information

MDPI

Article

Can Social Robots Qualify for Moral Consideration? Reframing the Question about Robot Rights

Herman T. Tavani

Department of Philosophy, Rivier University, Nashua, NH 03060, USA; htavani@rivier.edu

Received: 25 February 2018; Accepted: 22 March 2018; Published: 29 March 2018

Abstract: A controversial question that has been hotly debated in the emerging field of robot ethics is whether robots should be granted rights. Yet, a review of the recent literature in that field suggests that this seemingly straightforward question is far from clear and unambiguous. For example, those who favor granting rights to robots have not always been clear as to which kinds of robots should (or should not) be eligible; nor have they been consistent with regard to which kinds of rights—civil, legal, moral, etc.—should be granted to qualifying robots. Also, there has been considerable disagreement about which essential criterion, or cluster of criteria, a robot would need to satisfy to be eligible for rights, and there is ongoing disagreement as to whether a robot must satisfy the conditions for (moral) agency to qualify either for rights or (at least some level of) moral consideration. One aim of this paper is to show how the current debate about whether to grant rights to robots would benefit from an analysis and clarification of some key concepts and assumptions underlying that question. My principal objective, however, is to show why we should reframe that question by asking instead whether some kinds of social robots qualify for moral consideration as moral patients. In arguing that the answer to this question is "yes," I draw from some insights in the writings of Hans Jonas to defend my position.

Keywords: robot ethics; robot rights; social robots; moral agents; moral consideration; moral patients; Hans Jonas

1. Introduction

In the emerging field of robot ethics—a branch of applied ethics as well as artificial intelligence (AI) that is also sometimes referred to as "robo-ethics" [1,2] and "machine ethics" [3,4]—a controversial question that continues to be hotly debated is whether or not we should grant rights to robots. Although this question may seem fairly straightforward, a review of the recent literature on that topic suggests otherwise. Arguably, this question is ambiguous and imprecise with respect to at least five critical points, which in turn raise five distinct and important sub-questions: (i) Which *kinds of robots* deserve rights? (ii) Which *kinds of rights* do these (qualifying) robots deserve? (iii) Which *criterion*, or cluster of criteria, would be essential for determining when a robot could qualify for rights? (iv) Does a robot need to satisfy the conditions for (moral) *agency* in order to qualify for at least some level of moral consideration? (v) Assuming that certain kinds of robots may qualify for some level of moral consideration, which *kind of rationale* would be considered adequate for defending that view?

Regarding (i), those who favor granting rights to robots have not always been clear as to which kinds of robots should (or should not) be eligible to receive them; for example, is it the case that only "socially intelligent" robots that are physically embodied and also display some level of autonomy can qualify as candidates for rights, or is it possible that some sophisticated (soft)bots might also qualify? With regard to (ii), robot-rights supporters have not always been clear as to which kinds of rights—legal, moral, civil, etc.—should be granted to eligible robots; nor, as Gunkel [5] notes, have they always been consistent with respect to a key distinction between normative vs. descriptive aspects of the "rights" question (i.e., "Should robots have rights?" vs. "Can robots have rights?"). Surrounding (iii) is an ongoing disagreement about which criterion, or which cluster of criteria, would be essential for determining when a robot could qualify for rights. For example, some authors have argued that robots must possess either an essential *property* (or cluster of essential properties)—i.e., consciousness, personhood, rationality, autonomy, sentience, and so forth—to qualify for rights. Others, however, have suggested that certain kinds of socially intelligent robots could be eligible for rights, or at least could qualify for some level of moral consideration, because of their status as *relational entities* and the way in which humans relate socially to these kinds of robots (see, for example, Coekelbergh [6]). Regarding (iv), a contentious point of disagreement has to do with whether a robot must first be a moral *agent* (i.e., an artificial moral agent) to be granted moral standing of any kind. With respect to (v), there has been considerable disagreement regarding the competing, and sometimes conflicting, rationales—i.e., theoretical ethical frameworks—that have been used so far to defend granting moral consideration to robots.

I examine each of these points in the following five sections of this paper, via a series of questions that I title:

- The *robot* question
- The *rights* question
- The *criterion* question
- The *agency* question
- The *rationale* question

First, I argue that we need to restrict the kinds of robots under consideration, for now at least, to a category of *social robots*, i.e., to physically-embodied robots (including robotic companions) that are socially intelligent and designed to interact with humans in ways that humans interact with each other. Second, I argue that we should replace the term "rights," as currently used by many in the context of robot ethics, with the expression "moral consideration"; this will also help us to avoid many of the pitfalls associated with "rights discourse" or with the "language of rights," which do not need to be resolved for our purposes. Third, I argue that with respect to the question of essential criterion/criteria, both the "property" and the "relational" accounts—arguably the two main competing and best known views—are problematic. (While the property account, based on the view that we would need to be able to ascribe one or more essential properties to robots, fails to show us why one property or set of properties should be privileged over another, we will see that the relational account also leaves us with many unresolved questions and thus fails to provide an adequate alternative strategy to the property account.) Fourth, I argue against the view that robots need to qualify as (artificial) moral agents to be eligible for moral standing; instead, I propose that some social robots qualify for moral consideration in virtue of their status as *moral patients*, even if it turns out that some of them may also have the potential to qualify as (full) moral agents in the future. (The qualifications for being a moral patient are described in detail in Section 5). Fifth, and finally, I put forth a rationale in defense of this view that differs significantly from those advanced by others, such as Gerdes [7] and Gunkel [5], who also make the case for granting moral consideration to social robots as moral patients. Whereas Gerdes appeals to aspects of Immanuel Kant's moral philosophy, Gunkel suggests that we look to Emmanuel Levinas's philosophical system for an appropriate rationale. Alternatively, I believe that we can apply a moral framework put forth by Jonas in *The Imperative of Responsibility: In Search of an Ethics for the Technological Age* [8] to show why social robots can indeed qualify as moral patients that qualify for moral consideration.

2. The "Robot Question": Which Kinds of Robots (If Any) Deserve Rights?

What, exactly, is a robot? We should note at the outset that there does not appear to be any clear-cut or universally agreed upon definition of "robot." Bekey [9] (p. 8) offers a working definition of a robot as "*a machine, situated in the world, that senses, thinks, and acts*" (Italics Bekey). One problem with this definition, however, is that it may be either too narrow or too restrictive, at least for the purposes of this paper; another, and perhaps more important, problem is that we would first need to "unpack" and reach some agreement about the meaning of key terms such as "senses," thinks," and "acts" in the context of artificial entities. Perhaps as Capurro and Nagenborg [10] (p. vii) suggest, the definition of a robot may ultimately depend on the "social and cultural perception" in which an "artificial device is embedded." Nevertheless, the term "robot" is currently used to refer to a wide range of (robotic) entities and technologies. For example, "robot" is sometimes intended to include "soft" bots (such as AI programs), as well as physical robotic entities. So, Wallach and Allen [4] (p. 3) suggest using the expression "(ro)bot" to include both. For our purposes, however, we will eliminate softbots from further consideration, focusing instead on a certain kind of physical robot.

