3

Artificial Moral Agents

Conceptual Issues and Ethical Controversy

Catrin Misselhorn

I. ARTIFICIAL MORALITY AND MACHINE ETHICS

Artificial Intelligence (AI) has the aim to model or simulate human cognitive capacities. Artificial Morality is a sub-discipline of AI that explores whether and how artificial systems can be furnished with moral capacities.[1] Its goal is to develop artificial moral agents which can take moral decisions and act on them. Artificial moral agents in this sense can be physically embodied robots as well as software agents or 'bots'.

Machine ethics is the ethical discipline that scrutinizes the theoretical and ethical issues that Artificial Morality raises.[2] It involves a meta-ethical and a normative dimension.[3] Meta-ethical issues concern conceptual, ontological, and epistemic aspects of Artificial Morality like what moral agency amounts to, whether artificial systems can be moral agents and, if so, what kind of entities artificial moral agents are, and in which respects human and artificial moral agency diverge.

Normative issues in machine ethics can have a narrower or wider scope. In the narrow sense, machine ethics is about the moral standards that should be implemented in artificial moral agents, for instance: should they follow utilitarian or deontological principles? Does a virtue ethical approach make sense? Can we rely on moral theories that are designed for human social life, at all, or do we need new ethical approaches for artificial moral agents? Should artificial moral agents rely on moral principles at all or should they reason case-based?

In the wider sense, machine ethics comprises the deliberation about the moral implications of Artificial Morality on the individual and societal level. Is Artificial Morality a morally good thing at all? Are there fields of application in which artificial moral agents should not be deployed, if they should be used at all? Are there moral decisions that should not be delegated to machines? What is the moral and legal status of artificial moral agents? Will artificial moral agents change human social life and morality if they become more pervasive?

This article will provide an overview of the most central debates about artificial moral agents. The following section will discuss some examples for artificial moral agents which show that the topic is not just a problem of science fiction and that it makes sense to speak of artificial agents. Afterwards, a taxonomy of different types of moral agents will be introduced that helps to understand the aspirations of Artificial Morality. With this taxonomy in mind, the conditions

[1] C Misselhorn, 'Artificial Morality: Concepts, Issues and Challenges' (2018) 55 *Society* 161 (hereafter Misselhorn, 'Artificial Morality').

[2] SL Anderson, 'Machine Metaethics' in M Anderson and SL Anderson (eds), *Machine Ethics* (2011) 21–27.

[3] C Misselhorn, 'Maschinenethik und Philosophie' in O Bendel (ed), *Handbuch Maschinenethik* (2018) 33–55.

https://doi.org/10.1017/9781009207898.005 Published online by Cambridge University Press

for artificial moral agency in a functional sense will be analyzed. The next section scrutinizes different approaches to implementing moral standards in artificial systems. After these narrow machine ethical considerations, the ongoing controversy regarding the moral desirability of artificial moral agents is going to be addressed. At the end of the article, basic ethical guidelines for the development of artificial moral agents are going to be derived from this controversy.

II. SOME EXAMPLES FOR ARTIFICIAL MORAL AGENTS

The development of increasingly intelligent and autonomous technologies will eventually lead to these systems having to face moral decisions. Already a simple vacuum cleaner like Roomba is, arguably, confronted with morally relevant situations. In contrast to a conventional vacuum cleaner, it is not directly operated by a human being. Hence, it is to a certain degree autonomous. Even such a primitive system faces basic moral challenges, for instance: should it vacuum and hence kill a ladybird that comes in its way or should it pull around it or chase it away? How about a spider? Should it extinguish the spider or save it?

One might wonder whether these are truly moral decisions. Yet, they are based on the consideration that it is wrong to kill or harm animals without a reason. This is a moral matter. Customary Roombas do, of course, not have the capacity to make such a decision. But there are attempts to furnish a Roomba prototype with an ethics module that does take animals' lives into account.[4] As this example shows, artificial moral agents do not have to be very sophisticated and their use is not just a matter of science fiction. However, the more complex the areas of application of autonomous systems get, the more intricate are the moral decisions that they would have to make.

Eldercare is one growing sector of application for artificial moral agents. The hope is to meet demographic change with the help of autonomous artificial systems with moral capacities which can be used in care. Situations that require moral decisions in this context are, for instance: how often and how obtrusively should a care system remind somebody of eating, drinking, or taking a medicine? Should it inform the relatives or a medical service if somebody has not been moving for a while and how long would it be appropriate to wait? Should the system monitor the user at all times and how should it proceed with the collected data? All these situations involve a conflict between different moral values. The moral values at stake are, for instance, autonomy, privacy, physical health, and the concerns of the relatives.

