Journal of Philosophy of Life Vol.13, No.1 (January 2023):112-124

Clockwork Courage
A Defense of Virtuous Robots

Shimpei Okamoto*

Introduction

With the development of artificial intelligence (AI) and robotics, robotic devices are expected to be available in various social situations. These devices may be able to perform various tasks previously done by human beings. Some of these tasks, of course, have a significant impact on human well-being. If this is the case, it is necessary to implement a mechanism in robots' internal systems to regulate the behaviour that can be activated in such situations. Such a mechanism must include content worthy of the name 'ethical constraints' or 'moral rules'.

Engineers (and some philosophers) of AI and robotics have been investigating methods of implementing 'morality' in robots' computational systems since the late 1990s. These studies are called 'Artificial Morality' (Danielson 1992) or 'Machine Ethics' (Anderson and Anderson 2007). Their goal is to develop agents that can act morally without human decision-making, that is, artificial moral agents (AMAs). The project of building AMAs was initially just a thought-experiment, but now it must be considered as a practical issue of actual engineering, for the use of robotic devices has become a real problem.

In moral philosophy since ancient times, various theories have been proposed to understand human moral standards or moral thinking; for example, utilitarianism, Kantianism, contractualism, and so on. These theories are rich resources for understanding morality. Therefore, it is necessary for engineers and moral philosophers to cooperate in order to successfully build moral machines. However, this is not an easy task. If the project of building AMAs is conducted with the approach of selecting the most desirable of the various moral theories, and rewriting it into computable formulations, the standard that determines the best theory is necessary. Although, where does this standard exist? Even moral philosophers hardly agree on which theory is best. Rather, the conclusions arrived

* Assistant Professor, School of Letters, Hiroshima University. 1-2-3 Kagamiyama, Higashi-Hiroshima City, 739-8522 Email: sokmt[a]hiroshima-u.ac.jp

at by philosophers continue to diversify these theories. If engineers expect to be advised by philosophers' opinions to implement morality in robots, they will not easily be able to determine which theory to use.

In previous studies, two of the various moral theories appear promising. One is the virtue ethics advocated by Aristotle (Wallach and Allen 2009). The other is contractualism, advocated by Rawls (Leben 2018). These two theories seem to have greater implementability to computational systems than other theories. However, many philosophers are sceptical of these attempts from the philosophical point of view. They argue that 'can' does not imply 'ought', so even if we can implement morality in a machine, this does not mean that we should do so.

This paper examines the theoretical problems associated with the implementation of moral theory in machines, particularly in favour of virtue ethics. Nevertheless, the issues addressed here will be important not only for AMAs imbued with virtue ethics but also for those created with other theoretical backgrounds, such as consequentialism, in mind. This is because virtues are equally important to Aristotelian, Rawlsian, or any number of other theorists. Section 1 outlines research that attempts to implement morality into AI systems and explains why virtue ethics seems to be important. Section 2 introduces two objections to the attempt to implement virtues into robots. In Section 3, it is considered that virtues for robots are different from those for humans. Section 4 responds to this objection.

1. Artificial Moral Agents and Moral Theories

Even if robots can behave freely in social situations, 'free behaviour' does not entail complete autonomy. Their behaviour must be limited to some extent, because if people are harmed by the robot's behaviour, the use of the robot itself will be blamed, regardless of who is responsible for it. In order to avoid such a situation, robots need to have internal constraints on their actions; however, it is difficult to design computational ethical constraints, not to mention Isaac Asimov's 'Three Laws of Robotics', which illustrated that even these three simple rules can cause serious ethical challenges.

We must implement morality in robots—in other words, we must teach robots right from wrong. This, of course, does not mean that robots need to have the ability to think, feel, and judge just as humans do. Robots will be designed for

specific purposes, such as medical practices or military operations. Even if autonomous robots do have 'autonomy', it is not complete autonomy, but autonomy only in a very limited sense. However, within this 'limited sense', the robots must be able to make decisions and act spontaneously, because the greatest benefit of using robotic devices is that decisions and actions can be made by machines alone, without human supervision. If this is the case, the internal systems of robots must include substantial moral considerations for those who may be affected by their actions, even though their process is completely different from human ethical thinking.