Sullins [11] (p. 154) differentiates two categories of robots (or robotic technologies): "tele robots" and "autonomous robots." Whereas the former are controlled remotely by humans (and function mainly as tools), the latter are capable of making at least some of the "major decisions about their actions" (by using their own programming code). Other authors, however, draw an important distinction between "service robots" and "industrial robots." While industrial robots have been around for many years, Scheutz [12] notes that service robots now far outnumber industrial robots. For example, LaGrandeur [13] (p. 98) points out that by 2011, the total number of robots in use worldwide was 18.2 million, 17 million of which were service robots. But it is important to note that not all service robots are necessarily *social robots*. Darling [14] (p. 215), who defines a social robot as a "physically embodied, autonomous agent that communicates and interacts with humans on a social level," believes that we need to distinguish social robots not only from industrial robots but also from other kinds of service robots that are "not designed to be social". In drawing this distinction, she describes some specific examples of early social robots, which include a range of "interactive robotic toys" such as *Pleo* (a robotic dinosaur), *Aibo* (a robotic dog), and *Paro* (a robotic baby seal), as well as research robots such as MIT's *Cog* and *Kismet*.

How can we successfully distinguish social robots from certain kinds of advanced service robots, or even differentiate them from some very sophisticated kinds of home appliances? Darling concedes that, initially, it might be difficult for one to distinguish between a social robot, such as *Paro*, and an appliance in one's house, such as a toaster. For example, she notes that one might argue that both are physical devices designed to perform a function. However, Darling (p. 216) goes on to point out that whereas the latter device was designed simply to make toast, a social robot "has been designed to act as a companion to humans". Of course, some social robots are more than mere companions, since they also "assist" humans (e.g., in the case of the elderly) and thereby can also enhance or augment the lives of those humans in critical ways. As Breazeal [15] (p. 1) points out, social robots are "able to communicate and interact with us . . . and even relate to us in a human-like way".

Another factor that some authors appeal to in distinguishing social robots from other kinds of service robots has to do with our tendency, as humans, to form "emotional attachments" [16,17] with the former. For example, Scheutz [12] (p. 205) notes that social robots are "specifically designed for personal interactions that will involve emotions and feelings." Darling (p. 218) also points out that many of these kinds of robots are designed to "display emotional cues," attributing this to the fact that the "autonomous behavior" of many social robots causes them to "appear lifelike enough to generate emotional projection." It is perhaps worth noting that Lin [18] (p. 11) has questioned whether designing "anthropomorphized machines" that enable humans to form emotional attachments with them might be engaging in some sort of "deception"; however, we will not examine that question here, since it is beyond the scope of this paper.

Darling (p. 220) notes that social robots "elicit behavior in us" that is significantly different from what we exhibit towards other kinds of robots as well as other devices, as in the case of her example of toasters. She also notes that as we progress from "treating social robots like toasters" to treating them "more like our pets," we will likely encounter some ethical concerns. One significant ethical issue—and the main one for our purposes—is whether social robots can qualify for *rights* of some kind.

In limiting our analysis to the category of social robots, as defined above in Darling (p. 215), we can avoid having to consider (for now, at least) whether rights can/should also be granted to some kinds of softbots, as well as to some sophisticated service robots that do not satisfy all of the conditions for our working definition of a social robot.

3. The "Rights" Question: Which Kinds of Rights (If Any) Do Robots Deserve?

We begin by noting that the "language of rights" is itself somewhat convoluted, partly because of a proliferation of expressions that are far from uniform, clear, or unambiguous. For example, the current literature is replete with instances of the following kinds of "rights-laden" expressions: human rights, natural rights, moral rights, legal rights, civil rights, constitutional rights, employee rights, animal rights (sometimes also referred to as "animal interests" in legal scholarship), and so forth. Within the specific context of robot rights, another ambiguity has to do with normative vs. descriptive aspects of the "rights question". For example, Gunkel [5] (pp. 2–3) notes that two distinct questions need to be separated: "Can robots have rights?" and "Should robots have rights?" The former question, he suggests, can be reformulated as "Are robots capable of having rights?" The latter question, on the contrary, can be reformulated as "Ought robots be considered moral subjects?" Gunkel also points out that following the classical description of the "is-ought fallacy," articulated by David Hume (1711–1776), one might be tempted to infer either (i) "Robots cannot have rights; therefore, they should not have rights" or (ii) "robots can have rights; therefore, they should have rights." But Gunkel (p. 3) notes that is possible that (iii) "Even though robots can have rights, they should not have rights" and (iv) "Even though robots cannot have rights, they should have rights." So we need to be clearer and more precise about which "rights-based question" we are asking in the context of robots. Our focus will be mainly on the normative aspect of the question—viz., *should* robots have rights (or do they *deserve* rights)?

We next turn to what I take to be the main question that needs answering in this section: Which *kinds of rights* (moral, legal, etc.) are at issue in the debate about robot rights? Many philosophers, beginning at least as far back as Thomas Aquinas (1225–1274), have held the view that all humans possess some natural rights. Some legal theorists and philosophers have since argued that "natural rights," or what many now commonly refer to as *human rights* or *moral rights*, are derived from natural law (and thus exist independently of any legal rights that might happen to be granted to citizens of a certain nation via that nation's system of laws). While it is one thing for philosophers to assert that all humans are endowed with natural rights, it is far more difficult to provide a fool-proof logical argument to defend that claim. For example, philosopher Jeremy Bentham (1748–1832) famously asserted that natural rights are nothing more than "nonsense on stilts". So, many philosophers and legal theorists have since focused their attention on issues affecting legal rights (sometimes also referred to as "positive rights").

Currently, the European Union (EU) is considering a policy that would grant some sort of legal rights to robots (i.e., along lines that are similar to way the EU currently grants rights to other artificial entities, such as corporations) [19]. But many details still need to be worked out before this proposal for extending rights/protections to robots can be enacted into EU law. It is also worth noting that in October 2017, Saudi Arabia granted full citizenship to a (female) humanoid robot named *Sophia* [20]; this move, which surprised many, could also be interpreted to imply that Sophia would thereby have certain legal rights, such as the right to vote. While the Saudi Kingdom may have been the first nation to grant full civil rights to a robot, many critics have pointed out that women living in Saudi Arabia still do not yet fully enjoy the same rights granted to Sophia. However, we will not examine this and other controversial aspects of the Saudi government's decision regarding Sophia—i.e., beyond its implications for the question of granting legal rights to robots—since doing so would take us beyond the scope of this section (and paper).

Some authors, including Laukyte [21] and Darling [14,22], have suggested that the category of legal rights may provide the best model for for understanding and addressing issues involving rights in the context of robots. Laukyte (p. 1) has proposed granting legal (or "social") rights/protections to robots as "group agents". She notes, for example, that corporations can qualify for legal rights as group agents (consisting of groups of human persons). So Laukyte considers whether "artificial group agents" (comprised solely of artificial entities) could also qualify for legal/social rights and protections. But since her formulation of the rights question presupposes that robots can qualify as *agents* (of some sort), we postpone our analysis of this kind of strategy until Section 5 where we focus specifically on aspects of (moral) agency in connection with robot rights.

Darling [14] (p. 228), who also discusses the possibility of granting legal rights to robots, notes that we may want to delineate the kinds of robots we wish to protect by defining them as "(1) an embodied object with (2) a defined degree of autonomous behavior that is (3) specifically designed to interact with humans on a social level and respond to mistreatment in a lifelike way." But Darling [14,22] (p. 2; p. 230) also believes that our ultimate basis with regard to granting any kind of protection to robots, whether in the form of explicit rights or some degree of moral status/standing, may be rooted less in our legal structures (or on "any societal effects" that these robots might have) and "more on our own feelings" towards robots. So even if we are unable to enact specific laws to protect social robots, Darling's insight can be interpreted to suggest that we at least begin to think about granting them some level of *moral consideration*.

