AI & SOCIETY (2022) 37:319–330
https://doi.org/10.1007/s00146-021-01179-z

OPEN FORUM

In search of the moral status of AI: why sentience is a strong argument

Martin Gibert[1] · **Dominic Martin**[2]

Accepted: 11 March 2021 / Published online: 8 April 2021
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2021

Abstract
Is it OK to lie to Siri? Is it bad to mistreat a robot for our own pleasure? Under what condition should we grant a moral status to an artificial intelligence (AI) system? This paper looks at different arguments for granting moral status to an AI system: the idea of indirect duties, the relational argument, the argument from intelligence, the arguments from life and information, and the argument from sentience. In each but the last case, we find unresolved issues with the particular argument, which leads us to move to a different one. We leave the idea of indirect duties aside since these duties do not imply considering an AI system for its own sake. The paper rejects the relational argument and the argument from intelligence. The argument from life may lead us to grant a moral status to an AI system, but only in a weak sense. Sentience, by contrast, is a strong argument for the moral status of an AI system—based, among other things, on the Aristotelian principle of equality: that same cases should be treated in the same way. The paper points out, however, that no AI system is sentient given the current level of technological development.

Keywords Ethics · Artificial intelligence (AI) · Moral status · Sentience · Pathocentrism · Biocentrism

Is it OK to lie to Siri?[1] Is it bad to mistreat a robot for our own pleasure? Should we treat all artificial intelligence (AI) systems with moral consideration? While it has often been assumed that AI systems exist for the sole purpose of serving the interests of humans—and that they can be instrumentalized to this end—the question of the moral status of AI systems is a matter of increased interest. Robert Sparrow, in 2004, suggested that the capacity for reasoning, self-consciousness, or undertaking projects would be a relevant criterion for making an AI system worthy of moral respect, but that the capacity to feel pain may be the most morally relevant capacity. He did not, however, develop an argument to defend any one criterion over another, and many different contributions have been made on this question since the early 2000s.

Taking stock of this literature, this paper looks at different arguments for granting moral status to an AI system.[2] In each but the last case, we find unresolved issues with the particular argument, which leads us to move to a different argument that would avoid these issues. First, we claim that we may have indirect duties concerning an AI system, but this is insufficient to ground a moral status because this does not lead us to consider an AI system for its own sake. Second, according to the relational argument, an AI system can be considered for its own sake, but the argument raises issues in terms of its lack of consistency or objectivity in the attribution of moral status. The argument from intelligence is more consistent or objective, but we argue that

✉ Dominic Martin
martin.dominic@uqam.ca

Martin Gibert
martin.gibert@umontreal.ca

[1] Affiliated to the Centre de Recherche en Éthique, University of Montreal, 2910, Boul. Édouard-Montpetit, bureau 313, Montréal, QC H3T 1J7, Canada

[2] School of Management, Université du Québec À Montréal (UQAM), Downtown Station, PO 8888, Montréal, QC H3C 3P8, Canada

[1] The software agent, developed by Apple Inc., that can answer questions, make recommendations, and perform actions using voice queries and a natural language interface (Hoy 2018).

[2] The notion of having a moral status is very close to the notions of having moral standing, having moral considerability, and/or having moral patiency. A deeper analysis could reveal some differences among these notions, but we will use them interchangeably in this paper and we will prioritize the expression *moral status* when possible. Let us also note that this question of the moral status of AI belong to the fields of AI ethics and robot ethics, which are concerned not only with moral agency but also with moral patiency (Loh 2018).

intelligence is not a relevant criterion for moral status among humans.

We then consider the idea that an AI system could have a moral status if it is a living or an information system. The property of being alive—broadly construed—may ground an AI system's moral status, but only in a weak sense. Thus, the argument from sentience is the strongest in our view because it avoids the issues identified with the previous arguments. We point out, however, that no AI systems exhibit the sentience that would be necessary for having a moral status, given the current state of technological development. The paper is divided in six sections. We introduce preliminary remarks in the first section and then we deal with each of the five arguments in the subsequent sections.

1 Preliminary remarks

We understand an *AI system* as a system that uses new technologies in AI to reproduce the properties associated with biological intelligence: skills, behaviors, affects, and so on. This category includes virtual assistants such as Siri, humanoid robots and advanced psychology software. AI systems today rely on one or more algorithms (Hill 2016) to reproduce intelligence—for example, a neural network trained to process natural language or to classify visual inputs—as well as on other structural elements.

This conception of AI is similar to other popular conceptions of AI systems as "artifacts that extend any of the capacities related to natural intelligence" (Bryson 2019) or "any kind of artificial computational system that shows intelligent behaviour" (Müller 2020). We adopt a broad definition of artificial intelligence because we want to consider various arguments for the moral status of AI. There is a wide diversity of conceptions of AI in the literature (Bringsjord and Govindarajulu 2018) and we do not want to exclude any argument on the basis that it rests on a different conception.

However, we use the term "robot" in a more exclusive sense to point to an AI system designed to produce "something physical or tangible in the world": to weld metal pieces together, to wipe the floor, to emulate the human body, and so on (Martin 2017).[3] Robots often rely on AI systems, but not all AI systems are robots. This narrower conception of robots will be useful within the limited scope of this paper to draw a distinction between AI systems in general and AI systems—such as humanoid robots—that have a more clearly circumscribed physical dimension. AI systems that are mostly virtual may not elicit the same emotional or cognitive responses from humans and may change our perception about their moral status.

Sparrow (2004) nicely illustrates the significance of the question of moral status for an AI system with a scenario he calls the "Turing triage test." Let us imagine a situation where we have to choose (following a catastrophic loss of power in a hospital) between unplugging the machine supports of a human who is being kept alive by artificial means and unplugging an AI. We will know that machines have a moral status comparable to that of humans "when the replacement of one of these people with an artificial intelligence leaves the character of the dilemma intact. That is, when we might sometimes judge that it is reasonable to preserve the continuing existence of a machine over the life of a human being" (203).

Even if these kinds of life-or-death situations do not represent the whole range of what it means to have moral status, this is certainly a useful example through which to grasp what is morally relevant.[4] To have a moral status is, to use Warren's (1997, 3) formulation, "to be morally considerable, or to have moral standing. It is to be an entity towards which moral agents have, or can have, moral obligations. If an entity has moral status, then we may not treat it in just any way we please." Kamm (2007: 229) proposes a similar definition: "an entity has a moral status when, in its own rights and for its own sake, it can give us reason to do things such as not destroy it or help it." For instance, a work of art can count morally—we want to prevent the artwork from burning—but lack moral status, because we do not act for the sake of the artwork when we prevent it from burning. We do not think it would be good *for* the artwork to continue existence. In the same vein, we may have moral reasons to not burn a flag, but this do not imply the flag has a moral status. In the Turing triage test, we can say the AI system has a moral status if we decide to save it *for its own sake*—and not only for reasons similar to the ones about the work of art or the flag.[5]

Importantly, having a moral status is not the same thing as being a moral agent. An entity is a moral agent when it is morally responsible for what it does (Gruen 2017). Babies, people with cognitive disabilities and animals are usually not

[3] While a system such as a computer may also have a physical dimension (a screen, a case, electric circuits, etc.), its useful product may be only virtual: a hash value, the solution to a facial recognition request, a risk prediction, and so on.