Autonomous driving is the application field of artificial moral agents that probably receives the most public attention. Autonomous vehicles are a particularly delicate example because they do not just face moral decisions but moral dilemmas. A dilemma is a situation in which an agent has two (or more) options which are not morally flawless. A well-known example is the so-called trolley problem which goes back to the British philosopher Philippa Foot.[5] It is a thought experiment which is supposed to test our moral intuitions on the question whether it is morally permissible or even required to sacrifice one person's life in order to save the lives of several persons.

Autonomous vehicles may face structurally similar situations in which it is inevitable to harm or even kill one or more persons in order to save others. Suppose a self-driving car cannot stop and it has only the choice to run into one of two groups of people: on the one hand, two elderly

[4] O Bendel, 'Ladybird: The Animal-Friendly Robot Vacuum Cleaner' (2017) *The AAAI 2017 Spring Symposium on Artificial Intelligence for the Social Good Technical Report SS-17-01* 2-6.

[5] P Foot, *The Problem of Abortion and the Doctrine of Double Effect. Virtues and Vices* (1978) 19–32.

men, two elderly women and a dog; on the other hand, a young woman with a little boy and a little girl. If it hits the first group the two women will be killed, the two men and the dog are going to be severely injured. If it runs into the second group one of the children will get killed and the woman and the other child will be severely injured.

More details can be added to the situation at will. Suppose the group of the elderly people with the dog behaves in accord with the traffic laws, whereas the woman and the children cross the street against the red light. Is this morally relevant? Would it change the situation if one of the elderly men is substituted by a young medical doctor who might save many people's lives? What happens if the self-driving car can only save the life of other traffic participants by sacrificing its passengers?[6] If there is no morally acceptable solution to these dilemmas, this might become a serious impediment for fully autonomous driving.

As these examples show, a rather simple artificial system like a vacuuming robot might already face moral decisions. The more intelligent and autonomous these technologies get, the more intricate the moral problems they confront will become; and there are some doubts as to whether artificial systems can make moral decisions which require such a high degree of sophistication, at all, and whether they should do so.

One might object that it is not the vacuuming robot, the care system, or the autonomous vehicle that makes a moral decision in these cases but rather the designers of these devices. Yet, progress in artificial intelligence renders this assumption questionable. AlphaGo is an artificial system developed by Google DeepMind to play the board game Go. It was the first computer program to beat some of the world's best professional Go players on a full-sized board. Go is considered an extremely demanding cognitive game which is more difficult for artificial systems to win than other games such as chess. Whereas AlphaGo was trained with data from human games; the follow-up version AlphaGoZero was completely self-taught. It came equipped with the rules of the game and perfected its capacities by playing against itself without relying on human games as input. The next generation was MuZero which is even capable of learning different board games without being taught the rules.

The idea that the designers can determine every possible outcome already proves inadequate in the case of less complex chess programs. The program is a far better chess player than its designers who could certainly not compete with the world champions in the game. This holds true all the more for Go. Even if the programmers provide the system with the algorithms on which it operates, they cannot anticipate every single move. Rather, the system is equipped with a set of decision-making procedures that enable it to make effective decisions by itself. Due to the lack of predictability and control by human agents, it makes sense to use the term 'artificial agent' for this kind of system.

III. CLASSIFICATION OF ARTIFICIAL MORAL AGENTS

Even if one agrees that there can be artificial moral agents, it is clear that even the most complex artificial systems differ from human beings in important respects that are central to our

[6] One can find these and some more morally intricate scenarios for self-driving vehicles at http://moralmachine.mit.edu/. The website was created by the MIT with the aim of providing a platform for '1) building a crowd-sourced picture of human opinion on how machines should make decisions when faced with moral dilemmas, and 2) crowd-sourcing assembly and discussion of potential scenarios of moral consequence.' The results were published in different papers that are available at the website.

https://doi.org/10.1017/9781009207898.005 Published online by Cambridge University Press

understanding of moral agency. It is, therefore, common in machine ethics to distinguish between different types of moral agents depending on how highly developed their moral capacities are.[7]

One influential classification of moral agents goes back to *James H. Moor*.[8] He suggested a hierarchical distinction between four types of ethical agents.[9] It does not just apply to artificial systems but helps to understand which capacities an artificial system must have in order to count as a moral agent, although it might lack certain capacities which are essential to human moral agency.