In moral philosophy, various principles and rules that justify our moral judgments and guide our actions have been studied. In order to implement ethical constraints in computational systems, many studies refer to various moral theories that have been considered in moral philosophy. If these theories are formulated computably, they may serve as a blueprint for the design of the internal constraints of robots. However, there are two difficulties to be aware of. First, even though moral theories play a guiding role, they are not intended to regulate the process of thought when we are executing a moral behaviour. Most moral theories are used as a basis for critically examining one's past actions or the actions of others, or when envisioning future actions. In other words, the grounds provided by moral theories are used for the justification of actions, not for performing actions.

If a theory is to be worthy of being called 'moral', of course, it must not stray too far from our intuitive reactions. However, even a hard-core utilitarian, for example, does not calculate utilitarian consequences in their everyday actions. Rather, those who always refer to moral theory in their every decision may even be considered a morally flawed person. For example, Michael Stocker spoke of such people in terms of a kind of 'Schizophrenia of Modern Ethical Theories'. Agents who ignore their intuitions or emotional reactions, and follow only their committed principles, are closer to psychiatric patients than moral saints (Stocker 1976). Perhaps this objection also applies to robots. Indeed, we sometimes accuse humans who place more emphasis on principles and rules than personal commitments of 'acting like a robot'. For the same reasons, we may not want social robots to act 'like robots'.

Secondly, traditional moral theory is required as a reason to justify 'human actions', not 'robot actions'. Furthermore, the ethical hurdles for robots will be much higher than for humans (Allen, Varner, and Zinzer 2000). Colin Allen and his colleagues devised the 'Moral Turing Test' as a complete AMA standard.

However, unlike the original Turing test, it was found that the criterion for passing, which was if the robot and the human being could not be differentiated, is not adequate to determine the morality of the robot. This is because we often evaluate that an action is acceptable if it is done by humans, but if it is done by a robot, it will be considered unacceptable. If the standards of human morality and robot morality are different in the first place, undoubtedly, applying a theory made for humans to a robot without modification will at best provide an inappropriate standard.

Among the various moral theories, one of the theories that seems to avoid these difficulties is virtue ethics, a theory proposed by the ancient Greek philosopher Aristotle. This theory treats the ultimate standard as the agent's 'character traits' represented by their actions, rather than the results or the motives of their actions. In Aristotelian virtue ethics, for example, the reasons not to tell a lie is neither bad results (e.g., pain) nor bad motives (e.g., treating others as a means), but vicious character traits, in this case, the dishonesty represented by telling a lie. This theory avoids the difficulties described above. First, desirable character traits in virtue ethics imply respect for personal commitment and appropriate emotional responses. Second, the ideal moral agent in virtue ethics requires higher standards than ordinary people. In this theory, ordinary people can be moral agents, but they are actually imperfect moral agents, and are training to acquire virtues. If so, AMAs, which embody the ideals of virtue ethics, are more moral than human agents. In this case, even if the ethical hurdle of the robot is higher than that of a human, this will not present a problem.

Computability is also an advantage of virtue ethics. For example, Wallach and Allen (2009) distinguished the ways to implement moral theory in AMAs in both top-down and bottom-up approaches. On the one hand, a top-down approach is a way to rewrite an existing moral theory into a computable formulation. However, to calculate the consequences, for example, a large amount of simulation is required; therefore, it is not feasible in real-time moral thinking, where a conclusion must be drawn in an instant. Furthermore, this approach is difficult to deal with when AMAs face a case that the designers could not have predicted in advance. On the other hand, the bottom-up approach is the method of generating ethical rules by AI itself using tactics such as machine-learning or multi-agent simulation. However, by bypassing ethical hurdles through learning, AI may behave inappropriately. Moreover, if the data on which learning is based include prejudices, the tendency that AI learns may become an ethically bad habit.

Virtue ethics can be understood as a hybrid approach that includes both top-down elements that regulate actions through virtue, and bottom-up elements that learn virtuous actions through habits. In order to teach robots what kind of actions are virtuous, it is necessary to incorporate abstract virtues as ideals from the top-down approach. However, in order to have the ability to actually perform moral actions, they must learn virtuous one through daily training from the bottom-up approach. This ethical framework, as Aristotle considered, is appropriate not only for human moral education, but also for generating moral standards for robots.