[4] Sparrow (2012) himself considers this test to be an explanatory thought experiment, rather than a discriminatory test like the Turing test.

[5] That is why we should be careful when using moral dilemmas, such as the Turing test triage, to look at the moral status of an entity. It is totally conceivable that you have stronger reasons to save something without moral status (like a work of art) than an entity with a moral status (like a plant or an animal). Likewise, you may have stronger reasons to save a child than an elderly person, but that does not mean that the elderly have no moral status (of course they do).

regarded as morally competent and blamable. They are not moral agents, but they have a moral status. More precisely, they are moral patients because they can be morally wronged. Thus, in the case of human beings, moral patiency comes temporally before moral agency (babies are moral patients and they will become moral agents), and every human moral agent is also a moral patient.[6] However, different combinations are possible: an entity could be a moral agent without being a moral patient, could be both a patient and an agent, and so on.

It should also be noted that this understanding of moral status is open to threshold or scalar conceptions (Jaworska and Tannenbaum 2018). In the threshold conception, an entity has moral status or not depending on whether or not it meets certain criteria—for instance, possessing a capacity such as self-consciousness. In the scalar conception, there are degrees of moral status: for instance, the better an entity displays some capacity, the higher its moral status. Having *full moral status* means the entity reaches some threshold or possesses the highest moral status in the continuum.

2 Indirect duties

On August 2015, hitchBOT, a Canadian hitchhiking robot, was found decapitated in Philadelphia, after having successfully crossed Canada and a few European countries (Victor 2017). The robot was programmed to ask people to be picked up and to travel with them. hitchBOT consisted of a cylindrical body, made using a plastic bucket, with attached arms and legs, and an LED screen displaying eyes and a mouth. Vandals destroyed the robot and left him on the side of the street.

Did the vandals do something bad when they decapitated hitchBOT? Did they have a moral obligation not to interfere with its normal state of functioning and let him continue his journey across the world? We might say that the vandals violated a moral obligation if, first, they destroyed the property of David Harris Smith and Frauke Zeller, the creators of hitchBOT, without their consent. Their actions might have been bad if they encouraged other people to adopt violent behavior, if they promoted vandalism, or if they simply fostered bad dispositions or feelings among the people who followed the story. We may even wonder whether the vandals did not have self-destructive behaviors, and whether engaging in these destructive actions with hitchBOT contributed negatively to their own psychological well-being or other personal issues.

These considerations suggest that we may have indirect duties regarding AI systems—that is, duties that are not owed to an entity directly, but owed insofar as our treatment of such entity can affect our duties to other human beings or ourselves. Similar issues have been raised in animal ethics in regard to the obligations we may have towards other living creatures (Gruen 2017). Kant, who did not think that animals have a moral status, claimed famously that humans had some obligation to not harm them:

> If a man shoots his dog because the animal is no longer capable of service, he does not fail in his duty to the dog, for the dog cannot judge, but his act is inhuman and damages in himself that humanity which it is his duty to show towards mankind. If he is not to stifle his human feelings, he must practice kindness towards animals, for he who is cruel to animals becomes hard also in his dealings with men. (Kant 1997, Ak 27:459)

The notion of indirect duties comes out in different places in the animal ethics literature to clarify the moral obligations we could have towards nonhuman entities (Wilson 2002).[7] It also emerges in other literatures, such as the literature on international justice and human rights, on the basis of which it may be claimed that because we have a *direct* duty towards the bearer of these rights, we have an *indirect* duty to maintain and enhance international institutions that are working to protect human rights (Shue 1988).

If we apply the same argument to robots, this suggests that people who are *cruel* to them, for instance, may become cruel to other moral patients. Darling (2016) explains that certain behaviors towards a robot or a chatbot may traumatize or desensitize us. In the same vein, Danaher (2017) suggests that future sex robots may encourage their users to ignore norms of consent.[8] Typically, it is argued that the

[6] Building on this distinction, Hogan (2017) contested the claim that the machine question is the same as the animal question. She claims that the latter is patient-centered while the former originates from agency. The fact that robots may be moral agents before being patients would prevent us from addressing the question of the moral status of AI using the same arguments we would use for animals. However, we do not see why different types of entities should be considered differently because they acquired moral agency or patiency in a different order. An AI system can have moral status without being a moral agent if we can have reasons to act for its own rights and for its own sake.

[7] It should be pointed out that some animal ethicists, such as Korsgaard (2018, 102–5), criticize indirect duties on the basis "that it is almost incoherent." Or, at least, the idea of indirect duties invites us to have an attitude that is incoherent because we are compelled to treat animals with gratitude, or love, yet also to detach this attitude from any moral concern about animals.

[8] Note that there are at least two different theses in the idea of an indirect duty to an animal or an AI system. First, this implies that we owe the duty to treat an animal well to ourselves or other humans, rather than directly to the animal. Second, this implies that the basis of the duty resides in the effects on human behavior, dispositions, or moral character, and not on the effects on the animal. As pointed out by Korsgaard (1996, 101 ff.), these two theses are separate, logically speaking: we might owe it to ourselves, not directly to animals, to

following effects on humans would provide sufficient reason to ground an indirect duty to treat an AI system well:

- Behaving badly with an AI system can lead to bad behaviors perpetrated against other entities, such as human beings (who have a moral status). As pointed out above, destroying hitchBOT implied destroying the property of its owners. I might have an obligation to not desecrate an object you regard as sacred because I have a moral obligation to you (Shepherd 2018, 22).
- Behaving badly with an AI system can foster bad dispositions or habits, such as a tendency towards deception, violence, or questionable sexual practices, among humans. For instance, parents now have the option to program the chatbot Alexa in a way that strengthens the politeness habits of children (Gonzalez 2018). In Korea, Shelly, a robot tortoise, was developed to curb children's abuse of robots (Ackerman 2018).
- One may even claim that one has duties towards oneself, or ought to treat oneself in certain ways (Wood 2009). For instance, it is quite common to think that self-mutilation is bad, that self-preservation is good, that one must be healthy or well informed to live a fulfilling life. In that case, behavior such as lying to an AI system or destroying a robot for no reason—even if the AI system is one's property and even if this does not harm anyone else in any way—could be connected with problems of depression, denial or other self-destructive behaviors. In that sense, this behavior may violate duties we have towards ourselves.
- Mistreating or destroying AI systems can be a form of waste or a misallocation of resources.