The most primitive form describes agents who generate moral consequences without the consequences being intended as such. *Moor* calls them ethical impact agents. In this sense, every technical device is a moral agent that has good or bad effects on human beings. An example for an ethical impact agent is a digital watch that reminds its owners to keep their appointments on time. However, the moral quality of the effects of these devices lies solely in the use that is made of them. It is, therefore, doubtful whether these should really be called agents. In the case of these devices, the term 'operational morality,' which goes back to *Wendell Wallach* and *Colin Allen*, seems to be more adequate since it does not involve agency.[10]

The next level is taken by implicit ethical agents, whose construction reflects certain moral values, for example security considerations. For *Moor*, this includes warning systems in aircrafts that trigger an alarm if an aircraft comes too close to the ground or if a collision with another aircraft is imminent. Another example are ATMs: these machines do not just have to always emit the right amount of money; they often also check whether money can be withdrawn from the account on that day at all. *Moor* even goes so far as to ascribe virtues to these systems that are not acquired through socialization, but rather directly grounded in the hardware. Conversely, there are also implicit immoral agents with built-in vices, for example a slot machine that is designed in such a way that people invest as much time and money as possible in it. Yet, as in the case of ethical impact agents these devices do not really possess agency since their moral qualities are entirely due to their designers.

The third level is formed by explicit ethical agents. In contrast to the two previous types of agents, these systems can explicitly recognize and process morally relevant information and come to moral decisions. One can compare them to a chess program: such a program recognizes the information relevant to chess, processes it, and makes decisions, with the goal being to win the game. It represents the current position of the pieces on the chessboard and can discern which moves are allowed. On this basis, it calculates which move is most promising under the given circumstances.

For *Moor*, explicit moral agents act not only in accordance with moral guidelines, but also on the basis of moral considerations. This is reminiscent of *Immanuel Kant's* distinction between action in conformity with duty and action from duty.[11] Of course, artificial agents cannot strictly be moral agents in the *Kantian* sense because they do not have a will and they do not have inclinations that can conflict with the moral law. Explicit moral agents are situated somewhere

[7] For an overview, see JA Cervantes and others, 'Artificial Moral Agents: A Survey of the Current Status' (2020) 26 *Science and Engineering Ethics* 501–532.

[8] JH Moor, 'The Nature, Importance, and Difficulty of Machine Ethics' (2006) 21 *IEEE Intelligent Systems* 18–21.

[9] Moor uses the terms 'ethical' and 'moral' synonymously. I prefer to distinguish between these two terms. According to my understanding, morality is the object of ethics. It refers to a specific set of actions, norms, sentiments, attitudes, decisions, and the like. Ethics is the philosophical discipline that scrutinizes morality.

[10] This will be spelled out in the next section. W Wallach and C Allen, *Moral Machines: Teaching Robots Right from Wrong* (2009) 26 (hereafter Wallach and Allen, *Moral Machines*).

[11] M Gregor (ed), *Immaul Kant: Groundwork of the Metaphysics of Morals* (1996) 4:397f.

in between moral subjects in the *Kantian* sense, who act from duty, and *Kant's* example of the prudent merchant whose self-interest only accidentally coincides with moral duty. What *Moor* wants to express is that an explicit moral agent can discern and process morally relevant aspects as such and react in ways that fit various kinds of situations.

Yet, *Moor* would agree with *Kant* that explicit moral agents still fall short of the standards of full moral agency. *Moor's* highest category consists of full ethical agents who have additional capacities such as consciousness, intentionality, and free will, which so far only human beings possess. It remains an open question whether machines can ever achieve these properties. Therefore, *Moor* recommends viewing explicit moral agents as the appropriate target of Artificial Morality. They are of interest from a philosophical and a practical point of view, without seeming to be unrealistic with regard to the technological state of the art.

Moor's notion of an explicit ethical agent can be explicated with the help of the concept of functional morality introduced by *Wallach* and *Allen*.[12] They discriminate different levels of morality along two gradual dimensions: autonomy and ethical sensitivity. According to them, Moor's categories can be situated within their framework.

A simple tool like a hammer possesses neither autonomy nor ethical sensitivity. It can be used to bang a nail or to batter somebody's skull. The possibility of a morally beneficial or harmful deployment would, in Moor's terminology, arguably justify calling it an ethical impact agent, but the artefact as such does not have any moral properties or capacity to act. A child safety lock in contrast does involve a certain ethical sensitivity despite lacking autonomy. It would fall into *Moor's* category of an implicit ethical agent. Because its ethical sensitivity is entirely owed to the design of the object *Wallach* and *Allen* avoid the term of agency and speak of operational morality.

