SDU
University of Southern Denmark

The Issue of Moral Consideration in Robot Ethics

Gerdes, Anne

Published in:
Computers & Society (Online)

DOI:
10.1145/2874239.2874278

Publication date:
2015

Document version:
Accepted manuscript

Citation for pulished version (APA):
Gerdes, A. (2015). The Issue of Moral Consideration in Robot Ethics. *Computers & Society (Online)*, *45*(3), 274-280. https://doi.org/10.1145/2874239.2874278

Go to publication entry in University of Southern Denmark's Research Portal

Terms of use
This work is brought to you by the University of Southern Denmark.
Unless otherwise specified it has been shared according to the terms for self-archiving.
If no other license is stated, these terms apply:

- You may download this work for personal use only.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying this open access version

If you believe that this document breaches copyright please contact us providing details and we will investigate your claim.
Please direct all enquiries to puresupport@bib.sdu.dk

Download date: 25. Mar. 2023

The Issue of Moral Consideration in Robot Ethics

Anne Gerdes
Associate Professor
Department of Communication and Design
University of Southern Denmark
4565501323
Gerdes@sdu.dk

ABSTRACT
This paper discusses whether we should grant moral consideration to robots. Contemporary approaches in support of doing so centers around a relational appearance based approach, which takes departure in the fact that we already by now enter into ethical demanding relations with (even simplistic) robots *as if* they had a mind of their own. Hence, it is assumed that moral status can be viewed as socially constructed and negotiated *within* relations. However, I argue that a relational turn risks turning the *as if* into *if* at the cost of losing sight of what matters in human-human relations. Therefore, I stick to a human centered framework and introduce a moral philosophical perspective, primarily based on Kant's *Tugendlehre* and his conception of duties as well as the Formula of Humanity, which also holds a relational perspective. This enables me to discuss preliminary arguments for moral considerations of robots.

Categories and Subject Descriptors
K4 [**Computers and Society**]: Ethics.

General Terms
Design,Theory.

Keywords
Moral consideration, ethics of robotics, duties, as if.

1. INTRODUCTION
In a recent report on lethal autonomous robot systems, Heynes points to that personhood is what links moral agency to responsibility [11]. But is that necessarily the case, or is Heynes being species chauvinistic? The answer could well be a yes, since robots have started to come into our social lives and we interact with them in human-like ways, as if they had inner mental states. On this background, it seems that we have good reasons to dwell upon our concepts of moral agency and patiency. Especially since our interactions with, and reactions towards, robots also concerns our self-image. First, I discuss the possibilities of artificial moral agency and patiency and explore whether this counts in favour of anchoring the question of moral status in phenomenological observations of how we form relations with robots; the so called *relational turn*, favoured by Coeckelbergh [3] and Gunkel [9], who summarizes the idea as an alternative to standard explanations, which sets out to decide, who (or what) deserves moral standing on the basis of ascribing properties to the entity in question. Hence, according to Gunkel, the relational *"..alternative [..] approaches moral status not as an essential property of things but as something that is socially negotiated and constructed in face of others."* ([10]:13)

I sympathize with the relational turn, but still find that it is challenged by the fact that, over time, our human-human relations may be obscured by human-robot relations. Currently, it may seem reasonable to skip discussions about what a robot *really* is and instead focus on how it appears to us and how we engage with it by applying *as if* approaches. But in the long run, our experiences with robots may radically alter our *Lebenswelt*. Here, I'm in alignment with the ideas of Turkle [18], who fears that we may lose something of great importance if we turn to robots or even end up preferring robots over humans.

For that reason, I outline a Kantian moral argument in emphasizing his treatment of duties in the doctrine of virtues, *The Tugendlehre*, which is presented in the second part of *The Metaphysics of Morals* [13]. Related to Kant's analysis of duties, there is room for a relational perspective, which can be expressed via the Formula of Humanity. Moreover, I also make reference to virtue ethical reflections in general. Thereby, I am able to put forward preliminary arguments for granting degrees of moral consideration to robots without risking that we gradually lose sight of our folk intuition and lived experience with what it is to enter into social relations. As such, I prefer to stay within a human centered framework, even though I agree with the proponents of the relational turn that there are baffling problems inherent to this kind of mind-morality perspective. However, the mere fact that things are complicated and problems unsolved does not constitute a proper reason for rejecting a framework.