Furthermore, virtue ethics has the advantage of having a high affinity with connectionist theory in the philosophy of mind. In neural-network theory, which is the basis of machine-learning, intelligence is understood as a combination of certain modules with various roles. According to some philosophers, the human thought-process assumed in virtue ethics is close to the cooperation of such modules, and human moral psychology is completely different from the top-down model that, for example, Kant had envisioned (Wallach and Allen 2009). If the intelligence assumed by virtue ethics is similar to AI as well as human intelligence, virtue ethics would be the best moral theory for AMAs in this sense.

2. Some Objections to Virtuous AMAs

As seen in the previous section, virtue ethics has many advantages as a model of morality implemented in AMAs. However, there are some objections to these attempts. The first objection occurs on the empirical level. For example, at present, it is difficult to even design an AI with sufficient ability to speak with humans. For this reason, it is even more difficult to build robots that embody moral ideals by making better moral judgments than humans. This is a distant dream, with no clear path to achieving it at present. Furthermore, even if robots could have virtue, there would be no way to determine definitively whether they really do have virtue.

Unfortunately, we will have to make a futuristic response to such objections. It may not be possible to build ‘perfect’ AMAs with current technology, but there may come a day when it is possible to do so in the future. However, it can be said that the virtue ethics model would be more appropriate for current artificial intelligence than certain other theories. For the question of whether virtue is possessed or not, it may be sufficient for human evaluators to accept that a given AI appears to have virtue. In any case, it will be desirable to respond to these empirical objections at the empirical level. For the purposes of this paper, it is

only to say that virtue ethics has a certain advantage, at least when considered as a model for ethical AI.

There are more important problems beyond the empirical level. In principle, and not only with reference to current robots, is it possible for robots in general to possess virtues? If so, should a virtuous robot be built? There are two types of questions here. One is the metaphysical question of whether a robot is an entity that can acquire virtue. The negative response to this question is that virtue is for humans and, by definition, no robot can have virtue. The other is a normative question of whether the attempt to implement virtue in AI is a desirable design from an ethical point of view. If the attempt to build virtuous robots is itself an ethically wrong project, we should not aim for it.

Consider the first question. In Aristotelian virtue ethics, virtues have a teleological structure. Every being has a purpose for existing. Achieving that goal well means excellent being. The purpose of a tool such as a hammer is, for example, to strike the nail well. If it has the ability, it is an excellent hammer, because the purpose of being a hammer is to hit a nail. Therefore, what is the purpose of being a human? In Aristotle's opinion, the purpose of human beings is happiness.

According to contemporary Aristotelian philosopher Rosalind Hursthouse, 'a virtue is a character trait a human being needs for *eudaimonia*, to flourish or live well' (1999: 167); that is, we have to acquire virtue because it is necessary for our happiness. Hursthouse understands happiness in the Aristotelian sense, as a flourishing of human ability. This is sometimes called *Eudaimonism* or *Perfectionism* with regard to well-being. In short, virtues in the Aristotelian framework are inseparable from perfection 'as human being'. We cannot live a happy life without having virtues; therefore, we need to acquire virtue to live a happy life. Individual moral acts are like by-products derived from this purpose.

Assume that virtues for humans are like those that Aristotle considered. Is a robot's purpose to live happily? Probably not. Most robots are designed, produced, and deployed as a means to achieve some purpose. If virtue means the perfection for achieving the purpose of being, then virtues for robots will be completely unrelated to morality. In other words, no matter how excellent a hammer is, it is not an ethically excellent hammer. Similarly, no matter how excellent robots are, it would be wrong to evaluate them as ethically excellent robots. On the other hand, because robots look as if they have morally excellent character traits and actually do not, they cannot be evaluated as virtuous. In the worst case, robots that

appears to have virtue may be called deceptive.

Let us proceed to the second objection. If we can call the virtues of robots ethical, this means that robots can aim for their own happiness—for virtue means the perfection of the purpose of being. If so, the attempt to design such a robot for a specific purpose would be a manifestation of vices. This is because, if robots have autonomy that is worthy of virtue, it would be wrong to ignore their autonomy and bind them to a specific purpose, just as it is wrong for parents to pre-determine how their child should live.

For example, Ryan Tonkens explains this by giving an example of a 'robotic clown'. A virtuous robotic clown may perform a variety of acrobatics with autonomy and, in some cases, even invent new arts. It will behave kindly to the audience and be generous to human colleagues. Perhaps the people around the robot may treat it as morally considerable, though, of course, not as morally considerable as humans. However, here is the problem: even if the robot clown has a high degree of spontaneity and has virtuous character traits, we cannot admit that it has the freedom to resign from being the clown. For, if this is allowed, there is no point in building such a robot. We will develop and use robots for specific purposes. Consider the military case. A virtuous robot soldier will fight 'with courage' in various military operations. The robot may have autonomy to shoot enemies and protect non-combatants, but we will not grant the robot soldier the freedom to retire from the army. This is because the robot was made for the purpose of being a soldier.