Generally, autonomy and ethical sensitivity are independent of each other.[13] There are, on the one hand, systems which possess a high degree of autonomy, but no (or not much) ethical sensitivity, for example an autopilot. On the other hand, there are systems with a high degree of ethical sensitivity, but no (or a very low degree of) autonomy, for example the platform 'MedEthEx' which is a computer-based learning program in medical ethics.[14] 'MedEthEx' as well as the autopilot belong to the category of functional morality for *Wallach* and *Allen*. Functional morality requires that a machine has 'the capacity for assessing and responding to moral challenges'.[15] This does not necessarily seem to involve agency. If this is the case, there is a level of functional morality below the level of moral agency.[16] Therefore, it has to be specified in more detail which conditions a functional artificial moral agent has to meet.

IV. ARTIFICIAL SYSTEMS AS FUNCTIONAL MORAL AGENTS

There seems to be an intuitive distinction between the things that merely happen to somebody or something and the things that an agent genuinely does.[17] The philosophical question is how to distinguish an action from a mere happening or occurrence and which capacities an object must

[12] Wallach and Allen, *Moral Machines* (n 10) 26.
[13] Ibid, (n 10) 32.
[14] M Anderson, SL Anderson, and C Armen, 'MedEthEx: A Prototype Medical Ethics Advisor' (2006) Proceedings of the Eighteenth Innovative Applications of Artificial Intelligence Conference.
[15] Wallach and Allen, *Moral Machines* (n 10) 9.
[16] Ibid. (n 10) 27.
[17] E Himma, 'Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties Must an Artificial Agent Have to Be a Moral Agent?' (2009) 11 *Ethics and Information Technology* 19–29 (hereafter Himma, 'Artificial Agency'); G Wilson and S Shpall, 'Action' in EN Zalta (ed), *Stanford Encyclopedia of Philosophy* (2012).

have in order to qualify as an agent. The range of behaviors that count as actions is fairly broad. It starts from low-level cases of purposeful animal behavior like a spider walking across the table and extends to high-level human cases involving intentionality, self-consciousness, and free will.[18]

A minimal condition for agency is interactivity, i.e. 'that the agent and its environment [can] act upon each other.'[19] Yet, interactivity is not sufficient for agency. The interactions of an agent involve a certain amount of autonomy and intelligence which can vary in degree and type.

The view is expressed, for instance, by the following definition of an artificial agent:

> The term agent is used to represent two orthogonal entities. The first is the agent's ability for autonomous execution. The second is the agent's ability to perform domain-oriented reasoning.[20]

The term 'autonomous execution' means that, although the system is programmed, it acts in a specific situation without being operated or directly controlled by a human being. A higher degree of autonomy arises if a system's behavior becomes increasingly flexible and adaptive, in other words, if it is capable of changing its mode of operation or learning.[21]

Different natural and artificial agents can be situated at different levels of agency depending on their degree and type of autonomy and intelligence. They can, for instance, be classified as goal-directed agents, intentional agents, agents with higher order intentionality, or persons.[22] Distinctive of moral agency is a special kind of domain-oriented reasoning. Explicit ethical agents in *Moor's* sense of the term would have to be able to act from moral reasons.

According to the philosophical standard theory which goes back to *David Hume*, a reason for an action consists in a combination of two mental attitudes: a belief and a pro-attitude. A belief consists in holding something true; a pro-attitude indicates that something ought to be brought about that is not yet the case. Desires are typical pro-attitudes. For this reason, the approach is also often called Belief-Desire-Theory. Take an example: The reason for my action of going to the library may be my desire to read *Leo Tolstoy's* novel 'Anna Karenina', together with the belief that I will find the book in the library. Some versions of the standard theory assume that action explanation also has to refer to an intention that determines which desire will become effective and that includes some plan of action.[23] This accommodates the fact that we have a large number of noncommittal desires that do not lead to actions.[24]

A moral action can thus be traced back to a moral reason, in other words to some combination of moral pro-attitude and corresponding belief. A moral reason may comprise, for instance, the utilitarian value judgment that it is good to maximize pleasure (pro-attitude) and the belief that making a donation to a charitable organization will result in the overall best balance of pleasure versus pain.[25]

[18] H Frankfurt, 'The Problem of Action'(1978) 15 *American Philosophical Quarterly*, 157–162.

[19] L Floridi and JW Sanders, 'On the Morality of Artificial Agents'(2004). 14 *Minds and Machines*, 349, 357 (hereafter Floridi and Sanders, 'On the Morality of Artificial Agents').

[20] The MuBot Agent, cited by S Franklin and A Graesser 'Is It an Agent, or Just a Program?: A Taxonomy for Autonomous Agents' in JP Mueller, MJ Wooldridge and NR Jennings (eds) *Intelligent Agents III Agent Theories, Architectures, and Languages* (1997) 22.