This list is by no mean exhaustive, but it shows how, under certain circumstances, we can have moral obligations regarding an AI system in the form of indirect duties. Furthermore, indirect duties may be stronger in the case of robots as opposed to other AI systems if the robot has a more humanlike shape. For instance, mistreating a sexual robot may be more problematic precisely because its humanoid form may increase the risk that this will lead to the mistreatment of human sexual partners. However, the argument of indirect duty can also apply to other types of AI systems that are not embodied. As exemplified above, destroying a computer system or lying to a computer program can make us more prone to adopting self-destructive behavior or violating duties toward ourselves in other ways, in addition to destroying the property of another human.

Even if we do have indirect duties towards AI systems, this type of duty is not relevant to our main objective in this paper, which is to see whether an AI system is a valid candidate for moral status. Indirect duties imply, by definition, to not consider an entity directly for its own rights or for its own sake. Therefore, they cannot lead to the recognition of a moral status for an AI system. It was important to start with this distinction between indirect and direct duties to keep in mind that the moral status of an entity is not the only relevant consideration in determining how to behave morally with it. We will now look at arguments involving direct duties, which can ground a moral status for AI systems.

3 The relational argument

It may be argued that if a person has a valuable relationship with an entity e, then e is worthy of moral consideration, or that moral status is conferred only in the community of entities $e_1, e_2, e_3, \ldots$ that are in relationship with one another. If people develop special bonds or connections with other entities—a house cat, or a protected forest, for instance—then these entities should not be instrumentalized to serve other people's ends. By extension, some AI systems—a virtual assistant, a robot aid, a psychology program—could also be part of this network of relations, and they would have a moral status for that reason. Let us call this *the relational argument* for the moral status of AI systems.

This argument comes out frequently in the feminist, ecologist, and robot ethics literatures. According to Sherwin (2009), a relational approach is a more promising way to deal with ethical questions regarding the acceptability of abortion and euthanasia. Instead of trying to establish which criterion will tell us whether we can terminate the life of a fetus or a person in a persistent vegetative state, we should focus on the relations within which these entities are embedded. We should keep in mind that "societies create persons from the biological base of human existence though collective social practices" (152). Ascribing moral status to these entities involves more than determining physical or psychological characteristics; it "also requires determining the moral understanding of the community in which these beings reside" (153). For instance, many pregnant women experience pregnancy as a relational state with the fetus, and fetal movements will reinforce this experience. Family members often feel attached to a patient and will be quite willing to give care, even if the person is in a persistent vegetative

Footnote 8 (continued)

treat them well, and nonetheless the duty could be to treat animals well for their own sake, rather than for the effect it would have on humans.

state. We must take these relationships into account when we establish the moral status of these entities.[9]

These ideas have also resonated in the robot ethics literature. Coeckelbergh (2010, 217) builds on different views in ecological ontologies and social ecology to show that entities are "inter-related" and "inter-dependent," which suggests that the moral consideration of AI systems "is bound up with social relations between human and robots." Similarly to personal identity, which is often defined in terms of the relationship a person maintains with other members of a given community, moral status is defined within these communities. Gunkel (2012, 2014) advocates a viewpoint similar to Coeckelbergh's relational approach, claiming that "one is assigned the position of moral person (whether that be an agent, patient, or both) as a product of social relations that precede and prescribe who or what one is" (2012, 172). In his view, moral personhood, including moral agency and patiency, is conferred by one's social community.[10]

There are at least two interpretations of the relational argument. The first is that a valuable relationship with an entity implies the moral status of that entity. If we love, say, our cat Sophia, it would seem contradictory or incoherent for us to claim that her well-being does not matter, irrespective of our love for her. If that is the case, the argument states, then our cat has a moral claim on us, she can be wronged and she has a moral status.[11]

The second interpretation of the relational argument emphasizes the role that a community plays in the development and assignment of social identity, personhood, and moral status. In other words, moral status exists or, at least, is defined only within communities or networks of entities that enter into relationship with one another. This interpretation comes out particularly strongly in Coeckelbergh's and Gunkel's views. We agree with them when they say that communities have an impact on the definition and evolution of identities, rights, moral status, and so on. However, this reads more like a metaphysical view about the nature of moral status; it does not tell us which entities should be included in the circle of entities with moral status, or that we need to include any entity that is not included.

The first interpretation of the relational argument raises issues of consistency and objectivity. First, the meaning of a valuable relationship is subjective: different persons can experience valuable relationships with different entities for various reasons, some of them being very contingent. For instance, I may develop a special relationship with a dog or a goldfish because it reminds me of another dog or goldfish, or place or an event. If the existence of a valuable relationship serves as an enabling condition for moral status, then moral status becomes very subjective, contingent, or even arbitrary, and cannot be easily universalized.[12]

Closely connected to the first issue, the second issue is that the relational argument may introduce variability or fluctuations in the moral status of nonhuman entities. If we claim that an entity ought to be considered morally because a valuable relation has been developed, what happens if the relation ceases to exist? Does this mean that an entity is not worthy of moral consideration anymore? This seems counterintuitive. Third, how ought we to deal with token/types distinctions? Say we develop a valuable relationship with our cat Sophia. Does this mean that all cats are now worthy of moral consideration—in which case, moral status is granted to cats as a type of entity—or only Sophia, a specific cat token? If it seems preferable to include entity types instead of tokens, what are the implications for the expansionist tendency of the view? Perhaps these problems are not fatal objections to the relational argument, but they do invite us to be cautious before using that argument for defending the moral status of AI systems.[13]

4 The argument from intelligence

At first glance, AI systems are different from other human artifacts or tools because they are *intelligent*. Intelligence is one of the most strikingly common features between AI systems and typical entities with a moral status such as human beings. Furthermore, it is often argued that having sophisticated cognitive capacities, or the capacity to develop these capacities, is a necessary and sufficient condition for having a moral status (Jaworska and Tannenbaum 2018). The most famous articulation of this view was given by Kant (1785), according to whom autonomy and rationality ground the dignity of all human beings.[14] Thus, it seems reasonable to

[9] Little (1999) makes a similar argument to the one made by Sherwin, though her focus is on abortion specifically. See also the work of Hester et al. (2000) for the application of the relational approach to questions of environmental ethics (for example, the moral considerability of land-related entities).

[10] His most recent work takes a somewhat different direction, however, framing the issues in terms of the debate between normative and descriptive claims in moral philosophy, and suggesting that we should "deconstruct" the "is-ought inference" using the work of Emmanuel Levinas (Gunkel 2018, 159).

