No.340/February 2019
Analyses & Arguments
Digital Society

Konrad Adenauer Foundation

Robot Ethics

Catrin Misselhorn

- In a restricted sense, autonomous machines (computers, robots) are morally relevant agents. They can be delegated decisions but they do not take responsibility. For this, they lack freedom of will, causality, intention and knowledge.

- In certain situations, autonomous systems can undermine the ascribed responsibility. It becomes unclear who bears the responsibility. A "responsibility gap" arises in which neither programmer nor user can be held fully responsible.

- Especially in matters concerning human life,decisions should not be left to machines. Beyond that, there are many fields for robots to be put to use.

- For robot ethics three guidelines can be laid down: (1) Artificial systems should always further and not hinder the self-determination of humans. (2) They should not decide on the life or death of humans. (3) It must be ensured that humans always exercise control and take responsibility.

Table of Contents

Robotics is an industrial sector of growth promising a lot. According to the data of the *World Robotics Report 2018*, the sale of robots increased by ca. 30 percent in the last year as compared to the preceding year, and the trend is climbing.[1] Besides technical, economic and legal challenges, this also entails ethical problems. This is an issue robot ethics, a new discipline at the interface of information technology, philosophy and robotics, centres around.[2] It has as its object the moral problems in the development, production and use of robots as well as the relationship between humans and robots and observes the societal consequence of a growing robotization.

As is understood in this paper, a robot is an electric-mechanical machine which consists of a processor, sensors and effectors.[3] The term ethics is used in the same meaning as moral philosophy, in the sense of a philosophical discipline which deals with the moral qualities of actions, judgements, traits of character, attitudes and rules as well as institutions. Here *normative* ethics enjoys prominence which does not consider only the description of what people deem as morally right but rather offer well-founded recommendations of what is morally right or wrong.[4]

1. Classification of Robot Ethics

Machine Ethics

The specific ethical questions that robots raise deal with two aspects: on the one hand, with increasing intelligence and autonomy and, on the other, with their outward form and ways of interacting with human beings. The first aspect leads to *machine ethics*, a sub-discipline of robot ethics. In this context, it is about the development of an ethics for machines as opposed to *an ethics for humans dealing with machines.* . In analogy to "Artificial Intelligence" (AI) one also speaks of "Artificial Morality".[5] Whereas "Artificial Intelligence" serves the aim of modelling or simulating the cognitive faculties of humans, "Artificial Morality" is about equipping artificial systems with the ability for moral decision-making and acting. Not any machine can be trusted with the task, but only computers. The idea is to program computers in such a way that they can act morally. Here the software forms a kind of the robot's "brain". The sensors correspond to the sense organs feeding the robot with data about its environment and status. By means of the effectors, the robot can have an effect on its environment and change its position.

The second aspect rather results from the outer form of robots and their interaction with humans. One can also speak of an *Ethics of the Man-Machine-Interaction*. People tend to humanize intelligent systems. We easily regard robots which seemingly behave autonomously and intelligently, possibly showing a human or animal-like form, as beings with similar thoughts, motives and sensations of a human being. This raises several issues.

On the one hand, it raises the question if the interaction with such robots is based on a problematic form of deception or manipulation because it affects us in a way as if they possessed human qualities they do not befit.[6] On the other, there is the question if such machines that evoke such reactions in humans restrict us in moral terms in dealing with them even if they do not really command respective thoughts, motives or sensations.[7] Machine ethics and ethics of man-machine-interaction overlap when the question is if machines possess the capability of moral decision-making and acting, on their part are also subjects of moral demands that humans must heed when dealing with them. The contribution at hand focuses on the aspect of machine ethics.[8] In an open letter signed by an impressive number of AI-researchers and scientists, this is highlighted as one of the most important and pressing fields of AI research.[9]

2. Motivation of Machine Ethics

There is a lot of work we would like to leave to machines because it is too heavy, dangerous or simply unpleasant. In some cases, there are not enough people able or willing to shoulder it. Finally, machines simply can do certain things better and quicker than humans. In many cases, this requires machines to operate as autonomously as possible, i.e. without a human being directly interfering causally. In this way, a robot vacuum cleaner is such a relief from work because it need not be guided by a human being but autonomously moves around in your home when you are not in. This ability to autonomously orientate, navigate and act requires intelligent systems.[10]