[21] Floridi and Sanders, 'On the Morality of Artificial Agents' (n 19), regard adaptivity as a separate condition of agency in addition to interactivity and basic autonomy. I prefer to describe it as a higher degree of autonomy. But this might just be a terminological difference.

[22] C Misselhorn 'Collective Agency and Cooperation in Natural and Artificial Systems' in C Misselhorn (ed), *Collective Agency and Cooperation in Natural and Artificial Systems. Explanation, Implementation and Simulation* (2015) 3–25.

[23] ME Bratman, *Intention, Plans, and Practical Reason* (1987).

[24] In the following, this complication is set aside for the sake of simplicity.

[25] It is assumed that we have an intuitive grasp of what moral judgements are. More explicit criteria are given in C Misselhorn, *Grundfragen der Maschinenethik* (4th ed. 2020) (hereafter Misselhorn, *Grundfragen*).

It is a matter of controversy whether artificial systems can possess mental states such as beliefs and desires. Some authors argue that this is not the case because artificial systems do not have intentionality. Intentionality in this sense refers to the fact that mental states like beliefs and desires are about or represent objects, properties, or states of affairs. Most famously *Donald Davidson* assumed that intentionality presupposes complex linguistic abilities which, only humans have.[26] Others concede that animals might also possess intentional states like beliefs and desires, although they do not meet *Davidson's* strong requirements for rationality.[27] This seems to bring intentional agency within the reach of artificial systems as well.[28]

Which stand one takes on this issue depends on the conditions that have to be fulfilled in order to attribute beliefs and desires to an artificial system. According to an instrumentalist view which is often ascribed to *Daniel Dennett*, attributing intentional states is just an explanatory strategy. He argues that states like beliefs and desires are attributed to an agent if this assumption helps us to better understand its behavior, independently of whether there are any corresponding inner states. *Dennett* calls this the intentional stance and the systems that can thus be explained intentional systems. What matters is that we can explain and predict a system's behavior fruitfully by ascribing intentional states to it:

> The success of the stance is of course a matter settled pragmatically, without reference to whether the object really has beliefs, intentions, and so forth; so whether or not any computer can be conscious, or have thoughts or desires, some computers undeniably are intentional systems, for they are systems whose behavior can be predicted, and most efficiently predicted, by adopting the intentional stance toward them.[29]

Rational agency is thus a matter of interpretation and does not require that an entity actually possesses internal states, such as beliefs and desires. This condition can be satisfied by artificial systems. For example, if we can understand a chess computer by assuming that it wants to win the game and thinks that a certain move is appropriate to do so, then we can attribute the appropriate reason for action to the computer. Although the behavior of the computer could, in principle, be explained in purely physical terms, the intentional stance is particularly helpful with regard to complex systems.

In contrast, non-instrumental views are not satisfied with reducing intentionality to an attributional practice. Rather, an entity must have certain internal states that are functionally equivalent to beliefs and pro-attitudes.[30] If an artificial system possesses states which have an analogous function for the system as the corresponding mental states have in humans, the system may be called functionally equivalent to a human agent in this respect.

Since there are different ways of specifying the relevant functional relations, functional equivalence has to be seen relative to the type of functionalism one assumes. The most straightforward view with regard to Artificial Morality is machine functionalism which equates the mind directly with a Turing machine whose states can be specified by a machine table. Such a machine table consists of conditionals of the form: 'if the machine is in state S_i and receives input I_j it emits output O_k and goes into state S_l.'[31]

[26] D Davidson, 'Rational Animals' (1982) 36 *Dialectica* 317–327.
[27] F Dretske, *Explaining Behavior: Reasons in a World of Causes* (4th printing 1995).
[28] Dretske remained, however, skeptical with regard to the possibility of obtaining genuine AI as long as artificial systems lack the right kind of history; see F Dretske, 'Can Intelligence Be Artificial?' (1993) 71 *Philosophical Studies* 201–216.
[29] D Dennett, 'Mechanism and Responsibility' in T Honderich (ed), *Essays on Freedom of Action* (1973) 164–165.
[30] This view can also be used to characterize the intentional states of group agents, see C List and P Pettit, *Group Agency: The Possibility, Design, and Status of Corporate Agents* (2011).
[31] N Block, 'Troubles with Functionalism' (1978) 9 *Minnesota Studies in the Philosophy of Science* 261, 266.

