Manuscript version

M. Smakman, "Robots and moral obligations," in *Frontiers in Artificial Intelligence and Applications*, 2016, vol. 290, no. What Social Robots Can and Should Do, pp. 184–189.

The final publication is available at IOS Press through
http://dx.doi.org/10.3233/978-1-61499-708-5-184

Robots and moral obligations

Matthijs Smakman [a,]
[a] *Lecturer at Institute of ICT, HU University of Applied Sciences Utrecht*

Abstract. Using Roger Crisp's [1] arguments for well-being as the ultimate source of moral reasoning, this paper argues that there are no ultimate, non-derivative reasons to program robots with moral concepts such as moral obligation, morally wrong or morally right. Although these moral concepts should not be used to program robots, they are not to be abandoned by humans since there are still reasons to keep using them, namely: as an assessment of the agent, to take a stand or to motivate and reinforce behaviour. Because robots are completely rational agents they don't need these additional motivations, they can suffice with a concept of what promotes well-being. How a robot knows which action promotes well-being to the greatest degree is still up for debate, but a combination of top-down and bottom-up approaches seem to be the best way.

Keywords. Robots, Machine ethics, Moral obligation, Moral concepts, Moral judgment, Ethics, Decision making, Well-being

1. Introduction

In this paper, I will assess arguments for and against the use of moral concepts and see if these arguments can help understand how moral concepts should be applied to the field of robotics. It might seem strange to talk about moral concepts and robotics, but robots are no longer the "stupid" machines they once were. Contemporary robots don't have to be under the constant control of humans. They can even learn on their own. Because of this, outcomes can arise that the engineer could not have thought about when he or she designed the machine.

The first part of this paper will give a description of the current state of the field of robotics: what robots are, how they learn, and consequently, what kind of ethical questions arise. I will give a sufficient background to explain why robots will find themselves in moral situations. Secondly, I will examine why humans need moral concepts using Crisp [1] and McElwee [2]. Crisp [1] argues that concepts, such as moral obligation, are not the ultimate source of moral reasoning, but well-being is. McElwee [2] reacts to this by showing that the function of moral concepts goes beyond giving agents reason to perform an act. Thirdly, this paper will give an insight in how current approaches try to integrate morality in robots and how these could be combined with the results of this paper.

2. Introduction to robotics and Machine ethics

A robot is an "engineered machine that senses, thinks and acts" [3]. A machine uses sensors to obtain data about the external world, for example by cameras, GPS-receivers or other ways to acquire data. This data needs to be processed if the robot is to react. This process is called thinking. Although it can be argued that this classification of software process as 'thinking' is false, the classification will be adequate for the question of this paper. The software process that I call thinking uses rules to process the data from its sensors and to make decisions on how to react. At the basis of these rules is a human, but on the basis of the human programmed code, the robot can teach itself how to react to new unknown situations.

2.1. Machine learning

Machine learning software enables machines to learn from past experiences. The definition of machine learning is: "A computer program is said to learn from experience 'E' with respect to some class of task 'T' and performance measure 'P', if its performance at tasks in 'T', as measured by 'P', improves with experience 'E'" [4]. In other words, if a robot can improve how it performs a certain task based on past experience, then it has learned. As robots that are programmed to use machine learning improve their actions, situations can arise that the engineer or programmer of that robot could possibly not have been aware of.

2.2. Machine ethics

Next to the abilities of a robot to sense and think it must be able to react. This reaction follows from its ability to sense and to think. The reaction of the robot happens in the real world, the world humans are living in. Since robots can act autonomously without direct control from a human and can make decisions based on sense data, a number of ethical questions arise. For example, how do we ensure that robots don't harm humans? Lin et. al. [2] describe an interesting list of ethical questions like: "whose ethics and law ought to be the standard in robotics", "if we could program a code of ethics to regulate robotic behaviour, which ethical theory should we use" "should robots merely be considered tools, such as guns and computers, and regulated accordingly" and "will robotic companionship (that could replace human or animal companionship) for other purposes, such as drinking buddies, pets, other forms of entertainment, or sex, be morally problematic?".