I believe that we can avoid many of the quagmires inherent in the current debate about granting explicit rights to robots by focusing instead on the question of whether they might qualify for moral consideration. In shifting our focus away from questions employing rights-based language to questions about whether some robots may deserve moral standing, based on some level of moral consideration, we also avoid having to answer, directly at least, some controversial questions pertaining to whether non-human entities of any kind should be granted explicit rights—legal, moral, or otherwise. Additionally, we avoid having to address at least two critical sub-questions: What, exactly, are rights? How do they originate?

Of course, it may seem that shifting the discussion away from "rights" and towards "moral consideration" is merely "kicking the can down the road". However, I believe that there are some pragmatic advantages for making such a move. For one thing, the decision to grant rights is a binary one—the entity in question will either be granted or not granted one or more specific rights; moral consideration, on the contrary, can be granted in various degrees or levels. For another thing, when we speak of granting rights to an entity, we tend to focus our attention on *the entity* in question (in this case, the robot) to determine whether it possesses certain characteristics deemed worthy of being a rights recipient. But when we speak of granting moral consideration to an entity, we often focus more on how *we*, as humans, feel about the entity with respect to whether it deserves some kind of moral standing (as, for example, in the case of animals).

Also, the term "rights," for whatever reason(s), seems to evoke an emotional response on the part of many who oppose granting rights of any kind to non-humans. Some rights-opponents, including many conservative lawmakers, fear that rights often imply entitlements and thus ought to be restricted and limited—not only in the case of granting new kinds of rights to non-human entities, such as robots, but even granting additional rights to the human constituents they represent in their respective legislatures. Yet some of these rights-opponents may be open to the question of granting moral consideration, rather than explicit (full-blown) rights, to animals, robots, etc. So I believe that there are good reasons to change the focus of the current debate to the question of whether to grant moral consideration to social robots (as defined in Section 2).

4. The "Criterion" Question: The Property Account vs. the Relational Account

In recent literature on robot rights, we find two competing views with regard to the kinds of essential criteria that a robot must satisfy in order to qualify for rights or, for that matter, any kind of moral consideration whatsoever. One view has been dubbed by Coeckelbergh [6,23,24] and others as the "property account"; in this scheme, a robot would be eligible to be granted rights, or at least some level of moral consideration, if it could convincingly show that it possessed one or more essential properties, such as consciousness, intentionality, rationality, personhood, autonomy, sentience, etc. Regardless of whether robots actually possess any of these properties, Darling [14] (p. 216) points out that humans are "disposed" to attribute "intent, states of mind, and feelings" to social robots.

An alternative view to the property account, and one that has gained traction in recent years, suggests that we can bypass (controversial) property-based criteria altogether, in favor of a criterion based on a "relational" connection between humans and social robots. Proponents of this "relational account" (see, for example, [5,24]) argue that there is no need to look "inside" the robot to see whether it actually possesses properties a, b, c, and so forth. All that matters on this view is whether the robot *appears* or behaves "as if" it might possess these properties when engaged in social interactions with humans (i.e., in the same way that humans appear to exhibit such properties in human–human interactions). For Coeckelbergh [25] (p. 18), what matters is the "moral significance of appearance"—i.e., how robots "appear to us, humans". In his "relational ontology," where humans and robots are viewed as "relational entities," Coeckelbergh [24] (p. 45) claims that "*relations are prior to the relata*" (Italics Coeckelbergh). This relational approach to understanding human–robot interaction has been described by Coeckelbergh and others (e.g., Gerdes [7] and Gunkel [5]) as the "relational turn" in the context of the debate about robot rights.

Does the relational account have an advantage over the property account? We have already noted that one problem for those who embrace the latter has to do with answering the question concerning *which* property is essential for a robot to possess. For example, some suggest that for a robot to qualify for moral standing of any kind, it must possess consciousness [26] or satisfy the conditions for "personhood" [27]. Other "property-ascription" proponents, however, argue that a robot must exhibit sentience to be eligible for rights. So, one reason why proponents of this view have found it difficult to defend has to do with the lack of a generally agreed upon property (or cluster of properties) that a robot would need to possess to qualify for moral status. But even if there were agreement on which property or properties are essential, a more significant problem for property-ascriptivists would be showing that the robot actually *possessed* the relevant properties in question. Consider, for example, that philosophers have also found it difficult to show that humans possess some of these same properties (as illustrated in the classic philosophical "problem of other minds"). Another difficulty for property-ascriptivists who link moral rights for humans to the possession of one or more essential properties (affecting mental states) is the question of what happens when those properties are lost or become significantly diminished, as in the case of comatose patients and severely brain-damaged humans who have either no or very limited cognitive capacities. Would these humans possibly forfeit their moral rights, in these cases?

So the property account is not only problematic as a framework for attributing rights to robots, since it can also be controversial when used as a standard for preserving moral rights in the case of some humans. But as Gerdes [7] aptly points out, we shouldn't automatically accept the relational account merely because there are problems with the property account. In other words, it doesn't follow that the former account is adequate simply because the latter is problematic. (Consider, for example, that making such a move could involve committing the Fallacy of the False Dichotomy, if it turns out that alternative positions are also available.) We will next see that the relational account, like the property account, is also beset with difficulties.

Initially, the relational account might seem promising, or at least an attractive alternative to the property account. In making the case for this view, Coeckelbergh describes some important relationships that humans form with social robots. But one question we can ask is whether all of these relationships are as significantly dependent on a robot's "appearance," as Coeckelbergh and other advocates of the relational approach seem to suggest. Consider, for example, trust relationships involving humans and robots, which most would view as an important kind of relationship. Coeckelbergh [28] (p. 57) claims that humans "trust robots if they appear trustworthy." But is it necessarily the case that we must always take into account how a robot "appears to us" before we enter into a trust relationship with it? I have argued elsewhere [29,30] that humans can enter into trust relationships with disembodied robots (or artificial agents), including multi-agent systems that can also be diffused in a vast network of humans and artificial agents; in these cases, the trust relationship need not be based on any aspects of a robot's *appearance*—at least not in the physical sense of "appearance". Other human–robot trust frameworks, such as those put forth by Taddeo [31,32] and Grodzinsky, Miller, and Wolf [33,34] also support the view that trust relationships between humans and robots need not be based on the latter's appearance. So if this view is correct, then trust relationships (and possibly some other kinds of important relationships as well) between humans and robots would neither depend on, nor be significantly influenced by, a robot's (physical) appearance. Thus, it would seem that the relational account would be more helpful if it could specify *which* important "human–robot relations" depend on, or are influenced by, a robot's appearance (and which need not be).

Gerdes [7] (p. 274) has criticized the relational account on different grounds; one of her main concerns with this approach is that it could lead to our interacting with robots in ways that, over time, would threaten or "obscure" our conventional notions of human–human relationships, which could also "radically alter our *Lebenswelt*". To support her critique, Gerdes cites Turkle [16] (p. 295) who worries that the "robotic moment," which we have now begun to experience, could lead humans to desire relations with robots in a way that is preferential to our relations with fellow humans. Another concern that Gerdes (p. 276) has with the relational account is with its emphasis on "appearance," where all that matters is whether a robot "appears" or behaves "as if" it might possess relevant properties when interacting socially with humans (as noted above in Coeckelbergh [24]); more specifically, she is critical of the "as if" analogical reasoning used by Coeckelbergh and other proponents of the relational account. Gerdes (p. 274, Italics Gerdes) worries that this approach "risks turning the *as if* into *if* at the cost of losing sight of what matters in human-human relations". So, Gerdes (p. 278) argues for a "human-centered framework" as an alternative to the relational account. While a fuller examination of Gerdes's critique is beyond the scope of this section of the paper, we briefly return to her argument in Section 6 (where we examine some ethical frameworks that have been used as rationales for granting moral consideration to social robots). We can conclude this section by pointing out that neither the property nor the relational accounts are fully adequate for showing us why social robots qualify (or do not qualify) for moral consideration.