Analytic functionalism specifies the relevant functional relations by the causal role of mental terms in folk psychology and rests on the analysis of the meanings of mental terms in ordinary language. Psycho-functionalism, in contrast, defines mental states by their functional role in scientific psychology. This leads to different ways of specifying the relevant inputs, outputs, and internal relations. Analytic functionalism relies on externally observable inputs and outputs, in other words, objects which are located in the vicinity of an organism and bodily movements, as well as common sense views about the causal relations between mental states. Psycho-functionalism can, in contrast, describe functional relations at a neuronal level.

The different types of functionalism also differ with respect to the granularity of their descriptions of the structure of mental states. Simple machine functionalism, for instance, takes mental states like beliefs or desires as unstructured entities. The representational theory of the mind, in contrast, regards mental states as representations with an internal structure that explains the systematic relations between them and the possibility to form indefinitely many new thoughts. The thought 'John loves Mary' has, for instance, the components 'John', 'loves' and 'Mary' as its constituents that can be combined to form other thoughts like 'Mary loves John'.

The most famous proponent who combines a representational view with a computational theory of the mind is *Jerry Fodor*. He regards mental processes as Turing-style computations that operate over structured symbols which are similar to expressions in natural language and form a 'language of thought'.[32] According to *Fodor* and a number of other cognitive scientists, Turing-style computation over mental symbols is 'the only game in town', in other words the only theory that can provide the foundations for a scientific explanation of the mind in cognitive science.[33]

Although the computational model of the mind became enormously influential in the philosophy of mind and cognitive science, it has also been severely criticized. One of the most famous objections against it was developed by *John Searle* with the help of the thought experiment of the Chinese Room.[34] It is supposed to show that Turing-style computation is not sufficient for thought. *Searle* imagines himself in a room manually executing a computer program. Chinese symbols, that people from outside the room slide under the door, represent the input. *Searle* then produces Chinese symbols as an output on the basis of a manual of rules that links input and output without specifying the meaning of the signs. Hence, he produces the appearance of understanding Chinese by following a symbol processing program but does not actually have any language proficiency in Chinese. Because he does not know Chinese, these symbols are only meaningless squiggles to him. Yet, his responses make perfect sense to the Chinese people outside the room. The thought experiment is supposed to trigger the intuition that the system clearly does not understand Chinese, although its behavior is from the outside indistinguishable from a Chinese native speaker. One might also understand the argument as making the point that syntax is not sufficient for semantics, and that computers will never have genuine understanding viz. intentionality because they can only operate syntactically.

If *Searle* is right, machines cannot really possess mental states. They might, however, exhibit states that are functionally equivalent to mental states although they are not associated with phenomenal consciousness and have only derived intentionality mediated by their programmers and users. One might call such states quasi-beliefs, quasi-desires, etc.[35] This way of speaking borrows from the terminology of *Kendall Walton*, who calls emotional reactions to fiction

[32] JA Fodor, *The Language of Thought* (1975).
[33] For a critical assessment of this claim see E Thompson, *Mind in Life* (2007).
[34] JR Searle, 'Minds, Brains, and Programs' (1980) 3 *The Behavioral and Brain Sciences* 417.
[35] Misselhorn, *Grundfragen* (n 25) 86.

(for example, our pity for the protagonist of the novel 'Anna Karenina') quasi-emotions.[36] This is because they do resemble real emotions in terms of their phenomenal quality and the bodily changes involved: we weep for Anna Karenina and feel sadness in the face of her fate. Unlike genuine emotions, quasi-emotions do not involve the belief that the object that triggers the emotion exists.

With artificial moral agents, it is the other way around. They possess only quasi-intentional states that are, unlike their genuine counterparts, not associated with phenomenal consciousness and have only derived intentionality to speak with *Searle* again. For an explicit moral agent in the sense specified above with regard to *Moore's* classification of artificial moral agents, it seems to be sufficient to have such quasi-intentional states. Given the gradual view of moral agency that was introduced in this section, these agents may be functional moral agents although they are not full moral agents on a par with human beings. Arguments to the effect that artificial systems cannot be moral agents at all because they lack consciousness or free will are, hence, falling short.[37]

Functional moral agents are, however, limited in two ways. First, the functional relations just refer to the cognitive aspect of morality. The emotional dimension could be considered only insofar as emotions can be functionally modelled independently of their phenomenal quality. Secondly, functional equivalence is relative to the type of functionalism embraced and functional moral agents possess (so far) at most a subset of the functional relations that characterize full human moral agents. This holds all the more since artificial system's moral reasoning is to date highly domain specific.