Even a simple model, such as a vacuum cleaner robot, now faces an ethical decision, namely: should it suck in a ladybug, chase it away or move around it? What about spiders? The question whether one is allowed to kill an insect for cleaning purposes has become a basic moral issue.
However, common vacuum cleaner robots do not yet possess the ability to make such a decision. There are, however, first attempts to create a version of the popular model *Roomba* extended by an ethical module which takes into account the lives of insects (the prototype is fitted with an optional "kill button" for spiders.)[11]

The more complex the operational fields of autonomous systems, the more demanding the moral decisions to be made. For example, the operational field of moral machines caring for the elderly. Due to the demographic change, the

proportion of people needing special care will strongly increase in the next decades. Artificial systems are frequently cited as a means to cope with the alarming situation in geriatric care. But systems to be applied in this context must face up to moral decision-making, e.g. how frequently and intensively should a system of care remind the patient of needed food and drink as well as taking his or her medication? After how long a period of time should a system of care inform next of kin or call the medical service if someone does not move for a while. Should the system supervise the user around the clock and how is the thus generated data flow to be handled?

Weighing Moral Values

In all these situations, an artificial system must weigh certain moral values: in the first case, for example, between self-determination of the user and certain health risks which emerge if medication is not taken as prescribed. In the second case between the self-determination of the user and the next of kin concerned, who, perhaps, would like to be informed immediately and, again, for reasons of health. The third case also deals with the self-determination of the user, health, next kin's concern and privacy of personal data.

Another widely discussed example of the necessity of moral sensitive machines shows up in autonomous driving. Completely automated vehicles face moral decision-making. Thus they are to be programmed in such a manner that, in unavoidable situations of danger, the protection of human life ranks higher than any possible damage inflicted upon things and animals. As far as possible, the lives of animals should be spared. Under certain circumstances, special difficulties pose as moral dilemmas in this field of application. What about an autonomous vehicle having as its exclusive possibility the killing of a human being at the end of his/her life or that of a small child? What about such a vehicle being able to save five human lives by running over one pedestrian on the pavement? From a moral perspective, is the special protection of the passengers in the vehicle legitimate or do other road users rank higher?[12]

Finally, think of the military users. The dream is about no soldiers having to risk their lives on the battlefield but autonomous machines deployed to fight in their stead.[13] These are to be installed with the international law on war and context-specific rules of deployment which restrict their range of operation and ensure that they operate flawlessly in legal and moral terms. They must decide if and when a military action is necessary and appropriate and how combatants can be distinguished from civilians.

3. Moral Action and Responsibility

Lack of Control and Predictability

However, one could argue that it is not the system caring for the elderly, the autonomous car and combat robots making moral decisions in above mentioned cases but rather the programmers of such appliances. But the greater progress of artificial intelligence, the less are developers able to plan and anticipate what decisions a system will make in certain situations. A chess program plays much better than its programmers who cannot predict each singular move made by the

system. This applies even more so to such a complex system as Alpha Go Zero which, at the beginning, knows only the basic rules of the game and then, by playing against itself a number of games, finds optimal strategies for decision-making. Within a very short time, this system succeeded in beating its precursor Alpha Go which, as the first artificial system, defeated some of the world's best Go players. It is this lack of control and predictability that gives ground to the most important objections to machine ethics.

Difficulty in Ascribing Responsibility

Machines are not totally able to act morally as it befits a human being. Among other things, it lacks consciousness, free will and the ability of self-reflection.[14] As these characteristics are essential for taking up moral responsibility, machines cannot be held accountable for their actions. Nevertheless, one should discuss to what extent the use of machines undermines responsibility ascribed to humans so that possibly, in the end, nobody is responsible for their actions. In this context, there is talk of an emerging *responsibility gap*.[15]

Responsibility Gap

The criteria for ascribing responsibility comprise free will, causality, intentionality and knowledge. Accordingly, an agent is responsible for his/her action only if it is based on his/her free will, if the action had not come about without his/her participation, s/he had carried it out intentionally (or at least accepted its consequences) and had been in the know about its consequences (s/he could have anticipated them or gained respective knowledge with a reasonable effort). Undoubtedly, machines cannot meet all these requirements. They possess no free will; but the requirement of intentionality and knowledge raise problems for the ascription to machines. Therefore they cannot bear any responsibility but rather create a responsibility gap.