2. ROBOTS IN THE MORAL SPHERE
The role of robots in moral discourse has been widely debated both within science fiction, philosophy and science. Hence, The World Robot Declaration was issued in Japan in 2004 and within the last decade, humans have increasingly interacted with care bots, pet bots, robot toys and robots for various therapeutic purposes (see for instance [18], [6], [1]).

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
ETHICOMP, September 7-9, 2015, Leicester, United Kingdom.
Copyright 2010 ACM 1-58113-000-0/00/0010 …$15.00.

One of the first to include robots in the moral sphere was Asimov, who issued his famous laws of robotics, which he used in science fiction novels to illustrate ethical dilemma situations in human robot interaction. From an engineering point of view, in *Moral Machines – Teaching Robots Right from Wrong*, Wallach and Allen [21] present the promises of machine morality from an engineering perspective by distinguishing between top-down, bottom-up and hybrid approaches to programming morality. Here, the first mentioned system suggests the implementation of formalizations of a given moral philosophical theory, whereas a bottom-up system requires neural network models, which gradually build up moral understanding by trial and error based performance optimization techniques. However, pure bottom-up systems are challenged by the lack of a guiding ethical theory, and as such there is no guarantee that a robot will develop a preferred kind of moral maturity. On the other hand, a hybrid model, which Wallach and Allen speak in favour of, combines these ideas from a virtue ethical outlook: Here, artificial moral agency might be obtained by integrating bottom-up learning scaffolded by top-down rules.

By the same token, from a philosophical angle, Verbeek [20] grasps the possibility of artificial moral agency by viewing technologies as mediating devices, which serve as morally active in shaping human understanding and action in the world. Consequently, even though technological artifacts do not hold human-like intentions, it can make sense to refer to distributed or hybrid intentionality and hence assign intentionality to technology in the sense that technological artifacts may play a directing role in our actions and experiences ([20]:57). Correspondingly, in moving beyond an anthropocentric understanding of agency, Floridi and Sanders [8] reject free will and mental states as necessary conditions for moral agency. On the contrary, they argue that moral agency may be assigned to intelligent artificial agents (AAs) to the extent that such AAs are interactive, i.e., able to react to stimuli by changing state, and capable of adaptive behavior as well as autonomous responses to the environment. What matters is whether an agent can perform good or evil actions, that is, whether its actions are morally qualifiable ([8]:371).

If we include robots in the moral sphere by assigning moral agency and responsibility to them, a next reasonable step would be to discuss if the time has come where we ought to discuss whether robots are worthy of moral consideration? Among others, Gunkel thinks the answer to that question might be a yes. In *The Machine Question – Critical Perspectives on AI; Robots, and Ethics*, Gunkel [9] argues that already by now the term "person" has been stretched out to include non-human agents, such as corporations. As such, we might benefit from including machines into the category of persons. If we do so, the question arises whether the kind of responsibilities we have towards robots would be on pair with the kind of responsibilities we have towards animals, corporations or other human beings?

A lot has been written about machine agency in trying to lay out how robots ought to treat humans. Typically interest centers on how we may protect ourselves from possible harm caused by robots. At the same time little has been said about machine patiency. ([9]:103]). Hence, according to Gunkel, a claim to moral consideration, or even rights, may arise based on our social interactions with robots. We design artificial companions with whom (or which) we do engage and bond. Our machines are no longer tools, but have instead gradually turned into social actors or social interactive objects. Consequently, it may be about time we begin to think about moral obligations towards robots, maybe even in the strong form of robot rights. The mere fact that Paro, the seal care robot, is not a consciousness being with inner mental states does not automatically justify that we should not grant moral consideration to Paro. Moreover, our ways of living with robots is not just about what we do with robots, but also concerns our self-perception – what do I become through the kind of relations I form with robots?