[11] See also Korsgaard (2018, 102–5) and Thomas Scanlon (1998, 164–65). Although their work is not rooted in a relational perspective, they provide further elaborations on this idea.

[12] On the importance of universalizability in moral philosophy and the challenges associated with nonconsequentialist approaches, see Pettit (2000).

[13] For other critical perspectives on the relational approach, see Anne Gerdes (2016).

[14] Contemporary accounts of similar views can be found in the work of Quinn (1984) and Stone (1987).

ask whether intelligence or cognitive capacities could not ground the moral status of an AI system.

The general argument here is that once an entity is above a specific threshold of intelligence, it should be recognized as having a moral status. One may believe, for instance, that humans are permitted to exploit some animal species because they do not reach a minimal threshold for intelligence. There is also a scalable version of the argument: we have *more* moral obligations towards an entity if it is *more* intelligent. Following that logic, it would be acceptable to disregard the interests of rocks or bacteria, but this would be more problematic with chimpanzees, ravens or octopuses.

But what is intelligence? After a comprehensive survey of the literature, Legg and Hutter (2007, 12) define what seems to be the core of intelligence as "an agent's ability to achieve goals in a wide range of environments." They explain that intelligence is about the ability to deal with some range of unanticipated possibilities. Giving a nonanthropocentric definition of intelligence is, of course, necessary if we do not want to arbitrarily exclude machine intelligence.[15] Considering this nonanthropocentric definition, there is no doubt that some AI systems are intelligent, as well as chimpanzees, ravens and octopuses.

The argument of intelligence is not prone to the same weakness as the relational argument because it is possible to define intelligence in way that is not contingent on the subjective experience of a human being or a community of human beings. However, there is a good reason to reject intelligence as a criterion for moral status. That is, the property of being intelligent seems totally disconnected from the possibility to be treated morally. Indeed, reaching a threshold of intelligence is generally not required for a human being to have moral status. We usually consider that babies or mentally disabled people can be wronged, even if they do not have the cognitive capacities of a typical human adult.[16]

Singer (2011) gives an example of a slavery society based on intelligence quotient (IQ), where the low IQ individuals would be slaves of the higher. In his view, there is almost no doubt that this society would be considered unjust. This is because intelligence seems to be as arbitrary as race or gender, as a foundation on which to ground unequal treatment.

> Intelligence has nothing to do with many important interests that humans have, like the interest in avoiding pain, in satisfying basic needs for food and shelter, to love and care for any children one may have, to enjoy friendly and loving relations with others and to be free to pursue one's projects without unnecessary interference from others. (Singer 2011: 21)

Of course, intelligence may be instrumentally valuable, in particular if this allows one to achieve valuable "complex goals," such as doing some good. Cognitive capacities may also be morally relevant in some circumstances. However, this does not mean that intelligence grounds moral status. There is no reason to posit that a more intelligent entity should have a higher moral status.[17]

Finally, it should be pointed out that the argument from intelligence may have some undesirable implications. The argument may imply, under its scalable version, that a hypothetical superintelligent AI system would have a higher moral status than every human and nonhuman animal.

If intelligence is not an appealing criterion for grounding the moral status of an AI system, the structure of the argument is nevertheless similar to those of the last two arguments we will consider. The basic idea is to state something like this: if we give a moral status to people because they possess the property *X*, then an AI system with the property *X* should be granted an equivalent moral status. There is a sense in which this argumentative structure derives from the Aristotelian principle of equality (Aristotle 2000, V.3. 1131a10–b15). According to that principle, each similar case should be treated in a similar way. Thus, if two entities are similar in the sense that they share the same property, they should be granted the same moral status. In the last two sections of the paper, we will consider the property of being alive or carrying information, and the property of being sentient, that also lead to an argument with the same structure.

5 The arguments from life and information

A few currents of thought suggest that life is a relevant criterion for moral status because being alive seems to ground the possibility of all moral considerations. How an entity that is not alive, such as a rock or a hammer, could, in its own right and for its own sake, give us reasons to do things such as to not destroy them? On the other hand, it is easy to see that a

[15] Legg and Hutter (2007) even propose a formalization of this "universal intelligence," which may be applied to measure a machine's intelligence—implying, among other things, a reward function and the principle of Occam's razor.

[16] For an overview of the arguments on the importance of treating incapacitated humans with decency, see the work of Hill (1993), Frankena (1986) and Cranor (1975), among others.

[17] More intelligent humans *often* have a higher *social status*, though, which also tends to be heavily criticized. The egalitarian trend of thought, exemplified by political philosophers such as John Rawls (1971), argues that we should reduce, as much as possible, the effect of natural endowments (including a genetic predisposition to intelligence) on people's liberties and opportunities in life. Under this formulation, intelligence is also considered an irrelevant condition for higher social status in addition to being an irrelevant condition for moral status.

living entity may have a basic interest not to be killed and to continue its life.

This view is especially important among environmental ethicists, and *biocentrism* is precisely the normative theory according to which all and only living creatures have a moral status. For instance, Taylor (1981) argues that every organism that achieves its own good is a "teleological center of life" and, consequently, has intrinsic value. This also captures widely held views in non-Western cultures, such as Indigenous cultures, which can ascribe agency and moral status to a river, a forest, a tree, or other entities (Harvey 2005 and Stone 1985). For biocentrists, trees or plants have to be treated morally, and this is not because they are useful to us (because of their instrumental value), but for their own sake. That is why we can say that they have a moral status.

Biocentrist arguments may be extended to AI systems if there are reasons to think that an AI system could be "alive" in a way that biocentrists would accept. As it was the case with the criterion of intelligence, answering this question depends largely on our definition of life. A narrow definition of life can exclude nonorganic entities by stating, for instance, that having a metabolism is an absolute criterion for life (Bedau and Cleland 2010). Indeed, only organic entities seem able to instantiate metabolic processes that convert food to cellular energy or building blocks for the organism, and that eliminate nitrogenous waste. While it is not impossible that one day an AI system will rely on these kinds of metabolic processes to produce the energy it needs, this is not the case today.