The Australian machine ethicist Robert Sparrow, originator of this term (originally *responsibility gap*), uses the example of an autonomous battle robot for his arguments. The core of his argument is as follows: a responsibility gap emerges if: (1) a battle robot has not intentionally been programmed as to violate ethical, respectively legal, norms for the conduct of war; (2) one could not predict the outcome by using the battle robot; and (3) there was no human control over the machine from the very start of the operation.

The problem is that the existence of these three conditions leads to the fact that moral responsibility cannot be ascribed to a human if the machine kills people when in conflict with ethical, respectively legal, norms of warfare. No human intended this, it was not predictable, and nobody had causally the possibility to prevent such a result. A responsibility gap arises at the very point when the machine itself is not responsible but its deployment undermines the requirements of responsibility ascribed to humans. To Sparrow this is the reason to reject the deployment of battle robots as immoral. But, basically, it can be transferred to other areas, especially autonomous driving.

This could serve as a reason to demand that humans are not allowed to completely yield control. In military contexts, there is the distinction between *In-the-Loop-*

Systems, On-the-Loop-Systems and Out of-the-Loop-Systems depending on the role a human plays in this control loop.[16] In the *In-the-Loop-Systems* a human operates the system and makes all the decisions, be it via remote control. *On-the-Loop-Systems* are programmed, but can operate independently of human interference in real time. Supervision is up to the human and s/he has the possibility to intervene any time. *Out-of-the-Loop-Systems* behave like *On-the-Loop-Systems* but there is no more possibility for control and intervention by humans.

The problem of the responsibility gap seems to be solved if the human remains On-the-Loop and perhaps must even consent to taking up responsibility by pushing a button before operating an artificial system.[17] But how realistic is the assumption that a human is capable of a permanent supervision? Can s/he sustain attention for a long time and is s/he ready to decide and intervene in the split of a second in stressful situations? If this were not the case, predictability and control would theoretically be possible, but not feasible for humans in real life.

Furthermore, there is a problem of findings as man depends on information provided by the system for the analysis of his situation. The question is if he can put data in doubt rationally if he has no access to independent information. Additionally, such a system must undergo a series of quality control processes in its development. For the users, this too might be a reason to regard the propositions made by the system to be superior to one's own doubts.

All in all, it seems unfair that the user is to take up total responsibility via a push button because at least part of the responsibility, if not the main part, should be attributed to the programmers whose algorithms are decisive for the actions of the system. The users are responsible only in a weaker sense because they have not hindered the system from acting. All three points make it appear doubtful if the conditions of predictability and control have been fully met. Therefore, the problem of the responsibility gap also looms up the On-the-Loop-Systems. Eventually, it also crops up if a human remains In-the-Loop. This supports the demand by human rights organisations for a human meaningful control.[18]

4. Decisions on Life and Death

Limits to Delegating Moral Decisions

Generally speaking, we should give careful thought to delegating moral decision-making to machines if it involves life or death of humans. A decisive argument against autonomous weapon systems states that there is no moral duty to kill in war.[19] There is only a single permission to kill that puts the general prohibition to kill in parentheses in certain situations. Therefore, the decision to kill a specific human being should always be the obligation of man and not of a machine.

This argument may also be transferred to autonomous driving. An analogy can be drawn between programming autonomous vehicles serving the purpose of accident optimization and targeting autonomous weapon systems.[20] To optimize accident results, it is necessary to cite cost-functions that, in cases of doubt, determine who is

injured and killed. Similar to autonomous weapons systems, legitimate targets must be fixed in the case of an unavoidable collision which would be injured wilfully or even killed.

That would require a moral duty to injure or even kill innocent people if this helps preventing harm. Such a duty obviously would run counter to German jurisdiction. Thus, in the year 2006, the German Federal Constitutional Court stated in its decision on the law of air safety about shooting down a hijacked passenger plane to be employed as weapons of mass destruction by terrorists that shooting it down would always contradict the human dignity of the passengers on board.[21] The Basic Law excludes deliberately killing innocent humans on the basis of a legal authorization. At least at first glance, this sentence contradicts an obligation to minimize damage which includes the deliberate harm or killing of innocent people.