A contrast to the relational view can be found in the work of Sparrow [17]. He presents a so-called Turing Triage Test which allows him to illustrate that we would always chose a human life over a robot's life, regardless of how advanced the robot might be. The mere fact that we can never know what the robot is *really* feeling, and if it feels anything at all makes it implausible to talk about, for instance, 'punishing' a robot: "*Our awareness of the reality of the inner lives of other people is a function of [..] "an attitude towards a soul"".* ([17]:211). According to Sparrow, there exist an unbridgeable gap between reality and appearance ([17]:210).

On the other hand, Coeckelbergh, like Gunkel, suggests a relational turn and continues by arguing in favour of replacing "*..the question about how "real" or how "moral" non-human agents are by the question about the moral significance of appearance."(*[5]:181).

He displays problems with what he coins "a property approach to moral status assignment", which seems to rest on the assumption that we can settle issues about moral significance with reference to a set of properties (e.g., mental states, speech, consciousness, intentionality). In this manner, we can supposedly establish a firm ground for separating out entities worthy of moral standing. But, Coeckelbergh points to problems inherent in this line of argument. Especially, it appears to be impossible to establish which properties we exactly need in order to be able to assign moral status to an entity. Also, the whole endeavor is challenged by "the other minds problem" - i.e.; the fact that we can never know for sure anything about the inner lives of others. Instead, Coeckelbergh focuses on our perceptions of robots and the way this affects our interactions with such entities:

"My suggestion is that we can permit ourselves to remain agnostic about what 'really' goes on 'in' there, and focus on the 'outer', the interaction, and in particular on how this interaction is co-shaped and co-constituted by how AAs [artificial agents] appear to us, humans ([5]: 188)

Coeckelbergh's phenomenological conception reflects a relational perspective, which takes departure in the observation of our mutual dependency. This fundamental precondition – with which everyone is actually familiar – forms a central point in Coeckelberg's so-called relational ontology, which assumes that "*relations are prior to the relata"*([3]:45), and thereby view robots and humans as "relational entities". For that reason, Coeckelbergh emphasizes a social-relational approach to moral consideration ([4]:219). But, here, unlike Coeckelbergh, I shall be arguing that we need not lean against appearance in combination with a social relational ontology. Instead, I point to a Kantian outset, which emphasizes how we can have duties *to* others and *with regard to* non-humans. Before moving forward, I find it important to stress that this paper does nothing else than provide a tentative outline of my preliminary ideas. In that respect, and all though I have reservations towards their positions, I find the work of Coeckelbergh and Gunkel highly inspiring and thought provoking.

3. AS IF

Appearance is closely related to the notion of 'as if', which is also explicitly noted by Coeckelbergh in mentioning that we interact with e.g., humanoid robots or artificial companions *as if* they could be trusted, blamed or loved. Therefore, Coeckelbergh calls for a phenomenological starting point in the investigation of human-robot relations, which takes departure in the "*observed or imagined"* human-robot relations ([5]:184).

It makes good sense to turn to analogical reasoning or to introduce *as if* constructions when confronted with unfamiliar territory. This kind of idealization, or way of using representations as tools, has been given a thoroughly treatment in Vaihinger's influential book *The Philosophy of as if* [19] in which he illustrates how fictions, i.e. *as if*-models and constructions may inform science and philosophy.

Fictions are applied due to their utility, meaning that they are justifiable when proving useful in practice. But, they are not on pair with hypotheses, which can be proved or verified ([19]: xlii). Obviously, there are shades of pragmatism in Vaihinger's work on the philosophy of *as if*. But we are not dealing with the pragmatic conception, which implies that what is useful to believe is true, since here "useful to believe" may involve *both* that which is true or false. In opposition to this, the guiding principle in Vaihinger's philosophy is the observation that fictions are not just false but contradictory. Hence, fictions are errors, but fruitful errors. Yet, Vaihinger warns us that the use of fictions may also lead us astray, hence in legal practice women used to be treated *as if* they wore minor, which caused grave injustice ([19]:148).