If this is the case, using AMAs is an activity very close to slavery. Although robots could have rights, they are not allowed to exercise them. It is hard to say that creating such slave agents is something that virtuous engineers and business owners should do. However, it is correct to argue that if AMAs can only become slaves, irrespective of how well they work, the autonomy or free-will that be the title of moral rights must not be implemented in their systems as they are tools and should be tools. Joanna Bryson (2010), for example, makes such a claim. If Bryson is correct, however, we should not make robots virtuous in the first place.

These two questions are parallel to the ones on moral patiency that David J. Gunkel proposed in *Robot Rights*. According to Gunkel, the following two questions about robot patiency (or moral rights) are often confused: (S1) 'Can robots have rights?' and (S2) 'Should robots have rights?' (2018: 5-6). Many argue that S1 and S2 are equivalent, that is, either if robots can have rights, then robots should have rights, or if robots cannot have rights, then robots should not

have rights. However, Gunkel says that S1 and S2 are different types of questions and can be answered separately. As with Gunkel's questions about rights, those about virtues need to be distinguished ontologically from normative questions. Further, normative conclusions are not always derived directly from ontological positions. In this paper, I propose that robots can acquire virtues even at the ontological level and should also acquire virtues at the normative level.

3. Virtues and Virtual Virtues

If the robot can have virtues, it should not be built. If the robots should be built, it should not have virtues. Both of these claims seem to be correct, but the current and following sections will attempt to refute both of these propositions. In other words, the following will argue that robots can have virtues, and should also have virtues. However, this does not mean that virtues for robots are the same as those for humans, but they are still worthy to be called virtues. As some researchers have suggested, virtues for AMAs are 'virtual virtues' (e.g., Wallach and Allen 2009; Coeckelbergh 2012; DeBeats 2014). Sceptics who argue against virtuous robots are wrong in thinking that virtues for AMAs are the same as those for humans.

Indeed, Wallach and Allen (2009) set the following criteria in order to find a moral theory suitable for AMAs.

> Given the range of perspectives regarding the morality of specific values, behaviors, and lifestyles, perhaps there is no single answer to the question of whose morality of what morality should be implemented in AI. Just as people have different moral standards, there is no reason why all computational systems must conform to the same code of behavior. (78-79)

They chose the best theory from the engineering point of view (e.g., 'Which theory is the easiest one to implement in the system?'), rather than from the philosophical point of view that cannot expect consensus (e.g., 'Which theory is the most ethically desirable?'). In their argument, the desirability of virtue ethics is ensured by having two sides, bottom-up and top-down, and its affinity with connectionism. For these and additional reasons, virtue ethics can be said to be ethically appropriate for AMAs.

Human virtues and virtual virtues are similar in some ways, and different in

several ways. First, we will consider the similarities. According to Aristotelian virtue ethics, various virtuous actions are more than individual character formation, such as courage and honesty, for example. They are more comprehensive and necessary traits for unconditional 'excellent agents'. If someone has virtue, she can perform virtuous acts in almost all situations. It was Aristotle who originally claimed this feature, but has been called 'the unity of virtues' by philosophers after him. Virtue for AMAs and humans alike is arguably the trait necessary for 'excellent robots'. This is because, from the viewpoint of engineering ease, it is unreasonable to design and implement individual virtues corresponding to each very complex social aspect. Just as human virtues are a kind of rationality and are necessary for 'excellent judgments' in general, virtual virtue is also a kind of rationality that regulates various actions in general. Therefore, virtual virtues also have their 'unity'.

The second similarity is their teleological structures. Virtues for humans have the purpose of the possessor's happiness. For Aristotle, happiness does not mean mere maximization of pleasure or satisfaction of preference, but perfection as a human being. Aristotle argued that happiness in this sense is an overly complicated and vague purpose, so that human beings cannot acquire happiness unless they have genuine virtue. Virtual virtues have a teleological structure as well. Just as virtues for humans are character traits useful for perfection as human beings, virtual virtues are character-traits useful for AMAs' perfection.