However, there are opinionators that do not yet see the final say in this matter. One proposition consists of starting from a grading in wrongdoing.[22] Accordingly, the killing of innocent humans is illegal (as well as immoral), at the same time there should be a legal and moral obligation to destroy as few lives as possible, even if innocent people are injured or killed deliberately. However, so far, such an obligation does not seem to be generally acknowledged in the dogmatics of law. Furthermore, such a construction appears to be questionable from a moral perspective. As the harming and killing of innocent people is still doing wrong, the people concerned have rights of defence. They are allowed to shoot down a plane that threatens to descend upon them if this option is the lesser evil. This implies that somebody is morally obliged to do something whilst somebody else has the moral right to put a stop to the very action.[23] This hardly seems congruent with the absoluteness and generalization marking moral obligations.

5. Conclusion

Even if one does not want to hand over to machines the decision on life and death, there remain many areas of application in which moral machines can be used meaningfully, such as care.[24] Here one must pay close attention to protecting the human right of self-determination. On the one hand, this applies to the decision if someone wants to be cared for by an artificial system in the first place; this should be optional to everybody. On the other hand, one must see to it that a care system can flexibly adjust to the value system of its user. In modern pluralistic societies one must assume that the value systems of the users are different, e.g. giving privacy more weight than avoiding health risks. A care system should be capable of individually calibrating with the moral standards of the respective user. In this case, a care system can help people, who want this, to live longer and self-determinedly in their homes. However, such a system is suitable only for people who are cognitively able to make basic decisions on their lives but are physically so challenged that they cannot live without care at home by themselves.

Coming to a conclusion, one can lay down the following three principles as basic guidelines of robot ethics:

1. Artificial systems should always further the self-determination of people and not hamper them.
2. They should not decide on life and death of people.
3. It must be secured that humans always exercise control and responsibility for the actions of the machines.

Footnotes

[1] Https://ifr.org/ifr-press-releases/news/industrial-robot-sales-increase-worldwide-by-29-percent (letzter Aufruf: 20.12.18).
[2] The term robot ethics dates back to Veruggio (2005) as the official originator of this discipline.
[3] Vgl. Misselhorn (2013).
[4] For further explication of what is to be understood by moral in this context, cf. Misselhorn (2018)
[5] Vgl. Misselhorn (2018): Artificial Morality
[6] Cf. Scholz (2008) talks of "subject simulating machines" whereas Turkle (2006) defines robots as relational artefacts
[7] Misselhorn 2009, 2018.
[8] As to the second aspect cf. Misselhorn et al. (2013)
[9] Https://futureoflife.org/ai-open-letter (letzter Aufruf: 20.12.18).
[10] Misselhorn (2015)
[11] Cf. Bendel (2017).
[12] On website http://moralmacine.mit.edu/(last call 20.12.18) is to be found a series of such scenarios with different constellations visitors can decide on by mouse click. The data is used,among other things, to develop a procedure of decision-making for moral machines (Awad et al. 2018).
[13] Cf. Arkin (2009).
[14] Cf. Misselhorn (2018).
[15] Cf. Sparrow (2007).
[16] Cf. United States Department of Defense (2011).
[17] Arkin (2009), for example, suggests such a concept for the weapons system he has developed.
[18] Cf. Roff und Moyes (2016).
[19] Cf.Misselhorn (2018).
[20] Cf. Lin (2016), S. 72.
[21] Cf. BVerfGE 115, 118, (160).
[22] Cf. Hilgendorf (2017), S. 155.
[23] Cf. Hilgendorf (2017), S. 155.
[24] Cf. Misselhorn (2019): Moral machines in care

Bibliography

A Arkin, Ronald (2009): Governing lethal behavior in autonomous robots. Boca Raton. Awad, Edmond et al. (2018): The Moral Machine experiment. In: Nature 563, S. 59–64.

B Bendel, Oliver (2017): Ladybird – The animal-friendly robot vacuum cleaner. In: The AAAI 2017 Spring Symposium on Artificial Intelligence for the Social Good Technical Report SS-17-01. Palo Alto, S. 2–6.