However, fictions are widely used in everyday thinking as well as in science, philosophy, economics, legal practice and in the description of abstract objects ect.. For instance, Vaihinger mentions Adam Smith's *Wealth of Nations*, which apply the fiction that human nature is driven by rational egoism. This fiction forms the foundation of Smith's theory. Likewise, Also, Kant, in his treatment of rational agency, requires us to act *as if* we were free even though this is not the case in the real, phenomenal world. By the same token, the categorical imperative demands that you "*act as if the principle of your action were, through your will, to become a general law of nature"* ([19]:292). Hence, according to Kant, our *vernunftbegriffe* are fictions since they do not refer to objects in the world of experience [14]:KrV B799). Actually, in explaining the role of *as* if, Vaihinger points to the fact that the term "heuristic fictions" was coined by Kant:

"Kant introduces a new term for what [..] he subsequently called "heuristic fictions": he calls the ideas "regulative principles of pure reason": they are not "constitutive" principles of reason, i.e. they do not give us the possibility of objective knowledge either within or outside the domain of experience, but serve "merely as rules" for understanding by indicating the path to be pursued within the domain of experience. By providing imaginary points on which it may direct its course but which can never be reached because it is outside reality." ([19]:273)

Also, Coeckelbergh notes that we can never have access to reality, mental states or the minds of others'. But, as noted above, instead of a mind morality approach, he suggests an alternative route. Rather than discussing the moral significance of either human or robot, we must turn to the study of appearance and relations in situations involving moral considerations in human-robot interactions ([4]:215). Consequently, when people, now or in a near future, start to treat humanoid robots as if they were moral agents, we could benefit from letting these observations guide our investigations by focusing on how humans experience and form interactions with robots through *as if* approaches.

Nevertheless, according to Vaihinger, fictions are only justifiable, not probable hypotheses. As such, I doubt that we need to take a full relational turn and introduce a social relational ontology. To me, it seems that the relational *as if* approach is challenged by the fact that, over time, our human-human relations may be obscured by human-robot interactions. Currently, it might seem reasonable to skip discussions about what robots *really* are and instead focus on how they appear to us and how we engage with robots in social situations by applying *as if* approaches and ascribe human-like agency to them. But in the long run, our experiences with robots may radically alter our *Lebenswelt* and by then we will no longer be able to make use of *as if* approaches, because we have forgotten what human-like relations are, that is: we have become unable to 'measure' experiences up against the benchmark of human relations. Here, I am in alignment with the ideas of Turkle [18], who fears that we may let go of fundamental values, such as trust and friendship, if we turn to robots or even end up preferring robots over humans:

"At the robotic moment, we have to be concerned that the simplification and reduction of relationships is no longer something we complain about. It may become what we expect, even desire."([18]:295).

Likewise, if philosophers take departure in observed and imagined human-robot relations, they risk turning the *as if* into *if* ([19], [7]:9) and thereby lose sight of what originally constituted human-human relations.

4. A HUMAN CENTERED PERSPECTIVE

In *Robot Futures* [16], Nourbakhsh describes a future scenario in which some kids act with great cruelty towards a robot dog. The scenario reminiscences about children's abusive behavior towards animals, and the son in Nourbakhsh's story remarks that: *"These people…they're sick. Let's go home!"* ([16]:54). By the same token, Nourbaksh reports a more recent experience with an autonomous tour-guide robot, which people would get great fun from teasing while it was guiding guests visiting a museum. Nobody seemed to care when it said: "please step out of my way", it was not until the engineering team changed the phrase to also include the people being guided by the robot, that people's attitudes towards the robot were changed to the better - *even slow robots will be treated well by people when they are wrapped into a human social context* ([16]:58).

As discussed above, a justification of moral consideration to robots may rest upon the observation that once we start ascribing agency to robots, we may possible become ethical obliged towards them. Moreover, the way we treat robots will have an impact on our moral habitus. In order to take this into account, I choose to I introduce Kant's distinction between two kinds of duties, as duties *to* human beings and duties *with regard* to non-human beings and entities [13].

Consequently, in what follows, I shall be introducing a perspective, which of course, within a relational ontology, is viewed as flawed due to problems derived from this kind of anthropocentric line and its inherent "property approach to moral status ascription" [3]. Both Coeckelbergh and Gunkel argue that we need to move beyond the assumptions of mind morality philosophers. They in particular point to the vagueness of metaphysical concepts and the fact that there is no consensus on

what these concepts designate. Moreover, complications also arise from the fact that we do not have access to others' minds. Hence, the argument goes that we must rethink moral agency and patiency by turning to their alternative relational paradigm ([9], [3]).