If robots are not programmed with the capacity to make moral decisions, disastrous situations can arise. Because of this Wallach & Allen [5] argue that robots should be equipped with the ability for ethical reasoning and ethical decision-making. There have been many, such as Anderson [6], Coeckelbergh [7], Crnkovic and Curuklu [8], Malle [9], Murphy and Woods [10], and Wallach [11] that have taken on the challenge of implementing moral decision making into robots.

One possible strategy for the implementation of morality and for answering some of the ethical questions formulated by Lin et. al. [3] is to program robots in such a way that they are guided by (the software equivalents of) moral concepts, such as moral obligation. The next section will assess whether we should or should not program robots to be guided by moral concepts by looking at arguments given by Roger Crisp [1].

3. Talking with(out) moral concepts

According to Crisp [1], we should talk about morality without moral concepts because morality provides only non-ultimate reasons. Well-being is the ultimate source of all moral reasons [1]. And because morality does not provide ultimate reasons to act, we should not start with the conception that people have any moral obligation when analysing morality. In this section, I will show the arguments Crisp [1] gives for demoralising ethics and how this can help understand why robots should or should not be programmed with moral concepts. Crisp [1] claims that: "The kind of reasons philosophical ethics should be most concerned with are ultimate or non-derivative in nature", actions which will advance well-being.

According to Crisp [1], nearly any human society has: "a set of cognitive and conative states, including beliefs, desires, and feelings, which leads its possessors among other things to (*a*) view certain actions as wrong (that is, morally forbidden) and hence to be avoided, (*b*) feel guilt and/or shame as a result of performing such actions, and (*c*) blame others who perform such actions." He calls this positive morality to correlate with the term positive law. He uses these terms to strengthen his claim that there is a strong analogy between morality and law.

3.1. Positive law and positive morality

The first analogy is that both law and morality are: "mechanisms for guiding human action towards similar kinds of goals" [1]. The focus here is firstly on criminal law, not civil law. He assumes that the two concepts have only one function, to tell what we have most reason to do, which will be argued against later in this paper. The second analogy is that both law and morality: "involve the forbidding of certain actions, and the infliction of sanctions on those who perform these actions" [1]. The third analogy is that both law and morality have a similar origin. Our ancestors developed morality so that the group would function better, and because of that survive. At first these would be just basic emotions such as anger and fear, but after humans developed language, more complex emotions became possible. With the analogies of positive law and positive morality Crisp [1] shows that positive morality does not provide non-ultimate or derivative reasons. So the moral reason to be kind does not come from the moral obligation to be kind, but because it increases well-being. But for morality to function effectively, people should

take it to be ultimately reason giving [12]. According to Crisp [1], this is clear because emotions of guilt and shame involve the thought of wrongness.

3.2. The function of moral concepts

The goal of normative ethics, says Crisp [1], is to tell what we have most reason to do. Ethical concepts such as moral obligation, moral wrongness, kindness, virtue, fairness, etc. are not the ultimate source of our reasons. They only provides non-ultimate of derivative reasons. The ultimate source for all moral reasoning is to promote well-being, according to Crisp [1]. Because of this: "there is no immediate need for us to consider those many philosophical views" [1]. Another reason to avoid moral concepts is that they come with strong normative emotions like shame, guilt and blame. These emotions can cloud our judgement and for that reason, they need to be avoided. To answer the question "what makes people's lives better" there is no need for concepts as moral wrongness, obligation, cruel, good, bad etc. All that we need to know, is what acts are going to best promote well-being.

So integrating moral concepts in robots is not only difficult, it is best to be avoided. Although robots can't feel emotions like humans do, at least not yet and not for the foreseeable future according to Lin et. al. [3], we should still try to avoid moral concepts as a cowardly robot or a loving robot. The robot should be aimed at increasing well-being and this is what the focus of the debate should be. Moral concepts, as shown, do not independently provide reasons for actions, and this is why they should be eliminated or at least be avoided.