However, it is possible to leave the sphere of biology for an information-based perspective and to propose a wider definition of life. Tegmark (2017, 39) defines life as "the ability [for an entity] to retain its complexity and replicate." Living entities at different scales tend to resist entropy and to behave in a way that reduces the gulf between their sensory input and a future state of the world—such as gene replication—which is similar to a goal (Heams 2019). According to this broad definition, some AI systems can be said to be (artificially) alive: an algorithm is a set of information and it can be designed to replicate itself. In the early 1990s the computer simulation Tierra, developed by the ecologist Thomas Ray, was considered to be artificial life. The code of the software was notably evolvable and able to self-replicate.[18]

Along those lines, an even more inclusive version of the argument from life rests on the premise that "information is the ultimate constituent of reality," or existence (Hogan 2017, 33). For instance, Floridi and Sanders (2002; Floridi 2010) argue that we should substitute life with existence in our moral understanding of the world, and that we should move from biocentrism to "ontocentrism." That is, we should grant a moral status to all beings because they carry information in one way or another. According to Floridi, any information entity "has a right to persist in its own status, and a right to flourish, i.e., to improve and enrich its existence and essence" (Floridi 2010, 112). Under this view, entropy is similar to pain or suffering, and it should be avoided as much as possible. The somehow counterintuitive implication is that any form of destruction, corruption, pollution, or depletion of an information object is wrong to some extent.

To sum up, there are at least three ways to expand the set of entities with a moral status based on the argument from life and information. According to a first, 'narrow' version, we can have moral obligations towards all entities that are alive, according to a narrow definition of life, which tends to exclude AI systems. According to a second, 'broader' version, we can have moral obligations towards all entities under a wide definition of life that can include some AI systems. The third version is Floridi's ontocentrism. It suggests we should give all information entities a moral status, which obviously includes all AI systems.

First, we will discuss the third version because ontocentrism raises many questions. If moral status is ascribed to all beings in the universe, regardless of their goal-oriented behavior, capacity to retain an informational state, or level of complexity, then there is a risk that moral status becomes a trivial concept that is not useful for solving any moral question.[19] Consider Floridi's (2010, 113) claims that ontocentrism concerns "any entity, understood informationally," which includes human beings, animals, and plants, but also "anything that exists, from paintings and books to stars and stones" and anything that will exist or has existed. He also adds:

> information ethics holds that every entity, as an expression of *being*, has a dignity, constituted by its mode of existence and essence (the collection of all the elementary proprieties that constitute it for what it is), which deserve to be respected (at least in a minimal and overridable sense), and hence place moral claims on the interacting agent and ought to contribute to the constraint and guidance of his ethical decisions

[18] However, it has to be admitted that artificial life is generally perceived as a simulation of life rather than as an authentic form of life (Boden 1996).

[19] See also Brey (2008) for other arguments against Floridi's ontocentrism. Brey rejects the idea that everything that exists has an intrinsic moral worth, but he suggests that inanimate things have a potential extrinsic, instrumental, or emotional value for persons. This argument falls back to something similar to either indirect duties or the relational argument we discussed in the previous sections of this paper.

> and behaviour. This *ontological equality principle* means that any form of reality (any instance of information/*being*), simply for the fact of being what it is, enjoys a minimal, initial, overridable, equal right to exist and develop in a way, which is appropriate to its nature. [emphasis in the original]

Floridi introduces an "ontological equality principle," whose formulation suggests some form of equality in the consideration of these entities. Elsewhere, he addresses the question directly, and claims that data have moral worth, that "they need not be viewed as someone's property in order for their unauthorized alteration to be ethically bad" (Floridi and Sanders 2004, 371). However, he also writes in the long quote above that all information beings should only be respected "in a minimal and overridable sense." We have more than indirect duties towards information objects, but this does not mean "that any destruction of data is evil, any more than it would mean that any destruction of life (deemed to have moral worth) in the real world is automatically evil. It simply means that the ethics of altering data in Cyberspace must be considered" (Floridi and Sanders 2004, 371). This make it clear that not all beings should be considered on par with all humans, but it leaves the question open as to what the proper moral status of each being is, in relation to other beings. Is a stone worth less than a book or a cat? Is a rare rock that detached from a comet worth more, morally speaking, than a more common rock on our planet? It is probably safe to say that we should not prioritize the interests of a cat against its human master, but even that question is not clearly answered. Thus, we will leave ontocentrism aside because the view might be self-defeating, in the sense that it does not help solve these questions.

We are left with the two first versions of the argument from life. Regarding the first version, we have not found good reasons to agree with the view that all living entities should have a moral status under a narrow definition of life—that is, reasons different than those considered in the other sections of this paper—but we have not found any good reason to disagree with that view either. Thus, let us grant that all living systems *may* have an intrinsic worth for the sake of the argument.

If, however, one accepts that all living systems have a moral status according to a 'narrow' definition of life, then we do not see why one would not also accept the second, 'broader' version of the argument. The boundaries of life are difficult to establish, and it is not clear why artificial entities that exhibit similar features, such as goal-oriented behavior, should not be included in the circle of moral consideration. Thus, it seems plausible that "living" AI systems, or networks of AI systems, could have a moral status similar to that of a plant or an ecosystem.

However, there is still an unresolved issue with this argument given that it is not clear how we should prioritize an AI system over other entities with a moral status. Similar questions also arise within the biocentrist view. After scrupulously examining different arguments for the moral status of nonhuman entities, the Swiss Ethics Committee on Non-Human Biotechnology (ECNH) also upheld the biocentric claim and considers "arbitrary harm caused to [a] plant to be morally impermissible. This kind of treatment would include, e.g., decapitation of wildflowers at the roadside without rational reason" (ECHN 2008, 20). But according to the committee, the reasons not to wrong living beings are only *prima facie* reasons and can be easily overridden by other considerations, as implied by the notion of an 'arbitrary harm.' This suggests we can prioritize human beings—or other entities, such as animals—over plants, and that our obligations towards these plants may be extremely minimal. Following the same logic, the argument from life applied to an AI system suggests there is some form of ranking in the moral consideration of different entities, and that our obligations towards an AI could be easily overridden by other considerations. It follows that the argument may ground only a minimal moral status, one that is easily overridable.

6 The argument from sentience

The final argument we want to consider runs as follows: if an AI system is sentient, then it ought to be granted moral status. Sentience is the ability to have subjective experience, which includes perceiving and experiencing things (Broom 2016). The concept has been developed to distinguish the ability to *think* and the ability to *feel*. Roughly, being sentient is similar to being conscious, even if you can imagine an entity having conscious mental states likes perceptions, without feeling anything. Sentience is a cognitive ability, which is very common in the animal realm. The *Cambridge Declaration on Consciousness* (Low et al. 2012) states that "the weight of evidence indicates that humans are not unique in possessing the neurological substrates that generate consciousness. Nonhuman animals, including all mammals and birds, and many other creatures, including octopuses, also possess these neurological substrates."

Not all animals are sentient: for instance, there are reasons to think that mussels, oysters or sea sponges are not sentient, and plants are not considered sentient. They notably lack the complex nervous system we find in vertebrates, even if plants seem able to communicate and display a certain kind of intelligence. Other species, such as insects, will fall in a grey zone between nonsentient and sentient beings.