But, in contrast to their approach, I think that one cannot reject the role of metaphysical concepts, such as consciousness, intentionality and freedom, with reference to the fact that complicated issues have not yet been settled. This would be like discharging logic on the basis of Gödel's incompleteness theorems.

Hence, In *Facing up the problem of consciousness* [2], Chalmers notes that consciousness is the outmost puzzling problem in the science of mind ([2]:200). He has coined the terms *the easy problem* and *the hard problem* of consciousness in referring to the fact that we already know about the part of consciousness dealing with e.g., our ability to categorize, discriminate, associate and recognize patterns. Additionally, over time, our knowledge about brain processes will gradually increase, and we will probably end up knowing all there is to know about the complexity of the brain. This is *the easy* problem. But, *the hard problem* of consciousness is the problem of experience, that is, to learn why all that processing accompanies my consciousness experience. As such, mental qualia escape reduction to biophysical matters, and in modern dualism, property dualism holds that the mind has two fundamentally different types of properties, bio-physical and qualia. According to Chalmers, despite interesting and advanced cognitive science and reductionist models "*the mystery of consciousness will not be removed.*" ([2]:221). As an alternative, Chalmers sets out to outline a nonreductive theory of consciousness, which I'll not go further into here, where I only wish to point to Chalmers' observation that : "*The hard problem is a hard problem, but there is no reason to believe that it will remain permanently unsolved"* ([2]:218).

By itself, the observation that the concepts of mind pose baffling problems is no argument for dismissing the project of mind philosophy. I argue in favour of re-instantiating the mind-morality perspective, which allows me to move on to a Kantian and virtue ethical perspective, in which there is room for arguments for moral consideration of robots as different from humans, as well as from other artifacts or tools.

Moreover, Kant's Formula of Humanity reflects a relational perspective in describing how we ought to treat others (persons) as ends in themselves, where by "ends" Kant means "*only the concept of an end that is also a duty, a concept that belongs exclusively to ethics.[..]."* ([13]: 6:389). As such, we can only have duties *to* human beings, since duties require being capable of obligation ([13]:192). Meanwhile, Kant's *Tugendlehre* [13] allows for a description of moral obligations *with regard to* other beings or entities. Actually, Kant gives similar reasons as above in emphasizing that a prevalent argument for having indirect duties *with regard to* non-human entities and animals rest upon our duties *to* ourselves:

"§17 [..] a propensity to wanton destruction of what is beautiful in inanimate nature [..] is opposed to a human being's duty to himself; for it weakens and uproots that feeling in hum, which, though not of itself moral, is still a disposition of sensibility that greatly promotes morality or at least prepares the way for it[..]. With regard to the animate but non-rational part of creation, violent and cruel treatment of animals is far more intimately opposed to a human being's duty to himself, and he has a duty to refrain from this; for it dulls this shared feelings of their suffering and so weakens and gradually uproots a natural predisposition that is very serviceable to morality in one's relations with other men. [..] – Even gratitude for the long service of a horse or dog belongs indirectly to a human being's duty with regard to these animals; considered as a direct duty, however, it is always only a duty of the human being to himself."([13]: 6:443)

Thus, a Kantian perspective, as formulated in his doctrine of virtues, enables us to introduce degrees of moral consideration along a continuum stretching from, e.g. simple artifacts, such as tools, over to, for instance, paintings and historical buildings. We have varying degrees of duties *with regard to* such entities: One could say, that I have a duty towards tools, such as for instance my garden kit, in the sense that I handle these objects with care, i.e.; I clean them after use, oil them when needed and so on. In that sense, the practice surrounding gardening includes taking proper care of one's tools, and if I fail to do so, I will either feel bad about myself and improve my behavior or continue acting carelessly. In that case, others might blame me for neglecting my duties as a gardener. Here, we are of course dealing with moral consideration in a minimal sense thereof. But, from a virtue ethical perspective [15], the way I succeed or fail in my role as a gardener is nevertheless important for my personal flourishing.

