gordon JS (2014) Human dignity, human rights, and global bioethics. In: Teays W Renteln A (eds) Global bioethics and human rights: contemporary issues. Rowman & Littlefield, Lanham. pp 68–91.

Gordon JS (2016) Human rights. In Oxford bibliographies in philosophy, edited by Duncan Pritchard, published online 2016 (http://www.oxfordbibliographies.com/view/document/obo-9780195396577/obo-9780195396577-0239.xml?rskey=z2W9vS&result=47&q=)

Gordon JS (2017) Remarks on a disability-conscious bioethics. In: Gordon JS, Pöder JC, Burckhart H (eds) Human rights and disability. interdisciplinary perspectives. Routledge, London, pp 9–20

Grau C (2011) There is no 'I' in 'Robot': robots and utilitarianism. In: Anderson M, Anderson SL (eds) Machine Ethics. Cambridge University Press, Cambridge, pp 451–463

Guarini M (2006) Particularism and the classification and reclassification of moral cases. IEEE Intell Syst 21(4):22–28

Gunkel DJ (2012) The machine question: critical perspectives on AI, robots, and ethics. MIT Press, Cambridge

Gunkel DJ (2014) A vindication of the rights of machines. Philos Technol 27(1):113–132

Gunkel DJ, Bryson J (2014a) Introduction to the special issue on machine morality: the machine as moral agent and patient. Philos Technol 27(1):5–8

Gunkel DJ, Bryson J (2014b) The machine as moral agent and patient. Philos Technol 27(1):5–142

Hall JS (2011) Ethics for self-improving machines. In: Anderson M, Anderson SL (eds) Machine ethics. Cambridge University Press, Cambridge, pp 512–523

Hanna R, Thompson E (2003) The mind-body-body problem. Theoria Et Historia Scientiarum 7:24–44

Hernández-Orallo J (2017) The measure of all minds: evaluating natural and artificial intelligence. Cambridge University Press, Cambridge

Johnson DG (2011) Computer systems: moral entities but not moral agents. In: Anderson M, Anderson SL (eds) Machine ethics. Cambridge University Press, Cambridge, pp 168–183

Johnson AM, Axinn S (2014) Acting vs. Being Moral: The Limits of Technological Moral Actors. Proceedings of the IEEE 2014 International Symposium on Ethics in Engineering, Science, and Technology, 30:1–4. ETHICS'14. Piscataway, NJ, USA: IEEE Press

Kane R (ed) (2002) The Oxford handbook of free will. Oxford University Press, Oxford

Kant I (2009) Groundwork of the metaphysic of morals. Harper Perennial Modern Classics, New York

Knapton S 2017. AlphaGo Zero: Google DeepMind supercomputer learns 3,000 years of human knowledge in 40 days. https://www.

telegraph.co.uk/science/2017/10/18/alphago-zero-google-deepmind-supercomputer-learns-3000-years/. Accessed 27 March 2018
Koch T (2004) The difference that difference makes: bioethics and the challenge of 'disability'. J Med Philos 29(6):697–716
Levy D. 2007. Love and sex with robots: the evolution of human-robot relationships. Harper, New York
Lin P, Abney K, Bekey GA (eds) (2014) Robot ethics: the ethical and social implications of robotics. Intelligent robotics and autonomous agents. The MIT Press, Cambridge
Macklin R (2003) Dignity is a useless concept. BMJ Br Med J 327(7429):1419–1420
Meyer M (2001) The simple dignity of sentient life: speciesism and human dignity. J Soc Philos 32(2):115–126
Moor JH (2006) The nature, importance, and difficulty of machine ethics. Res Gate 21(4):18–21
Nadeau JE (2006) Only androids can be ethical. In: Ford K, Glymour C (eds) Thinking about android epistemology, pp 241–248. MIT Press, Cambridge
Nussbaum M (2006) Frontiers of justice. disability, nationality, species membership. The Belknap Press of the Harvard University Press, Cambridge
Picard R 1997. Affective computing. The MIT Press, Cambridge
Pothast U (ed) (1978) Seminar, Freies Handeln Und Determinismus, 1. Aufl. Suhrkamp, Suhrkamp Taschenbuch Wissenschaft 257. Frankfurt am Main
Rodogno R (2016) Robots and the Limits of Morality. In: Norskov M (ed) Social robots. boundaries, potential, challenges, Routledge. (http://pure.au.dk/portal/files/90856828/Robots_and_the_Limits_of_Morality.pdf). Accessed 03 Dec 2016
Rzepka R, Araki K (2005) What statistics could do for ethics? The idea of common sense processing based safety valve. AAAI Fall Symposium on Machine Ethics, Technical Report FS-05-06: 85–87
Searle J (1980) Minds, brains and computers. Behav Brain Sci 3(3):417–457
Searle J 1994. The rediscovery of mind. The MIT Press, Cambridge
Silver D et al. (2017). Mastering the game of Go without human knowledge. Nature 550: 354–359
Singer P (1975) Animal liberation. Avon Books, London
Singer P (1979) Practical ethics. Cambridge University Press, Cambridge
Singer P (2009) Speciesism and moral status. Metaphilosophy 40(3–4):567–581
Singer P (2011) The Expanding Circle: Ethics, Evolution, and Moral Progress. 1st Princeton University Press pbk. Princeton University Press, Princeton
Sullins JP (2011) When is a robot a moral agent?. In: Anderson M, Anderson SL (eds) Machine ethics. Cambridge University Press, Cambridge, pp 151–161
Torrance S (2005) A robust view of machine ethics. In: Technical Report—Machine Ethics: Papers from the AAAI Fall Symposium, FS-D5-06, American Association of Artificial Intelligence, Menlo Park, pp 88–93
Turkle S (2011) Authenticity in the age of digital companions. In: Anderson M, Anderson SL (eds) Machine ethics. Cambridge University Press, Cambridge, pp 62–76
United Nations (2016) Preliminary draft report of COMEST on robotics ethics. *SHS/YES/COMEST-9EXT*/16 (3): 1–31. (http://unesdoc.unesco.org/images/0024/002455/245532E.pdf.Accessed 03 Dec 2016
Wallach W, Allen C 2010. Moral machines: teaching robots right from wrong. Oxford University Press, Oxford
Warren M (1973) On the moral and legal status of abortion. Monist 57(1):43–61
Watson G (ed) (2003) Free will. In: Oxford readings in philosophy. Oxford University Press, Oxford
Whitby B (2011) On computable morality: an examination of machines as moral advisors. In: Anderson M, Anderson SL (eds) Machine ethics. Cambridge University Press, Cambridge, pp 138–150
Ziff P (1959) The feelings of robots. Analysis 19(3):64–68
Zuolo F (2016) Dignity and animals. Does it make sense to apply the concept of dignity to all sentient beings? Ethical Theor Moral Pract 19(5):1117–1130