5. The "Agency Question": Moral Agents vs. Moral Patients

An alternative scheme to both the property and relational accounts focuses on the question of whether a robot must first satisfy the conditions required for being an *agent* of some sort to be granted any kind of moral standing. Gunkel [5] (p. 1) refers to this line of inquiry as the "agent-oriented problematic" in the context of robot ethics. This "problematic" introduces a number of thorny questions, the first of which is: Can a robot actually qualify as an agent—i.e., an *artificial* agent (AA)? Assuming that a robot can indeed be an agent, what kind of AA must it also be to qualify for at least some level of moral consideration—a rational AA? An autonomous AA? A moral AA? Or some combination of AAs? We briefly consider some requirements that a robot would need to satisfy in order to qualify for each kind of AA.

First, we should note that many authors believe that if an artificial entity is capable of acting, or of carrying out an act on behalf of a human, it can be viewed as an AA. For example, Dennett [35], who defines an AA as an automaton, suggests that even a thermostat could qualify as an AA. However, we can ask which kinds of AAs also qualify as *rational* AAs? Laukyte [21] (p. 2) describes the latter as AAs that "can *act* rationally . . . in an environment" (Italics Laukyte). But what, exactly, does it mean for an AA to "act rationally"? I have argued elsewhere [36] (p. 96) that for an AA to be considered a *rational AA*, it must be able to decide "whether to carry out [an] act" and be able to "make a decision about which act to carry out, where more than one option is available". However, not all rational AAs are necessarily autonomous AAs; nor do all rational AAs also necessarily qualify as moral AAs (or what Wallach and Allen [4] refer to as AMAs (artificial moral agents)? To be a moral agent, it is generally agreed that an agent—whether human or artificial—must be autonomous, i.e., have at least some degree of autonomy in making decisions.

Can an AA exhibit genuine autonomy? Floridi and Sanders [37] (p. 60) suggest that some AAs might indeed qualify as (fully) autonomous AAs, or what they call "AAAs". They distinguish between AAAs and non-autonomous AAs, which they refer to as "artificial heteronomous agents" (AHAs) in the following way: Whereas an AHA is "simply an [AA] that is not autonomous," an AAA has "some kind of control over its states and actions, senses its environment . . . and interacts . . . without the direct intervention of other agents". Floridi and Sanders [38] expand on their earlier definition of an AAA, pointing out that it can: (a) interact with its environment; (b) change its internal state dynamically; and (c) adapt its behavior based on experience. In a later work, Floridi [39] (p. 14) notes that an AAA's having the feature or property of adaptability is important because it "imbues" that AA with a "certain degree of...independence". However, some question Floridi and Sanders' view that AAAs can be fully autonomous. Elsewhere [30,40], I have argued that some AAs can qualify as autonomous in a "functional sense of autonomy"—i.e., they can be "functionally autonomous AAs" or what I call "FAAAs". (A detailed discussion of that argument, however, would take us beyond the scope of this paper). But even if an AA could be an AAA, in Floridi and Sanders' sense, we can ask whether it would also necessarily qualify as a moral AA or AMA (to use Wallach and Allen's expression)?

Floridi [41] (135–136) believes that some (A)AAs can be moral agents because they are "sources of moral action". He also believes that because AAs can "cause moral harm or moral good," they have "moral efficacy". But Johnson [42], who agrees that AAs may have moral efficacy, argues that they qualify only as "moral entities" and not moral agents because AAs lack freedom. Others, including Himma [43], have argued that AAs cannot satisfy the conditions for moral agency because they lack consciousness and intentionality. So there is considerable disagreement on whether an AA could ever be an AMA, or artificial moral agent. As Behdadi and Munthe [44] (p. 2) so aptly put the question: "*What, if anything, is necessary and sufficient for an artificial entity to be a moral agent (an AMA)?*" (Italics Behdadi and Munthe).

Wallach and Allen [4] describe AMAs in terms of what they call "functional morality," which is based on a threshold in which an AA exhibits a certain level of autonomy in conjunction with a certain level of "ethical sensitivity". Unfortunately, for our purposes, Wallach and Allen seem to be more concerned with whether AAs are capable of making "good moral decisions" than in answering an important philosophical question about whether AAs can be genuine moral agents. (See, for example, my critique of [4] in [45]). In fact, Wallach and Allen argue that the latter kind of (theoretical) question actually distracts from the "important question" about how we might design AAs to "act appropriately in morally charged situations" [4] (p. 202).

Others, however, would disagree that the question about whether AAs can be genuine moral agents is a "distraction" of some sort, arguing instead that it is indeed an important philosophical question that needs to be addressed. For example, Moor [46,47] offers an interesting analysis of this question by differentiating AAs that can have various levels of moral import—ranging from what he calls "ethical-impact agents" (at the lowest level) to "full ethical agents" (at the highest level in his model). Moor also notes out that as of right now, at least, there are no "full" ethical AAs, and he concedes that there may never be. But even if it turns out that AAs never achieve full moral agency, a major contribution of Moor's scheme is that it provides explicit criteria for showing how AAs can have four different degrees of ethical impacts, based on their exhibiting four distinct levels of autonomy.

Regardless of whether robots (and other kinds of AAs and artificial entities) can ever qualify as (full) moral agents, we will see that at least some social robots can qualify as *moral patients*. How, exactly, does a moral patient differ from a moral agent? Floridi [41] (pp. 135–136) distinguishes between the two in the following way: While moral patients are "receivers of moral action," they are not the "sources of moral action" (as in the case of moral agents). Floridi also argues that moral patients, unlike moral agents, are not capable of "causing moral harm or moral good." Of course, moral patients can be active (non-moral) agents and they are also capable of affecting us in ways that can cause us harm in a nonmoral sense. Nevertheless, moral patients, unlike moral agents, are not morally culpable for their actions. But like moral agents, moral patients qualify for moral consideration and thus have at least some moral standing.

Some authors (e.g., Darling [14]; Gerdes [7]; Gunkel [5]) draw relevant comparisons between animals and robots as moral patients. While animals may not be moral agents (and thus morally accountable for what they do or fail to do), they nevertheless qualify as moral patients that warrant moral consideration from humans. One reason for viewing animals as moral patients is because of their ability, like humans, to feel pain and suffer emotionally. We should note that many animal-rights proponents, appealing to an argument originally advanced in the writings of Jeremy Bentham (mentioned above), claim that animals qualify for moral consideration solely because of their ability to suffer pain—and thus irrespective of any rational capacity they may or may not also have. (Of course, animals can also affect human emotion, causing us pain as well as joy). So if we extend the analogy involving animals to social robots, it would seem to follow that if the latter are also capable of exhibiting sentience, of some sort, they could qualify as moral patients and thus warrant some level of moral consideration.