H Hilgendorf, Eric (2017): Autonomes Fahren im Dilemma - Überlegungen zur moralischen und rechtlichen Behandlung von selbsttätigen Kollisionsvermeidesystemen. In: Eric Hilgendorf: Autonome Systeme und neue Mobilität - Ausgewählte Beiträge zur 3. und 4. Würzburger Tagung zum Technikrecht. Baden-Baden, S. 143–176.
Hilgendorf, Eric (2017): Autonomous Driving in Dilemma – Reflections on the moral and legal treatment of autonomous collision avoidance systems. In: Eric Hilgendorf: Autonomous Systems and New Mobility – Selected contributions to 3rd and 4th Würzburg Conferences on Technology Law. Baden-Baden, pp. 143-176

L Lin, Patrick (2016): Why ethics matters for autonomous cars. In: Markus Maurer und J. Christian Gerdes und Barbara Lenz et al. (Hrsg.): Autonomous driving – Technical, legal and social aspects. Berlin/Heidelberg, S. 69–85.

M Misselhorn, Catrin (2009): Empathy with inanimate objects and the uncanny valley. In: Minds and Machines 19, S. 345–359.

Misselhorn, Catrin (2013): Robots as moral agents. In: Frank Rövekamp und Friederike Bosse (Hrsg.): Ethics in science and society - German and Japanese views. München, S.30–42.

Misselhorn, Catrin (Hrsg.) (2015): Collective agency and cooperation in natural and artificial systems - Explanation, implementation and simulation. In: Philosophical Studies Series 122.

Misselhorn, Catrin (2018): Grundfragen der Maschinenethik. Ditzingen 2018.
Misselhorn, Catrin (2018): Basic Questions on Machine Ethics. Ditzingen 2018.

Misselhorn, Catrin (2019): Is empathy with robots morally relevant? In: Catrin Misselhorn und Maike Klein (Hrsg): Emotional machines – Perspectives from affective computing and emotional human-machine interaction. Wiesbaden. (Erscheint 2019).

Misselhorn, Catrin und Pompe, Ulrike und Stapleton, Mog (2013): Ethical considerations regarding the use of social robots in the fourth age. In: GeroPsych – The Journal of Gerontopsychology and Geriatric Psychiatry 26, S. 121–133.

R Roff, Heather M. und Moyes, Richard (2016): Meaningful human control, artificial intelligence and autonomous weapons. Briefing paper prepared for the informal meeting of experts on lethal autonomous weapons systems. UN Convention on Certain Conventional Weapons. Geneva.

S Scholz, Christopher (2008): Alltag mit künstlichen Wesen. Theologische Implikationen eines Lebens mit subjektsimulierenden Maschinen am Beispiel des Unterhaltungsroboter Aibo Göttingen 2008.
Scholz, Christopher (2008): Daily Life with Artificial Beings. Theological implications of a life with subject simulating machines as exemplified by the entertainment robot Aibo, Göttingen, 2008.

Sparrow, Robert (2007): Killer robots. In: Journal of Applied Philosophy 24, S. 62–77.

T Turkle, S., Taggart, W., Kidd, C., & Daste, O. (2006). Relational artifacts with children and elders: The complexities of cybercompanionship. Connection Science, 18, 347–361.

U United States Department of Defense (2011): Unmanned systems integrated roadmap FY 2011-2036. Reference Number 11-S-3613. URL: <http://www.dtic.mil/dtic/tr/fulltext/u2/a558615.pdf> (letzter Aufruf: 20.12.18).

V Veruggio, Gianmarco (2005): The birth of roboethics. In: Proceedings of IEEE International Conference on robotics and automation. Genova, S. 1–4.

Imprint

The Author

Prof. Dr. Catrin Misselhorn
Direktorin des Instituts für Philosophie, Lehrstuhl für Wissenschaftstheorie und Technikphilosophie, Universität Stuttgart

Konrad-Adenauer-Stiftung e. V./ Konrad-Adenauser-Foundation

Dr. Norbert Arnold
Leiter des Teams Bildungs- und Wissenschaftspolitik
Hauptabteilung Politik und Beratung
T: +49(0)30 / 26 996-3504
norbert.arnold@kas.de

Konrad-Adenauer-Foundation, 10907 Berlin

Published by: Konrad-Adenauer-Foundation 2018, Sankt Augustin/Berlin

ISBN 978-3-95721-500-0

Translated from the German by York R. Buttler 3/19