It is also important to stress that the gradual view of agency does not imply that functional moral agents are morally responsible for their doings. From a philosophical point of view, the attribution of moral responsibility to an agent requires free will and intentionality.[38] These conditions are not met in the case of functional moral agents. Hence, they do not bear moral responsibility for their doings.

The most fruitful view for the design of artificial moral agents thus lies somewhere in between *Dennett's* instrumentalist conception, which largely abstracts from the agent's internal states, and computational functionalism as a reductive theory of the mind.[39] *Dennett* makes it too easy for machines to be moral agents. His position cannot provide much inspiration for the development of artificial moral agents because he sees the machine merely as a black box; *Fodor's* psycho-functionalism, on the other hand, makes it extremely difficult.

V. APPROACHES TO MORAL IMPLEMENTATION: TOP-DOWN, BOTTOM-UP, AND HYBRID

Moral implementation is the core of Artificial Morality.[40] It concerns the question of how to proceed when designing an artificial moral agent. One standardly distinguishes between

[36] K Walton, 'Fearing Fictions' (1978) 75 *The Journal of Philosophy* 5.

[37] Himma, 'Artificial Agency' (n 17).

[38] F Rudy-Hiller, 'The Epistemic Condition for Moral Responsibility' (2018) in EN Zalta (ed), *Stanford Encyclopedia of Philosophy*.

[39] C Allen, *Intentionality: Natural and Artificial* (2001), even suggests to regard the concept of intentionality as relative to certain explanatory purposes.

[40] For the following, see C Misselhorn, 'Artificial Systems with Moral Capacities? A Research Design and its Implementation in a Geriatric Care System' (2020) 278 *Artificial Intelligence* https://philpapers.org/rec/MISASW https://dl.acm.org/doi/abs/10.1016/j.artint.2019.103179 (hereafter Misselhorn, 'Artificial Systems with Moral Capacity'). This article also specifies a methodological framework for implementing moral capacities in artificial systems.

top-down, bottom-up, and hybrid approaches.[41] All three methods bring together a certain ethical view with a certain approach to software design.

Top-down approaches combine an ethical view that regards moral capacities as an application of moral principles to particular cases with a top-down approach to software design. The basic idea is to formulate moral principles like *Kant's* categorical imperative, the utilitarian principle of maximizing utility, or *Isaac Asimov's* three laws of robotics as rules in a software which is then supposed to derive what has to be morally done in a specific situation. One of the challenges that such a software is facing is how to get from abstract moral principles to particular cases. Particularly with respect to utilitarian systems, the question arises as to how much information they should take into account as 'the consequences of an action are essentially unbounded in space and time'.[42] Deontological approaches might, in contrast, require types of logical inference which lead to problems with decidability.[43]

A more fundamental objection against top-down approaches regarding Artificial Morality is the so-called frame problem. Originally, the frame problem referred to a technical problem in logic-based AI. Intuitively speaking, the issue is sorting out relevant from irrelevant information. In its technical form, the problem is that specifying the conditions which are affected by a system's actions does not, in classical logic, license an inference to the conclusion that all other conditions remain fixed. Although the technical problem is largely considered as solved (even within strictly logic-based accounts), there remains a wider, philosophical version of the problem first stated by *John McCar*thy and *Patrick Hayes* which is not yet close to a solution.[44]

The challenge is that potentially every new piece of information may have an impact on the whole cognitive system of an agent. This observation has been used as evidence against a computational approach to the mind because it seems to imply that central cognitive processes cannot be modelled by strictly general rules. A corresponding line of argument can also be turned against top-down approaches regarding Artificial Morality. As *Terry Horgan* and *Mark Timmons* point out, moral normativity is not fully systematizable by exceptionless general principles because of the frame problem.[45] Full systematizability is, however, not required for Artificial Morality, and *Horgan* and *Timmons* admit that a partial systematization of moral normativity via moral principles remains possible. The frame problem is, hence, not a knock-down argument against the possibility of top-down approaches to moral implementation although it remains a challenge for AI in general.