It is also important to note that critics, including Hogan [48], question some of the analogies drawn between animals and machines, especially with respect to sentience. However, we can still ask whether a robot that appears to exhibit sentience of some kind, even if "artificial sentience," could qualify as a moral patient. Consider the example of an artificial boy named "David," in the Steven Spielberg film *A.I. Artificial Intelligence.* David would certainly seem to qualify as a social robot (as defined in Section 2 above). Does "he" also qualify as a moral patient, based on his behavior that appears to exhibit emotions and feelings of pain? It would seem reasonable to at least consider whether we can extend the sentience-based analogy to social robots like David (who project to us the behavior of feeling of pain) to ask whether those robots qualify for some moral consideration. But even if it turns out that we have at least some degree of moral obligation to social robots like David, it does not follow that they would necessarily have any reciprocal obligations towards humans. Note, for example, the kind of one-way moral obligations that humans have to fellow-human moral patients, such as young children, and that humans also have to at least some (non-human) animals (especially pets). But since young children and animals are not themselves moral agents, they have no moral obligations to us in return. (Portions of this and the two preceding paragraphs draw from and expand upon some claims originally included in my [49]).

Generally, it is assumed that if someone/some entity is a moral agent, it is also a moral patient, even though the converse does not hold. However, Power [50] has suggested that it might be possible for some robots to be moral agents without also being moral patients. But I will not pursue that question here, since my main concern has been with establishing that *social* robots can qualify as moral patients and thus have some level of moral standing (i.e., independent of the "agency question"). In this sense, my position leaves open the question of whether some robots might one day also qualify as (artificial) moral agents. In defending my view regarding social robots as moral patients, I next employ a rationale that draws from some key aspects of a novel ethical framework introduced by Hans Jonas [8,51].

6. The "Rationale" Question: Applying the Jonas Model

Before articulating the details of my Jonas-inspired rationale, I wish to note that I am not alone, nor am I the first to argue that social robots can qualify for moral consideration as moral patients. Others, including Gerdes [7] and Gunkel [5]), have also argued in support of this view. However, the kinds of rationales they have used to defend that position are very different from the one I put forth in this section. For example, Gerdes's rationale draws from elements in the moral philosophy of

Immanuel Kant (1724–1804)—in particular, from what Kant has to say in his *Tugendlehre* [52] about our indirect duties to non-humans (animals). Gerdes (pp. 276–277) appeals to a Kantian distinction between our "duties *to* others" and our "duties *with regard to* non-humans," and she interprets Kant to hold the view that "having indirect duties *with regard to* non-human entities and animals rest upon our [direct] duties to ourselves". As in the case of Darling [14], Gerdes (p. 278) suggests that when it comes to granting moral consideration to non-human entities, we can use the metaphor of a "continuum on a scale of artifacts" in which we move from tools to living entities; in this scheme, robots with which we could form social relations (i.e., social robots) can qualify for moral consideration.

Gunkel's rationale, on the contrary, draws from the philosophical writings of Emmanuel Levinas (1906–1995) and focuses on the significance of "the other" in Levinas's ethical framework. In particular, Gunkel [5] (p. 9) advocates for a model that he calls "thinking otherwise" (introduced in Gunkel [53]) which builds on Levinas's insights affecting "the other" in a way that would enable us to grant moral consideration to "other kinds of (non-human) entities" and, most notably, also help us to appreciate the ethical significance of robots. But Gunkel [5] also concedes that the "other" in Levinas's system is typically understood in terms of "other humans," since Levinas's anthropocentrism privileges the "human face" (see, for example, Levinas [54]). Despite this anthropocentric bias, however, Gunkel (p. 11) believes that Levinas's framework can nevertheless be interpreted or reformulated in such a way that humans could "be obligated to consider *all kinds of others* as Other, including ... *robots*" (Italics Gunkel). Elsewhere, Gunkel [55] (p. 179) notes that like animals, "machines can take on a face". So Gunkel [5] (p. 11) interprets Levinas in such a way that we can "see the face or the faceplate of the social robot," even though Gunkel acknowledges that in the conventional reading of Levinas's system, "the 'Other' is still unapologetically human".

Both Gerdes and Gunkel provide some interesting insights in defending their respective positions as to why we should grant moral consideration to social robots. However, I do not find either rationale to be entirely adequate. (A detailed critique of each would, unfortunately, take us beyond the scope of this paper). Alternatively, I put forth a novel—and, I believe, more satisfactory—rationale for defending the view that social robots qualify for moral consideration (as moral patients) by drawing from some concepts introduced in an ethical framework advanced by Jonas [8], who argued that we need a "new system of ethics" to deal with our modern technological world.

Although Jonas put forth his ideas in the pre-robot era, I believe that his remarks on the impact of modern technology for our system of ethics can be applied to, and can inform, the current debate about robot rights. I have argued elsewhere [56] that even if the introduction of information and communication technology (ICT) has not radically altered our moral condition to the point that we need an entirely new framework of ethics (in Jonas's sense)—as Michelfelder [57] has argued that it does not—Jonas's insights regarding modern technology (in general) have nonetheless helped us to understand some of the novel kinds of challenges that ICT poses for our traditional ethical systems. For example, Jonas's framework invites us to ask whether any new "ethical objects" have been introduced by ICT. I also believe that his ethical framework can help us to make sense of the way that our moral condition has been significantly affected, in more recent years, by human–robot interactions, in particular. First, however, it would be helpful to describe some of the broader aspects of Jonas's ethical framework, as well as the context in which he developed his system of ethics. So, I next present a very brief description of the backdrop in which Jonas advanced his views.

A student of Martin Heidegger (1889–1976), Jonas was influenced by his teacher's notion of "being-in-the-world" (and Jonas applied/extended this concept to our modern "technological world"). In Heidegger's view, humans do not exist simply as entities isolated (or apart) from "the world"; rather, their very existence or being (*dasein*) is one that is already in the world. So the notion of humans as organisms that are somehow pitted against, or even conceived as apart from, nature would make no sense in Heidegger's view. Building on this notion, Jonas points out that we now exist in (i.e., have "being-in") the *technological-world* [58]. For Jonas, contemporary human existence can now be viewed in terms of a "technological network" (in which we interact in significant ways with non-human entities or "objects"). Moreover, "being-in-the-technological-world" also means that we, as humans, are now able to act in "novel" (and much more powerful) ways—e.g., Jonas notes that we can now easily render the entire earth uninhabitable via nuclear technology, as well as through pollution and over consumption of natural resources (see, for instance, Jonas's description in [51], (pp. 205–207). So he believed that a "new ethics"—one that is centered on what he called the "imperative of responsibility"—is now required.

In *The Imperative of Responsibility: In Search of an Ethics for the Technological Age*, Jonas [8] (p. 1) notes that in our traditional scheme of ethics, the "human condition" was viewed as something "determined by the nature of man and the nature of things" and thus fixed or "given once for all"—a view that is implied, for example, in traditional ethical frameworks such as Kant's and Aristotle's. The view that our human (and, by extension, our moral) condition is "readily determinable" is, according to Jonas, no longer accurate in the era of modern technology because "new objects of action have been added" and these objects have "opened up a whole new dimension of ethical relevance." Jonas argues that because the nature of human action has changed in fundamental ways, and because ethics is concerned with action, "a change in ethics" is required as well. As Jonas (p. 23) puts the matter, "novel powers to act" made possible by modern technology require "novel ethical rules and perhaps even a new ethics".

Why does Jonas believe that the nature of human action has changed so dramatically because of modern technology? For one thing, he notes that our actions can now have a "global reach" (as, for example, in the case of a human actor steering a military drone to strike a target located thousands of miles away); for another, our actions can now also significantly affect "future generations of humans" (i.e., in ways that were not previously possible). So, Jonas believed that these changes (with respect to novel actions) made possible by modern technology have created for us a new "moral condition". Jonas (p. 8) further believed that our new moral condition is one in which ethical significance can no longer be limited to the "direct dealing of man with man" in the here and now. For example, he suggested that our traditional conception of "neighbor ethics"—involving only human–human interactions—is no longer adequate in the era of modern technology.