In the same way that biocentrism upholds that all living entities should have a moral status, sentientism (or pathocentrism) upholds that all sentient beings should have a moral

status (Giroux and Larue 2015). Sentientism extends the set of entities with a moral status to many nonhuman animals, but excludes plants and ecosystems. The view has its roots in the basic normative intuition that, because sentient beings have positive and negative feelings, they can be wronged. In other words, sentientism affirms that, because sentient beings have a subjective experience of the world, they can be affected positively or negatively. For instance, sentient animals may have an interest to not suffer, to stay alive, or to be free: this gives us (strong) reasons to act towards them in their own right and for their sake. That is why they deserve a moral status.[20]

In his book *Consciousness and Moral Status*, Joshua Shepherd (2018) provides a comprehensive explanation as to why sentience is the foundation of moral status. He first points out that we should not be biased by the assumption that healthy adult humans are the paradigm case of entities with moral status (15). Thought experiments about 'phenolectomy' or 'zombification' (where an entity is deprived of the ability to feel) suggest that we highly value phenomenal consciousness, even if we cannot conceive it as fully separate from functional features like believing, desiring, or hoping (21–2). We value the conscious mental life of an entity because we ascribe a certain kind of awareness to that entity. In other words, we affirm that there is something it is like for this entity to be aware of the things of which it is aware. For Shepherd, the core of this valuation comes from the affective mental state that consciousness allows. He claims that it "is necessary and sufficient for the presence of some (non-derivative) value in a subject's mental life that the mental life contains episodes with essentially affective evaluative phenomenal properties" (35). Note that these affective evaluative phenomenal properties can go beyond pleasure and pain: emotions such as fear, surprise, or disgust could make an entity aware.

The argument from sentience is also grounded in the Aristotelian principle of equality suggesting that similar cases should be treated in a similar way. If an AI system is sentient, we should give it the same moral status as other sentient beings, like animals. There are many illustrations in the literature to show that proponents of the argument follow a similar logic. For instance, Johnson and Verdicchio (2018) claim that sentience is indeed the right criteria for the moral status of AI (and they add that "animals are sentient beings and robots are not"). Bloom and Harris (2018) claim: "If we did create conscious beings, conventional morality tells us that it would be wrong to harm them—precisely to the degree that they are conscious, and can suffer or be deprived of happiness."

Finally, Bostrom and Yudkowsky (2014) apply something similar to the Aristotelian principle when they introduce what they call the '*Principle of Substrate Non-Discrimination*:' "if two beings have the same functionality and the same conscious experience, and differ only in the substrate of their implementation, then they have the same moral status." In the case of living AIs, we saw that it is possible to raise the objection that entities *produced* artificially are different from natural ones. In the same way, it is possible to argue that sentient AIs should not have the same moral status as sentient animals, because they are artificially produced. Therefore, to be convincing, the argument of sentience needs a Principle of Ontogeny Non-Discrimination: "if two beings have the same functionality and consciousness experience, and differ only in how they came into existence, then they have the same moral stratus" (Bostrom and Yudkowsky 2014). However, this principle does not seem controversial: the way in which a human is conceived (by sexual intercourse or by in vitro fertilization) makes no difference to his or her moral status.

Of course, empirical questions arise quickly: is sentient AI technically possible? Are some existing AI systems already sentient? Right now, there is no reason to believe that we have created a sentient AI. Some entity, such as Shelly, the robot tortoise designed to mimic pain (see section II above), or some simulated character in a video game, can react *as if* it were sentient in some way, but that is pure mimicry, without authentic feeling. In our view, there is no fully convincing proof that there will ever be a sentient AI system. Nor is there anything making it impossible (Dehaene et al. 2017). If we admit that the brain is a sort of (wet) machine, there are reasons to think that building a sentient machine from scratch is a technical challenge that can be met. It may be a more difficult step to make it possible for a "pure" AI system without a "body" to be sentient, but who knows? In any case, the possibility or impossibility of sentient AI is not a valid objection against the argument. The fact that an entity is not sentient at a given point in time is not a reason not to grant a moral status to that entity if it becomes sentient at a later point in time.

Another objection comes from the other minds problem. As Nagel (1987, 26) puts it, what "can you really know about the conscious life in this world beyond the fact that you yourself have a conscious mind?" We cannot know what it is like to be a bat; in the same way, we cannot know what it is like to be an AI. It may even be argued that it is more difficult to take the perspective of an AI than a bat because we share at least some biological element with the bat. How, then, can we be sure that a particular AI is really sentient? As Singer and Sagan (2009) note, the "hard question, of course, is how we could tell that a robot really was conscious, and not just

[20] Neely (2014) goes further and invites us to include all beings that have an interest, a criterion that she considers to be more inclusive than sentience.

designed to mimic consciousness." Answering this question with certainty requires that we better understand the mechanisms of consciousness—in particular, that we understand them at a level of abstraction that allows generalizing to non-carbon-based entities.

For Gunkel (2012, 141), this is an objection to the argument from sentience. Why? Because the other minds problem means that an ethics centered on moral patiency fails to adequately ground its claim from an epistemological perspective: "if something looks like it is in pain, we are, in the final analysis, unable to decide with certainty whether it is really in pain or not." However, the fact that it raises epistemic difficulties to assess whether something is sentient does not imply that sentience is not the right criterion for moral status. As a matter of fact, we must be careful to not create what Sebo (2018) calls the "moral problem of other minds": we do have to make moral decisions in cases of uncertainty about sentience, whether we are considering a lobster or an AI. However, changing the criteria of moral status (ignoring sentience) to avoid this uncertainty seems not to be a valid response to the challenge.

One could also reject the implications of the other minds problem altogether. Danaher (2019, 1, emphasis in the original) proposes a theory that he coined "ethical behaviorism," which states that "robots can have significant moral status if they are *roughly performatively equivalent* to other entities that have significant moral status." Applying this theory to the sentience criterion implies that "what's going on 'on the inside' *does not matter* from an ethical perspective." If an AI system behaves as if it is in pain, it is enough to grant a moral status to that system. However, ethical behaviorism does not imply that we must give moral status to an AI system designed to mimic pain, such as Shelly the robot tortoise. As Danaher (2019, 6) explains, 'behavior' should be interpreted broadly, with "all external observable patterns, including functional operations of the brain."[21] In fact, on Danaher's view, the normative question becomes the question of our standards for assessing a performative similarity and of our capacity to discover the functional operations of the brain.