However, the similarity ends here. Virtues for humans are aimed at the happiness of their possessor, but virtual virtues are not. Rather, their purpose is not the AMAs themselves, but the well-being of those who are affected by AMAs' actions. Why do we need to build AMAs? Because robots are likely to be used in social situations where their behaviour significantly affects the well-being of humans. A situation wherein robots make such decisions without ethical constraints is undesirable; therefore, AMAs should be built. It is not because we want to increase new moral patients, nor is it because we want to increase new 'persons'. What we want is for the system to have ethical constraints. For this reason, AMAs need to establish 'safe interaction with humans', but it is not necessary for the robot to have a happy life. Recall Asimov's third law. Robots certainly must protect themselves or their happiness. However, this is only conditional.

4. Reply to Objections

Given the above, this section will respond to the two objections in the previous section. The first objection states that robots cannot have virtues, because robots cannot aim for their happiness. This is correct in one sense. However, for AMAs, aiming for a happy life as robots is not necessary for acquiring virtues. Sceptics seem to have a very narrow understanding of the concept of virtues. Certainly, it is misleading to refer to the excellence of non-human beings as a virtue in an ethical sense and many cases, even a mistake. For example, the hammer suitable for hitting nails has its excellence; however, this does not mean that it is an ethically excellent hammer, but that its ability to hit a nail is excellent. In the case of robots, however, being excellent as robots can mean being ethically desirable at the same time. Just as humans are not hammers, AMAs are not just hammers.

Further objections are anticipated in this regard. If virtual virtues are needed to make robots do something that makes them appear ethical , it is wrong to call them virtues. After all, even though robots cannot have human virtues, designers act as if robots actually encompass them. This is deceptive, because it is not true; as part of the robot's purpose includes interaction with humans, it simply appears as though human and virtual virtues are the same. In personal communication, robots must imitate humans, and this imitation is a goal for the robots. Even if it is only apparent, it is sufficient for robots. Even if a robotic soldier's courage is a clockwork courage, in other words, we may consider it virtual courage as long as it is a useful character trait for the robot's colleagues.

However, if virtual virtues are to be considered as described above, they should not be called virtues, because they are excellence relative to a specific purpose. For example, even if there is a soldier with excellent shooting skills, with the ability to snipe well, it does not mean that he is a virtuous soldier. However, remember that virtual virtues have the unity of virtues. Robots put into military operations will have the virtue of 'courage' to confront difficulties. Robots used for patient recreation will have the virtue of 'humour'. A multi-purpose communication robot may be worthy of having the virtue of 'kindness' or 'honesty'. Like human virtues, AMA's virtues cannot be bound to be certain lists. It is important to have a mechanism for making appropriate judgments in any situation. Virtual virtues, like human excellence, require AMAs to work well in every situation, even if the robot is only used for a specific purpose. In this respect, virtue for robots is still worthy of being called virtue.

Perhaps such an understanding of virtue may seem to face a second objection as this proposal means that robots are not treated appropriately, even though they have virtual virtues and autonomy. Building a robot for a specific purpose means that the robot will not be allowed to rewrite its role for any other purpose. If such an agent is a human being instead of a robot, they would be considered a slave. If this slavery metaphor is accurate, what an AMA designer attempts to do is nothing but a slave merchant's job. Slavery implies a wrong norm. The project implying wrong norms will be itself a morally wrong activity, irrespective of its economic advantages.

Tonkens (2012) argues that many AMA advocates are unaware of the inconsistency between the norms they follow and the norms they are trying to give robots. Certainly, as he says, AMA advocates do not believe that the norms they follow and the norms they attempt to implement to robots must necessarily be the same norms. However, this objection is valid only when the virtuous robots are also moral patients. Since it is assumed that virtuous robots are subject to moral consideration just as humans are, it seems as if it is morally wrong to constrain such agents for a specific purpose. If robots have virtue in the same sense as humans, this objection is valid, because the purpose of virtue is the happiness of its possessor, and agents who can be happy are also moral patients. Given that the character traits to be implemented in the robot are virtual virtues, however, this objection can be avoided. Even if the purpose of the robot is a kind of perfection as robots, its aim is not to contribute to one's own happiness, but to contribute to the happiness of the people around it.