Likewise, but on a more serious scale: when confronted with acts of vandalism, for instance the destroying of historical buildings by Islamic State, we find that such acts are wrongful due to the lack of moral consideration to these architectural pearls.

We do not have duties *to* animals, but we have duties *with regard to* animals. This is so, primarily because animals deserve moral consideration because they can suffer and because the way we treat animals will influence our self-perception. Moreover, according to MacIntyre:

"*To acknowledge that there are [..] animal preconditions for human rationality requires us to think of the relationship of human beings to members of other intelligent species in terms of a scale or a spectrum rather than of a single line of division between 'them' and 'us'"* ([15]:55)

Again, the question arises: what do I, or we, as a moral community, become if we abuse animals? This indirect argument for moral consideration has been criticized by Coeckelbergh [4]:213) with reference to that it seems contra-intuitive to justify moral consideration by referring to our own well-being rather than to the well-being of the receiver of moral consideration. But, as illustrated above, actually both Coeckelbergh and Gunkel stresses the importance of a relational turn (social relational ontology) with reference to that living with robots will change our lives, hence we need to reflect upon what we become from interacting with robots. By the relational turn Coeckelbergh de-individualizes the concept of a person and holds that we have to be viewed as *relational entities whose identity depends on their relations with other entities* ([4]:215).

In addition Coeckelbergh problematizes the fact that virtue ethics faces the problem of application. Hence, we cannot establish, or delimit, what the virtues are, which ought to guide our lives, and we cannot point out precisely which entities we should grant moral consideration by exercising virtuous behavior towards them. This is a classic line of argument against virtue ethics, which has been countered by Hursthorse [12] in arguing that an ethical normative theory does not necessarily have to deliver the right answers as such, or, in the case of virtue ethics, provide a

complete catalogue of virtues. As such, a plausible normative ethical theory should not give us universal rules to guide our behavior. Instead, it should be sufficiently flexible to allow for different moral outcomes by taking into consideration relevant elements in a particular context. Consequently, when faced with dilemma situations in real life contexts, it might well be the case that two persons solve a dilemma differently. This is not a relativist standpoint, since it does not imply disagreement about the fact that there is a conflict of values, rather it takes into consideration that, in the given context, there might be more than one solution, which is in accordance with that, which is virtuous.

Thus, from a virtue ethical perspective, we develop to become what MacIntyre calls *independent practical reasoners* [15]:158) through our upbringing and through participation in moral communities, which stand as morally robust and sound practices because they are open to critical reflective examination by members from in and outside the given community.

Within this kind of human based social framework, it might still be possible to grant moral consideration to robots by introducing a continuum on a scale above artifacts - such as tools and things, which we handle - over to animals. Probably below living entities, like animals, we may place robots with which we do form *as if* social relations.

I too hold that living with robots will change our lives. But I doubt that we need to take the relational turn.

5. CONCLUDING REMARKS

Since, we already by now interact with humanoid robots, and even rather simplistic types of robots, as if they were moral agents; we ought to start deliberating about moral status. This observation might lend support to a relational turn, which allows for viewing robots and humans as relational entities, rather than subjects and objects, thereby assuming that morality is always already situated in the social sphere and phenomenologically rooted in mutual dependency between social actors – "*relations are prior to the things related*" ([3]:110). Moreover, we ought to pay attention to how human-robot interactions actually unfold, that is, focus on *appearance* or how we apply *as if* approaches when we enter into human-like relations with robots. Thus, if we follow suit with the relational turn, we might benefit from not having to struggle with the problems of property ascription and mind-morality. Even better: Coeckelbergh holds that he does not want to give up on folk intuition reflected in the idea that there is a special relation between humanity and morality ([5]:181).

Yet, in the long run, our experiences with robots may radically alter our *Lebenswelt*. Therefore, by taking the relational turn, I think we risk losing sight of something of great value to our humanity, perhaps without recognizing that this has been the case. Instead, I suggest staying within a human-centered framework. Here, I present a Kantian relational perspective, which distinguishes between others, *to* whom we have duties, and non-humans, such as robots, with *regard to* which we have duties.