3.3. Other functions of morality

Brian McElwee [2] responds to Crisp's argument that moral concepts and ethics need to be reason-giving. If the goal of morality is to maximise well-being, then it seems strange to think that moral obligation would give additional reasons to act morally, reasons that are not exhausted by the reason to promote well-being. According to McElwee [2], the concept of moral obligation involves the idea of reason implying. So moral concepts are not reason providing, but reason implying. Therefore, moral wrongness does not give reason not to act in a certain way, but it implies that there are reasons not to do it. They don't come from the moral wrongness of it, they come from the facts that make it wrong. And the facts that make an act wrong are also the facts that give you reasons not to perform the act. The wrongness in itself does not provide any reasons, according to McElwee [2]. This view does not conflict with Crisp's view that moral concepts do not provide reasons for action. So, since ethics should point people to reasons to act, and moral obligations do not provide reasons to act, they still should play no role, and should be abandoned.

Crisp [1] assumes a situation where all the reasons to act in a certain way are known. In this situation, there would be no role for moral obligation. Moral obligation would then not be telling us anything extra, and because of this, it is not useful. But even with Crisp's argument, there are still reasons not to eliminate moral obligation. Agreeing that the function of ethics is to guide behaviour does not mean this is the only function. According to McElwee [2], there are other functions. The function of moral obligation is not only to give reason for actions, but its function can also be found in other areas.

The function of moral obligation, according to McElwee [2], is to tell us how to react to agents that behave in particular anti-social ways. He argues that: “even if morality does not itself provide reasons for action, the moral categories nevertheless have a central role to play in ethical theory: they allow us to make crucial judgements about how to feel about and react to, agents who behave in antisocial ways, and they help motivate us to act altruistically” [2]. So knowing something is morally required can help motivate the performance of actions. This means the concept of moral obligation helps an agent to make a choice of which actions he or she should perform in a certain situation. In McElwee’s view, moral obligations help agents to react to people in certain ways. There are three functions of moral obligation according to McElwee [2]. The first is that moral obligation offers an assessment of the agent. To say something is morally wrong is to say it is an unacceptable form of behaviour. The second is that we need the concept of moral obligation to take a stand. We need the concept to make a stand to which actions can be tolerated and which actions can not be tolerated. The final and third function is to play a motivational and reinforcing role. To believe something is morally required can reinforce our motivation to perform an action. These three functions can’t be reduced to their reason giving function.

3.4. Total rational agents

McElwee [2] has shown that the function of moral obligation can be wider than just a reason giving function. One response of Crisp [1] is to have a thought experiment. In a society with only complete rational agents, that know what actions will promote the most well-being, there is no need for morality to guide actions. This group of rational agents would not be missing anything if they had no concepts of moral obligation. The absence of for example moral wrongness, according to Crisp [1], doesn't seem to be problematic in this society.

This thought experiment shows that in a certain, ideal society there would be no need for moral obligation or other derivative moral concepts. In this society there is no need for the function McElwee [2] presents; an assessment of the agent, to take a stand or to motivate and reinforce behaviour. Inhabitants of such a society would use well-being to determine what they have most reason to do. But in reality, there is no such ideal society. Humans are not completely rational beings and are often guided by evolutionary and sociological processes in their moral decision making [13]. Robots do not suffer from these sometimes irrational processes and thereby can be programmed to be fully rational, like in the thought experiment stated above. This means that robots don’t need moral concepts as moral obligation and therefore there is no reason to program a robot to have a certain moral concept, like moral obligation, good or bad.

If there is no ultimate or derivative reason to program robots with moral concepts such as fairness or kindness, the concepts that guide us to the right moral attitude, there must be another way to integrate morality into these machines. This is important since we don’t have complete control over the learning process of a robot and its relationship to its environment. A robot can maybe learn using machine learning but its core purpose will be to maximise a score x. That x should be to well-being since this is the only non-derivative reason to act and therefore the only “ethical” concept a robot needs.