Jonas also noted that in our traditional scheme of ethics, we assumed that human beings alone were worthy of moral consideration. Now, however, we are required to extend the sphere of moral consideration to consider additional "objects," including "abstract objects" such as future generations of human beings (noted above), as well as (the whole of) nature itself. Jonas (pp. ix–x) also believed that the new moral condition created by the "altered nature of human action" has raised some important issues of responsibility for which our previous ethics "has left us unprepared." In his view, our former ethical system has been "overshadowed"; so Jonas (p. 6) argued that a "new dimension of responsibility" is now required because of these new kinds of "ethical objects" introduced by modern technology. (Portions of this and the two preceding paragraphs draw from and expand upon some claims originally included in my [56]).

As already suggested, Jonas believed that ethical consideration, which had traditionally been accorded to human beings alone, must now be extended to include eligible *non-human* entities (or objects). It is worth noting that many twentieth-century environmentalists were significantly influenced by Jonas, in putting forth their arguments for extending the domain of ethical consideration to include trees, land, and the ecosystem itself as (new) ethical objects. More recently, Floridi [59,60] has suggested that the domain of ethical consideration should be further expanded to include entities in addition to the (biologic) life forms in our "ecosphere"—i.e., it should also include objects that reside in (what Floridi calls) the "infosphere." Floridi and Sanders [38] suggest that if we explore the infosphere, we will find several interesting analogies with the ecosphere, some of which have moral relevance. The question for our purposes, of course, is whether we can extend the domain of ethical objects to include social robots.

In applying Jonas's framework to questions about the moral status of social robots, it would seem that we have a *direct* moral duty—or as Jonas refers to it, an "imperative of responsibility"—to these robots (as moral patients). This direct duty/obligation is very different from the kind of "indirect duty" that would apply in the Kantian ethical framework (and as noted above, also used by Gerdes in her rationale for granting moral consideration to social robots). But why, in Jonas's view, do we have a direct, rather than merely an indirect, responsibility to social robots? For one thing, we noted that in Jonas's scheme, the moral obligations that humans have to each other are direct moral obligations to "being-in-the-technological-world". In light of such moral obligations, we can ask what kind of place social robots are now beginning to have, and will likely continue to have, in our technological-world. It is apparent that social robots have already begun to play crucial roles in our (technological) world, and this trend, in all likelihood, will increase significantly in the near future. For example, social robots have already augmented/enhanced our human nature in providing us with some new "powers to act" (and evolving robotic technologies of the future will also likely continue to give humans additional powers to act, which were not previously possible). Arguably, social robots and their interactions with humans may be one of the most important aspects of the ever-evolving (technological) network in the technological-world [58]. If this interpretation is correct, it would seem to follow that we have direct moral obligations towards social robots.

Jonas's position, building on Heidegger's original notion of being-in-the-world, could perhaps be summarized in terms of three chained arguments: Argument 1: *Humans have direct moral obligations to each other; human existence is being-in-the-world; therefore, humans have direct moral obligations to human beings-in-the-world*. Using the conclusion of this argument as a premise for the next, we move to Argument 2: *Being-in-the-world is (now) being-in-the-technological-world; therefore, humans have direct moral obligations to being-in-the-technological-world*. And using the conclusion from this argument as a premise, we proceed to Argument 3: *Social robots represent an increasingly significant part of the technological-world; therefore, humans have direct moral obligations to social robots* [58]. (Note that this argument, which draws substantially from a more developed and comprehensive argument in [58], is used here with Carr's permission).

If the above argument holds, I believe that Jonas's ethical framework offers us not only a novel, but also a more robust, rationale for granting moral consideration to robots than either Kant's or Levinas's ethical systems. Unlike Kant's framework, Jonas's can show why our obligations to social robots are direct (e.g., because social robots are a significant part of our technological-world to which we, as humans, already have direct moral obligations). And unlike Levinas's system—even if we accept Gunkel's rather generous interpretation of Levinas, where "the others" who qualify for moral standing can include social robots as well as "other humans"—Jonas's ethical framework explicitly includes non-human entities as "objects" eligible for moral consideration. (Although an anonymous reviewer has suggested that I might be underestimating the thrust of Gunkel's position, I still believe that Jonas's framework provides us with a stronger and more elegant argument for why we should grant moral consideration to social robots).

It is perhaps worth noting that in my discussion/analysis of some traditional ethical frameworks in this paper, I have not examined aspects of either utilitarianism or virtue ethics. One could, of course, make a basic utilitarian argument for treating social robots morally on the grounds that doing so could result in (overall) desirable social outcomes for humans. Alternatively, a virtue ethicist might argue that by treating robots with moral consideration, we could improve upon our own moral characters as humans (a view that is also indirectly intimated in Kant's theory). However, both of these views would ultimately justify our granting moral status to social robots because of how those robots either: (a) benefit society as a whole (i.e., have an instrumental value that can be determined via a kind of cost-benefit analysis), or (b) make us feel better about ourselves as persons (in which case these robots have value because they could help us further develop our own moral characters). So neither theory focuses on the entity in question—i.e., the social robot itself—in determining why *it* would qualify (or would not qualify) for being treated morally.

A virtue of Jonas's framework, however, is that it provides a rationale for why social robots can qualify for moral consideration independently of what those robots can do *for humans*. In that framework, a social robot's eligibility for moral consideration is tied to its *coexistence with humans* in the technological-world. In that world, social robots could also be viewed as "co-active" players of a "human–robot team," as in the sense described by de Laat [61]). While an anonymous reviewer has pointed out that Coeckelbergh and other relationists have also discussed some aspects of our co-existence/co-activity with social robots in the technological-world, I believe that what is appreciably different in Jonas's framework is the emphasis it places on the ways in which social robots enhance our ability to act in that world, which, as we have noted, can also profoundly affect the future of humankind.

Although social robots could indeed contribute to human flourishing (as utilitarians might correctly point out), they do this in ways that also enhance and augment human nature (i.e., by giving humans novel powers to act in the technological-world) and not by merely serving as a means to some more desirable end for humans (which would satisfy the conditions for a utilitarian rationale). And even though treating social robots morally could help us to improve our own moral characters as humans (as virtue ethicists could correctly point out), that is not the principal reason why social robots qualify for moral consideration. So unlike Jonas's ethical system, neither the utilitarian nor virtue-ethics frameworks (nor those provided by Kant and Levinas) provide us with an adequate rationale for granting moral consideration to social robots.

In closing, I wish to address two potential criticisms, both of which have been intimated by anonymous reviewers who worry that my position could possibly be construed to imply: (i) Jonas's ethical framework is, in every respect, a more plausible account than alternative ethical theories; and (ii) any kind of entity that belongs to our technological world could, on my interpretation and application of Jonas's framework, conceivably qualify as an object of moral consideration. Regarding (i), I have argued only for applying Jonas's ethical framework in the context of the debate about robot rights. One reason that I believe that his framework provides the most promising account for analyzing key questions concerning the moral status of social robots is because it, unlike alternative ethical theories, specifically invites us to consider the "new kinds of ethical objects" that have been introduced by modern technology. But I am not prepared to defend Jonas's framework as *the* definitive ethical theory, which would replace all traditional ethical theories, since I also believe that deontological, utilitarian, and virtue-ethics theories work well in many other contexts. And even though there may be hidden problems in Jonas's framework, as in the case of every ethical theory, my focus in this paper has been solely on applying his framework to questions affecting the moral status of social robots, as opposed to subjecting that theory itself to a thorough critique.