Finally, one may object that, unlike a human born from in vitro fertilization, a sentient AI could be artificially designed *to* accomplish a specific function or a *role,* and they cannot have a moral status for that reason. Similarly, one could argue that it is acceptable to exploit farm animals because they are raised *for* that purpose, because that is the function we assigned to them. This thinking is misleading because—in cases of both sentient animals and sentient AI system—this does not follow the Aristotelian principle of equality. Indeed, even if a child were conceived to become a slave, that would not morally justify his or her exploitation. Among other things, Bostrom and Yudkowsky's Principle of Ontogeny Non-Discrimination should include a Non-Discrimination Principle regarding the intention of the parents, breeders, or designers of an entity.

Contrary to the arguments from life and information, the argument from sentience does not suggest we should grant a weak moral status to an AI system. Rather, an AI system could be considered on par with a human being if it were equally sentient. In the case of the Turing triage test, it is clear from the sentience argument that a sentient AI should have the same moral status as the sentient human plugged into a survival machine. Drawing of lots would then be a morally acceptable way the solve the dilemma, although there may be a lot of other relevant moral considerations to decide who or what should survive (e.g., survival prognostics, relational argument).

7 Conclusion

In the television show *Westworld*, created by Jonathan Nolan and Lisa Joy (2016), android robot 'hosts' cater to high-paying human 'guests' in a technologically advanced Wild-West-themed amusement park. The guests can indulge their wildest fantasies with the hosts, without fear of retaliation, because the robots are programmed not to harm humans. According to Jaquet and Cova (2018), it makes a great difference for the viewers when they realize (spoiler alert!) that the hosts are conscious. Since "they are conscious, the hosts are plainly members of the moral community, alongside human beings and other animals" (225).

Using the argument from sentience to defend the moral status of AI systems is a common theme in science fiction stories, and this shows that the argument captures widely held judgments. We end up with a similar view because the argument from sentience avoids the issues we identified with the other arguments this paper discussed. First, and contrary to the idea of indirect duties, the argument from sentience could lead us to consider an AI system for its own sake. Second, this does not raise the same issues of consistency or objectivity we identified in the relational argument. Third, sentience is also a relevant criterion of moral status for human beings, which is not the case with the property of being intelligent. Fourth, and although one can grant a moral status to an AI system on the basis that this system is a living or information entity, the argument may ground only a minimal moral status, one that is easily overridable. The argument from sentience, however, could lead us to grant moral status to an AI system at the point where it would be

[21] This theory has the advantage of respecting our epistemic limits—the main disadvantage being that it implies that we ought to treat philosophical zombies as human beings, which is another problem that would need to be further discussed.

considered equal to a human in Sparrow's triage test. The argument from sentience is the strongest for defending the moral status of an AI.

Ironically, however, the sentience argument has more important short-term implications for the way in which we treat nonhuman animals than AI systems. There is no reason to believe that any AI system is sentient today in such a way that it could be wronged for its own sake, but there are strong reasons to believe that other entities, like nonhuman animals, do have a subjective experience of the world, and that we are currently treating them in a way that we simply cannot justify.

References

Ackerman E (2018) Robotic tortoise helps kids to learn that robot abuse is a bad thing. IEEE spectrum, March 14 2018. https://spectrum.ieee.org/automaton/robotics/robotics-hardware/shelly-robotic-tortoise-helps-kids-learn-that-robot-abuse-is-a-bad-thing.

Aristotle (2000) Nicomachean ethics. Translated by Roger Crisp. Cambridge: Cambridge University Press.

Bedau MA, Cleland CE (eds) (2010) The Nature of life: classical and contemporary perspectives from philosophy and science. Cambridge University Press, Cambridge

Bloom P, Harris S (2018) It's Westworld: what's wrong with cruelty to robots? The New York Times, April 23, 2018, sec. Opinion. https://www.nytimes.com/2018/04/23/opinion/westworld-conscious-robots-morality.html.

Boden MA (1996) The philosophy of artificial life. Oxford University Press, Oxford

Bostrom N, Yudkowsky E (2014) The ethics of artificial intelligence. In: Frankish K (ed). The cambridge handbook of artificial intelligence

Brey P (2008) Do we have moral duties towards information objects? Ethics Inf Technol 10(2):109–114. https://doi.org/10.1007/s10676-008-9170-x

Bringsjord S, Govindarajulu NS (2018) Artificial intelligence. In: Zalta EN (ed) The stanford encyclopedia of philosophy (Summer 2020 edition). https://plato.stanford.edu/archives/sum2020/entries/artificial-intelligence/

Broom DM (2016) Considering animals' feelings: précis of sentience and animal welfare. Anim Sentience 2016:005

Bryson J. 2019. The past decade and future of AI's impact on society. In: Towards a new enlightenment: a transcendent decade (Turner-BVVA, pp. 127–169). https://www.bbvaopenmind.com/en/books/towards-a-new-enlightenment-a-transcendentdecade/

Coeckelbergh M (2010) Robot rights? Towards a social-relational justification of moral consideration. Ethics Inf Technol 12(3):209–221. https://doi.org/10.1007/s10676-010-9235-5

Cranor C (1975) Toward a theory of respect for persons. Am Philos Q 12(4):309–319

Danaher J (2017) The symbolic-consequences argument in the sex robot debate. In: Danaher J (ed) Robot sex. The MIT Press, Cambridge

Danaher J (2019) Welcoming robots into the moral circle: a defence of ethical behaviourism. Sci Eng 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 AM, Kerr I (eds) Robot law. Edward Elgar, Cheltenham

Dehaene S, Lau H, Kouider S (2017) What is consciousness, and could machines have it? Science 358:486–492

Ethics Committee on Non-Human Biotechnology (ECHN) (2008) The dignity of living beings with regard to plants. https://www.ekah.admin.ch/inhalte/ekah-dateien/dokumentation/publikationen/e-Broschure-Wurde-Pflanze-2008.pdf

Floridi L (2010) Information: a very short introduction. Oxford University Press, New York

Floridi L, Sanders JW (2002) Mapping the foundationalist debate in computer ethics. Ethics Inf Technol 4(1):1–9. https://doi.org/10.1023/A:1015209807065

Floridi L, Sanders JW (2004) On the morality of artificial agents. Mind Mach 14(August):349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Frankena WK (1986) The ethics of respect for persons. Philos Top 14(2):149–167

Gerdes A (2016) The issue of moral consideration in robot ethics. ACM Sigcas Comput Soc 45(3):274–279

Giroux V, Larue R (2015) Pathocentrisme. In: Bourg D, Papaux A (eds) Dictionnaire de la pensée écologique. Presses universitaires de France, Paris

Gonzalez R (2018) Hey Alexa, what are you doing to my kid's brain? Wired Magazine, 05/11/2018.

Gruen L (2017) The moral status of animals. In: Edward N, Fall Z (eds) The stanford encyclopedia of philosophy 2017. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2017/entries/moral-animal/.