4. How to program well-being

Crisp [1] defends hedonism, but not a psychological hedonism nor a hedonistic utilitarianism. He defends a hedonism as a theory of well-being: "of what is ultimately good for any individual" [1]. The hedonist, as he wants to understand her, will say that: "what makes accomplishment, enjoyable experiences, or whatever good for people is *their being enjoyable*, and that this is the only '*good-for*-making' property there is" [1]. How are robots supposed to figure out what would realise the most well-being? At first, it does not seem to be possible for a robot to feel enjoyable experiences. But robots don't have to "feel" enjoyment, or pleasure or pain, they only need the concept of what it enjoys, and what is enjoyable can be programmed.

How to program this might be problematic and there are different strategies at hand. The strategies fall into two main categories: top-down and bottom-up approaches [11]. Top-down approaches try to integrate consequentialist and deontological theories into robots and bottom-up approaches try to integrate the mechanisms out of which a capacity for moral judgement can originate. According to Wallach [11] neither of the approaches alone "is likely to be adequate for building autonomous (ro)bots with full capacity to make appropriate choices". It seems that neither bottom-up, nor top-down approaches will be sufficient to ensure to promotion of well-being in fully autonomous robots. But Wallach [11] gives an insight of how a not fully moral robot may learn by the support and direction of a community: while the virtues are acquired from the bottom-up through experience and habit, the virtues are supported and may be evaluated from the top-down. This approach seems to me to be the right one to program robots with what they have most reason to do, except without the use of the moral concepts for reasons stated above. So a combination of top-down and bottom-up approaches might be the best way to integrate morality into robots, but well-being should be the central concept, not moral obligation, good, bad, virtues or other moral concepts.

5. Conclusion

This paper set out to answer the question if and how moral concepts should be applied to the field of robotics. To answer this question this paper has argued that robots are engineered machines that can sense, think and act autonomously, without the direct control of humans. Well-being is the ultimate source of moral reason, not moral concepts. For humans there are additional reasons to retain moral concepts, namely: as an assessment of the agent, to take a stand or to motivate and reinforce behaviour. Because robots are completely rational agents they don't need these additional motivations and therefore there is no need for concepts as moral obligation to guide the actions of robots or tell robots what they have most reason to do. They should only be programmed to promote well-being, the ultimate source of moral concepts. How a robot knows which action will promote to most well-being is still up for debate but a combination of top-down and bottom-up approaches seem to be the best way.

References

[1] Crisp, R. (2006). R*easons and The Good.* Oxford: Clarendon Press

[2] McElwee, Brian (2010). Should We De-moralize Ethical Theory? *Ratio* 23 (3):308-321.
[3] Lin, P., Abney, K., Bekey, G., (2011) Robotic ethics: Mapping the issues for a mechanized world. *Artificial Intelligence* 175: 942–949
[4] Mitchell, T. (1997). *Machine Learning*, McGraw Hill. ISBN 0-07-042807-7, p.2.
[5] Wallach, W. & Allen, C. (2009) *Moral Machines, Teaching Robots Right from Wrong*. Oxford University press
[6] Anderson, S. L. (2008). Asimov's "three laws of robotics" and machine metaethics. *Ai & Society*, *22*(4), 477-493.
[7] Coeckelbergh, M. (2010). Moral appearances: emotions, robots, and human morality. *Ethics and Information Technology*, *12*(3), 235-241.
[8] Crnkovic, G. D., & Çürüklü, B. (2012). Robots: ethical by design. *Ethics and Information Technology*, *14*(1), 61-71.
[9] Malle, B. F. (2015). Integrating robot ethics and machine morality: the study and design of moral competence in robots. *Ethics and Information Technology*, 1-14.
[10] Murphy, R. & Woods, D. D. (2009). Beyond Asimov: the three laws of responsible robotics. *IEEE intelligent systems*, *24*(4), 14-20.
[11] Wallach, W. (2010). Robot minds and human ethics: the need for a comprehensive model of moral decision making. *Ethics and Information Technology* 12:243–250
[12] Joyce, R. (2001). *The Myth of Morality*, Cambridge University Press Ch. 1
[13] Haidt, J. & Kesebir, S. (2010). Morality. In S. Fiske, D. Gilbert, & G. Lindzey (Eds.) *Handbook of social Psychology 5th Edition.* Hobeken, NJ: Wiley. Pp. 797-832