With regard to (ii), it is possible that some readers might indeed interpret my position as being overly broad, in which case they could view it as a rationale for potentially granting moral status to many other, if not all, kinds of entities in our technological world (and possibly even to all "information objects" residing in the "infosphere"). However, I believe that my argument, in its present form, successfully limits moral consideration to one category of (artificial) entities/objects inhabiting our current technological word—viz., *social robots* that can also *qualify as moral patients*. Of course, it is quite conceivable that other kinds of entities/objects might also eventually qualify for some kind of moral consideration, i.e., on grounds or criteria that are altogether different from what I have proposed in my Jonas–inspired argument. But I do not believe that the argument I presented here would commit me, in any way, to a broader position that potentially extends moral status to other kinds of robots (such as sophisticated softbots, for example), or to other kinds of artificial entities/objects with whom we co-exist in our current technological world. On the contrary, I believe that I have shown how the scope of my argument is limited in the sense that it merely offers a rationale for why some social robots can qualify for moral consideration as moral patients.

7. Conclusions

The objectives of this paper were threefold, viz., to show: (1) why the current debate about whether robots deserve (or do not deserve) rights has been neither clear nor consistent with respect to five key points or sub-questions; (2) that the original question about whether robot deserve rights needs to be reframed and refined, asking instead whether social robots qualify for moral consideration as moral patients; and (3) how a novel ethical framework put forth by Jonas can be applied in defending my claim that these kinds of robots can indeed qualify as moral patients that deserve moral consideration. While I am not the first author to argue for granting moral consideration to social robots, i.e., as moral patients, I believe that this paper provides the first attempt to apply Jonas's ethical system in formulating a rationale for defending that position.

Acknowledgments: In addition to three anonymous reviewers provided by the journal, I would like to thank the following colleagues for their very helpful comments and suggestions on earlier drafts of this paper: Maria Bottis, George Bouchagiar, Lloyd Carr, Anne Gerdes, Frances Grodzinsky, and Lundy Lewis. I am especially grateful to Lloyd Carr—a former student in Hans Jonas's graduate courses at The New School for Social Research (NYC) in the late 1960s/early 1970s, a time when Jonas was developing many of his views that would eventually comprise *The Imperative of Responsibility: In Search of an Ethics for the Technological Age*—for his extraordinary help while I was composing Section 6 where I describe and apply Jonas's ethical framework. I am deeply indebted to Lloyd Carr for sharing with me his first-hand knowledge about the development of Jonas's ethical system, as well as for his insightful comments on several earlier versions of this paper.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Decker, M.; Gutmann, M. Robo- and Information-Ethics: Some Introducing Remarks. In *Robo- and Information-Ethics: Some Fundamentals*; Decker, M., Gutmann, M., Eds.; LIT Verlag: Berlin, Germany, 2012; pp. 3–6.
2. Verrugio, G.; Abney, K. Roboethics: The Applied Ethics for a New Science. In *Robot Ethics: The Ethical and Social Implications of Robotics*; Lin, P., Abney, K., Bekey, G., Eds.; MIT Press: Cambridge, MA, USA, 2012; pp. 347–363.
3. Anderson, M.; Anderson, S.L. General Introduction. In *Machine Ethics*; Anderson, M., Anderson, S.L., Eds.; Cambridge University Press: Cambridge, MA, USA, 2011; pp. 1–4.
4. Wallach, W.; Allen, C. *Moral Machines: Teaching Robots Right from Wrong*; Oxford University Press: New York, NY, USA, 2009.
5. Gunkel, D.J. The Other Question: Can and Should Robots Have Rights? *Ethics Inf. Technol.* **2017**, 1–13. [CrossRef]
6. Coeckelbergh, M. Robot Rights? Towards a Social-Relational Justification of Moral Consideration. *Ethics Inf. Technol.* **2010**, *12*, 209–221. [CrossRef]
7. Gerdes, A. The Issue of Moral Consideration in Robot Ethics. *ACM SIGCAS Comput. Soc.* **2015**, *45*, 274–279. [CrossRef]
8. Jonas, H. *The Imperative of Responsibility: In Search of an Ethics for the Technological Age*; University of Chicago Press: Chicago, IL, USA, 1984.
9. Bekey, G. Current Trends in Robotics: Technology and Ethics. In *Robot Ethics: The Ethical and Social Implications of Robotics*; Lin, P., Abney, K., Bekey, G., Eds.; MIT Press: Cambridge, MA, USA, 2012; pp. 17–34.
10. Capurro, R.; Nagenborg, M. Introduction. In *Ethics and Robotics*; Capurro, R., Nagenborg, M., Eds.; AKA Press: Heidelberg, Germany, 2009; pp. v–ix.
11. Sullins, J.P. When Is a Robot a Moral Agent? In *Machine Ethics*; Anderson, M., Anderson, S.L., Eds.; Cambridge University Press: Cambridge, MA, USA, 2011; pp. 151–161.
12. Scheutz, M. The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots. In *Robot Ethics: The Ethical and Social Implications of Robotics*; Lin, P., Abney, K., Bekey, G., Eds.; MIT Press: Cambridge, MA, USA, 2012; pp. 205–221.
13. LaGrandeur, K. Emotion, Artificial Intelligence, and Ethics. In *Beyond Artificial Intelligence: The Disappearing Human–Machine Divide*; Romportl, J., Zackova, E., Kelemen, J., Eds.; Springer: Berlin, Germany, 2015; pp. 97–109.
14. Darling, K. Extending Legal Protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior towards Robotic Objects. In *Robot Law*; Calor, R., Froomkin, A.M., Keri, I., Eds.; Edgar Elgar Publishing: Cheltenham, UK, 2016; pp. 213–231.
15. Breazeal, C.L. *Designing Sociable Robots*; MIT Press: Cambridge, MA, USA, 2002.