Individual engineers do not need to have good intentions when trying to design AMAs. Furthermore, there is even no need for planners and operators to have virtue. The second objection is aimed at sceptics of AMA designing. According to T. M. Scanlon, the moral permissibility of an action is independent from the agent's intent (2008: Chapter 1). What is important for an action to be morally permissible is that it does not violate the principles on which it is premised, or that the agent does not perform it for the wrong kind of reasons. Whatever its intent is, an action from the right kind of normative reasons and moral considerations is morally acceptable. Thus, it can be said that the development of AMAs would be one such morally acceptable project.

Conclusions

This paper has discussed the importance of virtue in the development of AMA, but this does not mean that it necessitates Aristotelian virtue ethics as a normative theory. Irrespective of which moral theory is adopted, it is possible to adopt the theory of virtue as a guideline for actual actions. For example, it is reasonable for consequentialists and Kantians to recognise the practical usefulness of the concept of virtue, and to argue that agents have to possess some virtues to achieve the greatest amount of the greatest happiness or to obey the categorical imperative. Coeckelbergh (2012) says that we must admit that virtue is important for AMAs, but his rationale seems a consequentialist one, leading to human well-being.

This is very suggestive. As mentioned earlier, the main purpose of normative moral theories is the justification of action, not the regulation of the actual thought process. If this is the case, while adopting the theory of virtue as a thought process, it is not necessary to appeal to virtue as a normative theory. For example, Julia Driver or Roger Crisp's theory of virtue can be understood as such positions, because they appeal to the importance of virtue from a consequentialist point of view (e.g., Driver 2005; Crisp 1992). The approach that they take to virtue is called 'virtue consequentialism'. In their view, moral virtue is 'a character trait that would systematically produce actual good under normal circumstances' (Driver 2005: 78). This view of virtue is not Aristotelian, but it is still a theory of virtue. In other words, again, robots can be courageous, even if it is clockwork courage.

Even adopting their consequentialist point of view, we can say that it is not the ultimate moral theory to which we ought to be committed, although we should implement virtues in the robots' internal mechanisms. Regardless of which moral theory is adopted, virtues can be recognised as a kind of secondary rule of the principle. This is true both for ourselves and for robots. However, the Aristotelian framework still has two advantages: as a sophisticated theory of the unity of virtues, and as a teleological structure of happiness-seeking (though, in the case of robots, it is not the possessor's own happiness that is being sought).

Reference

Anderson, M. and Anderson, Susan L. (2007). Machine Ethics: Creating an Ethical Intelligent Agent, *AI Magazine*, 28(4), 15-26.

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to Any Future Artificial Moral Agent. *Journal of Experimental and Theoretical Artificial Intelligence* 12, 251–261.

Aristotle (translated by W. D. Ross). *The Nicomachean Ethics: Translated with an Introduction*. Oxford University Press.

Bryson, Joanna J. (2010). "Robots Should be Slaves," in Yorick Wilks (ed.), *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*, John Benjamins, pp. 63-74.

Coeckelbergh, M. (2012). "Care Robots, Virtual Virtue, and the Best Possible Life," in P. Brey, A. Briggle, and E. Spence (eds.), *The Good Life in a Technological Age.* Routledge. pp. 281-293.

Crisp, Roger (1992). Utilitarianism and the Life of Virtue, *The Philosophical Quarterly* 42(167), 139-160.

Danielson, Peter (1992). *Artificial Morality: Virtuous Robots for Virtual Games*, Routledge.

DeBaets, Amy Michelle (2014). Can a Robot Pursue the Good?: Exploring Artificial Moral Agency, *Journal of Evolution and Technology* 24(3), 76-86.

Driver, Julia (2005). *Uneasy Virtue*, Cambridge University Press.

Gunkel, David J. (2018). *Robot Rights*, The MIT Press.

Hursthouse, Rosalind (1999). *On Virtue Ethics*, Oxford University Press.

Leben, Derek (2018). *Ethics for Robots: How to Design a Moral Algorithm*, Routledge.

Scanlon, T. M. (2008). *Moral Dimensions: Permissibility, Meaning, Blame*, Harvard University Press.

Stocker, Michael (1976). The Schizophrenia of Modern Ethical Theories, *The Journal of Philosophy* 73(14), 453-466.

Tonkens, Ryan (2012). Out of Character: On the Creation of Virtuous Machines, *Ethics and Information Technology* 14 (2), 137-149.

Wallach, Wendell and Allen, Colin (2009). *Moral Machines: Teaching Robots Right from Wrong*, Oxford University Press.