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

Gunkel DJ (2014) A vindication of the rights of machines. Philos Technol 27(1):113–132. https://doi.org/10.1007/s13347-013-0121-z

Gunkel DJ (2018) Robot rights. The MIT Press, Cambridge

Harvey G (2005) Animism: respecting the living world. Columbia University Press, New York

Heams T (2019) Infravies: Le vivant sans frontières. Le Seuil, Paris

Hester L, McPherson D, Booth A, Cheney J (2000) Indigenous worlds and Callicott's land ethic. Environ Ethics. https://doi.org/10.5840/enviroethics200022318

Hill TE Jr (1993) Donagan's kant. Ethics 104(1):22

Hill RK (2016) What an algorithm is. Philos Technol 29(1):35–59. https://doi.org/10.1007/s13347-014-0184-5 (**ABI/INFORM Collection**)

Hogan K (2017) Is the machine question the same question as the animal question? Ethics Inf Technol 19(1):29–38. https://doi.org/10.1007/s10676-017-9418-4

Hoy MB (2018) Alexa, siri, cortana, and more: an introduction to voice assistants. Med Ref Serv Q 37(1):81–88

Jaquet F, Cova F (2018) Of hosts and men: westworld and speciesism. In: South JB, Engels KS, Irwin W (eds) Westworld and philosophy: if you go looking for the truth, get the whole thing. Wiley-Blackwell, Hoboken, pp 219–228

Jaworska A, Tannenbaum J (2018) The grounds of moral status. In: Zalta EN (ed) The stanford encyclopedia of philosophy (Spring 2018 Edition).

Johnson D, Verdicchio M (2018) Why robots should not be treated like animals. Ethics Inf Technol Arch 20(4):291–301

Kamm FM (2007) Intricate ethics. Oxford University Press, New York

Kant I (1785) Groundwork of the metaphysics of morals. Gregor M (ed). Cambridge University Press. https://doi.org/10.1017/CBO9780511809590

Kant I (1997) Moral philosophy: collin's lecture notes. In: Lectures on Ethics (Cambridge Edition of the Works of Immanuel Kant), Heath P, Schneewind JB (ed. and trans.), Cambridge: Cambridge

University Press, pp. 37–222. Original is *Anthropologie in pragmatischer Hinsicht*, published in the standard *Akademie der Wissenschaften* edition, volume 27. https://doi.org/10.1017/CBO9781107049512
Korsgaard CM (1996) Sources of normativity. [S.l.]: Cambridge University Press. http://myaccess.library.utoronto.ca/login?url=http://books.scholarsportal.info/viewdoc.html?id=/ebooks/ebooks1/cambridgeonline/2012-07-31/1/9780511554476.
Korsgaard CM (2018) Fellow creatures: our obligations to the other animals. Uehiro series in practical ethics. Oxford University Press, Oxford
Legg S, Hutter M (2007) Universal intelligence: a definition of machine intelligence. Minds Mach 17:391–444
Little MO (1999) Abortion, intimacy, and the duty to gestate. Ethical Theory Moral Pract 2(3):295–312. https://doi.org/10.1023/A:1009955129773
Loh J (2018) Maschinenethik und Roboterethik. In: Bendel O (Hrsg.): *Handbuch Maschinenethik*. Wiesbaden: Springer VS (2018). pp 75–93.
Low P et al (2012) The cambridge declaration on consciousness. In: Publicly proclaimed in Cambridge, UK, on July 7, 2012, at the Francis Crick Memorial Conference on Consciousness in Human and Non-Human Animals.
Martin D (2017) Who should decide how machines make morally laden decisions? Sci Eng Ethics 23:951–967. https://doi.org/10.1007/s11948-016-9833-7
Müller VC (2020) Ethics of artificial intelligence and robotics. In: Zalta EN (ed) The stanford encyclopedia of philosophy (Winter 2020). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2020/entries/ethics-ai/
Nagel T (1987) What does it all mean? Oxford University Press
Neely EL (2014) Machines and the moral community. Philos Technol 27(1):97–111. https://doi.org/10.1007/s13347-013-0114-y
Nolan J, Joy L (2016) Westworld. HBO. http://www.imdb.com/title/tt0475784/.
Pettit P (2000) Non-consequentialism and universalizability. Philos Quart 50:175–190
Quinn W (1984) Abortion: identity and loss. Philos Public Aff 13(1):24–54
Rawls J (1971) A theory of justice. Belknap Press of Harvard University Press, Cambridge
Scanlon T (1998) What we owe to each other. Belknap Press of Harvard University Press, Cambridge
Sebo J (2018) The moral problem of other minds. Harv Rev Philos 25:51–70. https://doi.org/10.5840/harvardreview20185913
Shepherd J (2018) Consciousness and moral status. Routledge
Sherwin S (2009) Relational existence and termination of lives: when embodiment precludes agency. In: Campbell S, Meynell L, Sherwin S (eds) Embodiment and agency. Pennsylvania State University Press, University Park, pp 145–163
Shue H (1988) Mediating duties. Ethics 98(4):687–704
Singer P (2011) Practical ethics, 3rd edn. Cambridge University Press, New York
Singer P, Sagan A (2009) When robots have feelings. The Guardian, December 14, 2009.
Sparrow R (2004) The turing triage test. Ethics Inf Technol 6(4):203–213. https://doi.org/10.1007/s10676-004-6491-2
Sparrow R (2012) Can machines be people? Reflections on the turing triage test. In: Lin P, Abney K, Bekey G (eds) Robot ethics: the ethical and social implications of Robotics. MIT Press, Cambridge, MA, pp 301–315
Stone CD (1985) Should trees have standing revisited: how far will law and morals reach—a pluralist perspective. Southern California Law Rev 1(1986):1–156
Stone J (1987) Why potentiality matters. Can. J Philos 17(4):815 (**Periodicals Archive Online**)
Taylor P (1981) The ethics of respect for nature. Environ Ethics 3:197–218
Tegmark M. 2017. *Life 3.0: Being Human in the Age of Artificial Intelligence*. New York: Knopf.
Victor D (2017) Hitchhiking robot, safe in several countries, meets its end in Philadelphia. The New York Times, December 21, 2017, sec. U.S. https://www.nytimes.com/2015/08/04/us/hitchhiking-robot-safe-in-several-countries-meets-its-end-in-philadelphia.html
Warren MA (1997) Moral status: obligations to persons and other living things. Clarendon Press, Oxford
Wilson S (2002) Indirect duties to animals. J Value Inquiry 36(1):17–27. https://doi.org/10.1023/A:1014972803058
Wood A (2009) Duties to oneself, duties of respect to others. In: Hill TE Jr (ed) The blackwell guide to kant's ethics. Blackwell, Oxford

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.