Even though I place myself in (humble) opposition to the work of Coeckelbergh and Gunkel, I am deeply inspired by them. Compared to their thoroughly analyses in the field of ethics of robotics, my contribution represents nothing more than a preliminary note. For now, I have no fully fleshed out solution to offer regarding how to establish a continuum, which enables us to grant various degrees of moral consideration to non-humans. Nevertheless, when speaking about robots, I still find it worth being anthropocentric for the reasons given above, but also bearing in mind that morality is deeply linked with mortality.

6. ACKNOWLEDGMENTS

I am grateful to my dear colleague, Klaus Robering, for inspiring discussions about moral philosophy as well as for his suggestions, which helped me develop this paper.

7. REFERENCES

[1] Bartneck, C., Van der Hoek, M., Mubin, O., Al Mahmud, A. 2007. Daisy, Daisy, Give Me Your Answer Do! Switching off a Robot. *Proceedings of the 2nd ACM/IEEE International Conference on Human-Robot Interaction*. Washington DC. . DOI: 10.1145/1228716.1228746. 217-222.

[2] Chalmers, D.J. 1995. Facing up the Problem of Consciousness. *Journal of Consciousness Studies* (2): 3, 200-219.

[3] Coeckelbergh, M. 2012. *Growing moral relations: critique of moral status ascription*. Palgrave Macmillan, NY.

[4] Coeckelbergh, M. 2010. Robot rights? Towards a social-relational justification of moral consideration. *Ehtics Inf Technol*.12, 209-221.

[5] Coeckelbergh, M. 2009. Virtual moral agency, virtual moral responsibility: on the moral significance of the appearance, perception, and performace of artificial agents. *AI & Society*. 24, 181-189.

[6] Dautenhahn, K. 2007. Socially Intelligent Robots: Dimensions of Human-Robot Interaction. *Philosophical Transactions: Biological Sciences*, Vol. 362, No. 1480, (Apr. 29, 2007). 679-704.

[7] Fine, A. 1993. Fictionalism. *Midwest studies in philosophy,* XVIII.1-18.

[8] Floridi, L., Sanders, J. W. 2004. On the morality of artificial agents. *Minds and Machines*. 14(3), 349-379.

[9] Gunkel, D. J. 2012. *The Machine Question – Critical Perspectives on AI, Robots, and Ethics*. The MIT Press. MA.

[10] Gunkel, D. J. 2014. The Other Question: The Isssue of Robot Rights. *Proceedings of Robo-Philosophy 2014. Sociable Robots and the Future of Social Relations.* Frontiers in Artificial Intelligence and Applications. IOS Press

[11] Heynes, C. 2013. Report of the Special Rapporteur on extrajudicial summary or arbitrary executions on Lethal Autonomous Robot Systems. A/HCR/23/47 fhttp://www.ohchr.org/Documents/HRBodies/HRCouncil/RegularSession/Session23/A-HRC-23-47_en.pdf.

[12] Hursthouse, R. 1999. *On Virtue Ethics*. Oxford University Press. Oxford. NY

[13] Kant, I. 1991. *The Metaphysics of Morals*, transl. by M. J. Gregor. Cambridge University Press.

[14] Kant, I. 1785. Akademieausgabe, vol. IV *Grundlegung zur Metaphysik der Sitten*. http://www.korpora.org/Kant/aa04/Inhalt4.html

[15] MacIntyre, A. 1999. *Dependent rational animals: Why human beings need the virtues.* Carus Publ. Company. Chicago.

[16] Nourbakhsh, I. R. 2013. *Robot Futures*. MIT. Cambridge. MA.

[17] Sparrow, R. 2004. The Turing Triage Test. *Ethics and Information Technology*. 6, 203-213. DOI: 10.1007/s10676-004-6491-2.

[18] Turkle, S. 2011. *Alone Together – Why We Expect More From Technology and Less From Each Other*. Basic Books, NY.

[19] Vaihinger. H. 1924. *The Philosophy of as if*. Transl. by C. K. Ogden. London.

[20] Verbeek, P. P. 2011. *Moralizing Technology: Understanding and Designing the Morality of Things*. The University of Chicago Press.

[21] Wallach, W., Allen, C. 2009. *Moral Machines – Teaching Robots Right from Wrong*. New York: Oxford University Press.

.