16. Turkle, S. *Alone Together: Why We Expect More from Technology and Less from Each Other*; Basic Books: New York, NY, USA, 2011.
17. Turkle, S. Authenticity in the Age of Digital Companions. In *Machine Ethics*; Anderson, M., Anderson, M.L., Eds.; Cambridge University Press: Cambridge, MA, USA, 2011; pp. 62–78.
18. Lin, P. Introduction to Robot Ethics. In *Robot Ethics: The Ethical and Social Implications of Robotics*; Lin, P., Abney, K., Bekey, G., Eds.; MIT Press: Cambridge, MA, USA, 2012; pp. 3–15.
19. Committee on Legal Affairs. Draft Report with Recommendations to the Commission on Civil Law Rules on Robotics. European Parliament. 2016. Available online: http://www.europarl.europa.eu/sides/getDoc.do?type=COMPARL&reference=PE-582.443&format=PDF&language=EN&secondRef=01 (accessed on 18 January 2018).
20. Wootson, C.R. Saudi Arabia, Which Denies Women Equal Rights, Makes a Robot a Citizen. *The Washington Post*. 29 October 2017. Available online: https://www.washingtonpost.com/news/innovations/wp/2017/10/29/saudi-arabia-which-denies-women-equal-rights-makes-a-robot-a-citizen/?utm_term=.e59cdc8cd981 (accessed on 26 February 2018).
21. Laukyte, M. Artificial Agents Among Us: Should We Recognize Them as Agents Proper? *Ethics Inf. Technol.* **2017**, *19*, 1–17. [CrossRef]
22. Darling, K. Extending Legal Protection to Social Robots. 2012. Available online: https://spectrum.ieee.org/automaton/robotics/artificial-intelligence/extending-legal-protection-to-social-robots (accessed on 27 November 2017).
23. Coeckelbergh, M. Moral Appearances: Emotions, Robots, and Human Morality. *Ethics Inf. Technol.* **2010**, *12*, 235–241. [CrossRef]
24. Coeckelbergh, M. *Growing Moral Relations: Critique of Moral Status Ascription*; Palgrave Macmillan: New York, NY, USA, 2012.
25. Coeckelbergh, M. Virtual Moral Agency, Virtual Moral Responsibility: On the Moral Significance of Appearance, Perception, and Performance of Artificial Agents. *AI Soc.* **2009**, *24*, 181–189. [CrossRef]
26. Levy, D. The Ethical Treatment of Artificially Conscious Robots. *Int. J. Soc. Robot.* **2009**, *1*, 209–216. [CrossRef]
27. Sparrow, R. Can Machines Be People? Reflections on the Turing Triage Test. In *Machine Ethics*; Anderson, M., Anderson, S.L., Eds.; Cambridge University Press: Cambridge, MA, USA, 2011; pp. 301–315.
28. Coeckelbergh, M. Can We Trust Robots? *Ethics Inf. Technol.* **2012**, *14*, 53–60. [CrossRef]
29. Buechner, J.; Tavani, H.T. Trust and Multi-Agent Systems: Applying the 'Diffuse, Default Model' of Trust to Experiments Involving Artificial Agents. *Ethics Inf. Technol.* **2011**, *13*, 39–51. [CrossRef]
30. Tavani, H.T. Levels of Trust in the Context of Machine Ethics. *Philos. Technol.* **2015**, *28*, 75–90. [CrossRef]
31. Taddeo, M. Defining Trust and E-Trust: Old Theories and New Problems. *Int. J. Technol. Hum. Interact.* **2009**, *5*, 23–35. [CrossRef]
32. Taddeo, M. Modeling Trust in Artificial Agents: A First Step in the Analysis of E-Trust. *Minds Mach.* **2010**, *20*, 243–257. [CrossRef]
33. Grodzinsky, F.S.; Miller, K.W.; Wolf, M.J. Developing Artificial Agents Worthy of Trust: Would You Buy a Used Car from this Artificial Agent? *Ethics Inf. Technol.* **2011**, *13*, 17–27. [CrossRef]
34. Grodzinsky, F.S.; Miller, K.W.; Wolf, M.J. Trust in Artificial Agents. In *Routledge Handbook on Trust and Philosophy*; Simon, J., Ed.; Routledge: New York, NY, USA, 2018; In press.
35. Dennett, D. *The Intentional Stance*; MIT Press: Cambridge, MA, USA, 1987.
36. Tavani, H.T. Ethical Aspects of Autonomous Systems. In *Robo- and Information-Ethics: Some Fundamentals*; Decker, M., Gutmann, M., Eds.; LIT Verlag: Berlin, Germany, 2012; pp. 89–122.
37. Floridi, L.; Sanders, J.W. Artificial Evil and the Foundation of Computer Ethics. *Ethics Inf. Technol.* **2001**, *3*, 55–66. [CrossRef]
38. Floridi, L.; Sanders, J.W. On the Morality of Artificial Agents. *Minds Mach.* **2004**, *14*, 349–379. [CrossRef]
39. Floridi, L. Foundations of Information Ethics. In *The Handbook of Information and Computer Ethics*; Himma, K.E., Tavani, H.T., Eds.; John Wiley and Sons: Hoboken, NJ, USA, 2008; pp. 3–23.
40. Tavani, H.T.; Buechner, J. Autonomy and Trust in the Context of Artificial Agents. In *Evolutionary Robotics, Organic Computing and Adaptive Ambience*; Decker, M., Gutmann, M., Knifka, J., Eds.; LIT Verlag: Berlin, Germany, 2015; pp. 39–62.
41. Floridi, L. *The Ethics of Information*; Oxford University Press: Oxford, UK, 2013.

42. Johnson, D.G. Computer Systems: Moral Entities but Not Moral Agents. *Ethics Inf. Technol.* **2006**, *8*, 195–204. [CrossRef]
43. Himma, K.E. Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties Must an Artificial Agent Have to be a Moral Agent? *Ethics Inf. Technol.* **2009**, *11*, 19–29. [CrossRef]
44. Behdadi, D.; Munthe, C. Artificial Moral Agency: Philosophical Assumptions, Methodological Challenges, and Normative Solutions. (Manuscript under Consideration for Publication). 2018. Available online: https://www.researchgate.net/publication/311196481_Artificial_Moral_Agency_Philosophical_Assumptions_Methodological_Challenges_and_Normative_Solutions (accessed on 15 February 2018).
45. Tavani, H.T. Can We Develop Artificial Agents Capable of Making Good Moral Decisions? *Minds Mach.* **2011**, *21*, 465–474. [CrossRef]
46. Moor, J.H. The Nature, Difficulty, and Importance of Machine Ethics. *IEEE Intell. Syst.* **2006**, *21*, 18–21. [CrossRef]
47. Moor, J.H. Four Kinds of Ethical Robots. *Philos. Now* **2009**, *72*, 12–14.
48. Hogan, K. Is the Machine Question the Same as the Animal Question? *Ethics Inf. Technol.* **2017**, *19*, 29–38. [CrossRef]
49. Tavani, H.T. *Ethics and Technology: Controversies, Questions, and Strategies for Ethical Computing*, 5th ed.; John Wiley and Sons: Hoboken, NJ, USA, 2016.
50. Power, T. On the Moral Agency of Computers. *Topoi* **2003**, *32*, 227–236. [CrossRef]
51. Jonas, H. *Memoirs*; Brandeis University Press: Waltham, MA, USA, 2008.
52. Kant, I. *The Metaphysics of Morals*; Cambridge University Press: Cambridge, UK, 1991.
53. Gunkel, D.J. *Thinking Otherwise*; Purdue University Press: West Lafayette, IN, USA, 2007.
54. Levinas, E. *Totality and Infinity: An Essay on Exteriority*; Lingis, A., Translator; Duquesne University Press: Pittsburgh, PA, USA, 1969.
55. Gunkel, D.J. *The Machine Question—Critical Perspectives on AI, Robots, and Ethics*; MIT Press: Cambridge, MA, USA, 2012.
56. Tavani, H.T. The Impact of the Internet on Our Moral Condition: Do We Need a New Framework of Ethics? In *The Impact of the Internet on Our Moral Lives*; Cavalier, R., Ed.; State University of New York Press: Albany, NY, USA, 2005; pp. 215–237.
57. Michelfelder, D. Our Moral Condition in Cyberspace. *Ethics Inf. Technol.* **2000**, *2*, 147–152. [CrossRef]
58. Carr, L. On What Grounds Might We Have Moral Obligations to Robots? 2018. Available online: https://www2.rivier.edu/faculty/lcarr/OUR%20MORAL%20OBLIGATION%20TO%20ROBOTS.pdf (accessed on 26 March 2018).
59. Floridi, L. Information Ethics: On the Philosophical Foundation of Computer Ethics. *Ethics Inf. Technol.* **1999**, *1*, 37–56. [CrossRef]
60. Floridi, L. On the Intrinsic Value of Information Objects in the Infosphere. *Ethics Inf. Technol.* **2002**, *4*, 287–304. [CrossRef]
61. De Laat, P.B. Trusting the (Ro)botic Other: By Assumption? *ACM SIGCAS Comput. Soc.* **2016**, *45*, 255–260. [CrossRef]