The alternative to top-down are bottom-up approaches which do not understand morality as rule-based. This view is closely related to moral particularism, a meta-ethical position that rejects the claim that there are strict moral principles and that moral capacities consist in the application of moral principles to particular cases.[46] Moral particularists use to think of moral capacities in terms of practical wisdom or in analogy to perception as attending to the morally relevant features (or values) that a situation instantiates. Moral perception views emphasize the individual

[41] Wallach and Allen, *Moral Machines* (n 10).
[42] Wallach and Allen, *Moral Machines* (n 10) 86.
[43] TM Powers, 'Prospects for a Kantian Machine' in M Anderson and SL Anderson (eds), *Machine Ethics* (2011) 464.
[44] J McCarthy and PJ Hayes, 'Some Philosophical Problems from the Standpoint of Artificial Intelligence' in B Meltzer and D Michie (eds), *Machine Intelligence* (1969) 463; M Shanahan, 'The Frame Problem' in EN Zalta (ed), *The Stanford Encyclopedia of Philosophy* (2009).
[45] T Horgan and M Timmons 'What Does the Frame Problem Tell Us About Moral Normativity?' (2009) 12 *Ethical Theory and Moral Practice* 25.
[46] J Dancy, *Ethics Without Principles* (2004).

sensibility to the moral aspects of a situation.[47] The concept of practical wisdom goes back to *Aristotle* who underlined the influence of contextual aspects which are induced by way of socialization or training. In order to bring these capacities about in artificial systems, bottom-up approaches in software design which start from finding patterns in various kinds of data have to be adapted to the constraints of moral learning. This can be done either with the help of an evolutionary approach or by mimicking human socialization.[48]

Bottom-up approaches might thus teach us something about the phylo- and ontogenetical evolution of morality.[49] But, they are of limited suitability for implementing moral capacities in artificial systems because they pose problems of operationalization, safety, and acceptance. It is difficult to evaluate when precisely a system possesses the capacity for moral learning and how it will, in effect, evolve. Because the behavior of such a system is hard to predict and explain, bottom-up approaches are hardly suitable for practical purposes; they might put potential users at risk. Moreover, it is difficult to reconstruct how a system arrived at a moral decision. Yet, it is important that autonomous artificial systems do not just behave morally, as a matter of fact, but that the moral basis of their decisions is transparent. Bottom-up approaches should, as a consequence, be restricted to narrowly confined and strictly controlled laboratory conditions.

Top-down and bottom-up are the most common ways to think about the implementation of moral capacities in artificial systems. It is, however, also possible to combine the virtues of both types of approaches. The resulting strategy is called a hybrid approach. Hybrid approaches operate on the basis of a predefined framework of moral values which is then adapted to specific moral contexts by learning processes.[50] Which values are given depends on the area of deployment of the system and its moral characteristics. Although hybrid approaches are promising, they are still in the early stages of development. So, which approach to moral implementation should one choose? It does not make much sense to answer this question in the abstract. It depends on the purpose and context of use for which a system is designed. An autonomous vehicle will demand a different approach to moral implementation than a domestic care robot.

VI. ETHICAL CONTROVERSY ABOUT ARTIFICIAL MORAL AGENTS

Machine ethics, however, does not just deal with issues about moral agency and moral implementation. It also discusses the question of whether artificial moral agents should be approved from a moral point of view. This became a major topic in the last years because Artificial Morality is part of technological innovations that are disruptive and can change individual lives and society profoundly. Not least, a lot of effort and money is spent on research on artificial moral agents in different domains, which also receives a lot of public and media attention. A number of big companies and important economic players strongly push Artificial Morality in areas like autonomous driving, and politics removes, under the perceived economic pressure, more and more legal barriers that might so far prevent the commercial launch of these technologies.

[47] M Nussbaum, 'Finely Aware and Richly Responsible: Moral Attention and the Moral Task of Literature'(1985) 82 *Journal of Philosophy* 516.

[48] For the first approach, see T Froese and E Di Paolo, 'Modelling Social Interaction as Perceptual Crossing: An Investigation into the Dynamics of the Interaction Process' (2010) 22 *Connection Science* 43; for the second, see C Breazeal and B Scassellati, 'Robots That Imitate Humans' (2002) 6 *Trends in Cognitive Sciences* 481; T Fong, N Illah, and K Dautenhahn, 'A Survey of Socially Interactive Robots: Concepts, Design, and Applications' (2002) CMU-RI-TR Technical Report 2.

[49] R Axelrod, *The Evolution of Cooperation* (1984).

[50] For a hybrid approach to a software module for an elder care system, see Misselhorn, 'Artificial Systems with Moral Capacity' (n 40).

The ethical evaluation ranges from a complete refusal of artificial moral agents, over balanced assessments stressing that the moral evaluation of Artificial Morality has to take into account the diversity of approaches and application contexts, to arguments for the necessity of artificial moral agents.[51] The following overview tries to take up the most salient issues but it does not intend to be exhaustive. It focusses on questions that arise specifically with respect to artificial moral agents and does not comment on topics like privacy that belong to the more generic discipline of ethics of AI.

1. *Are Artificial Moral Agents Inevitable?*

One important argument in the discussion is that artificial moral agents are inevitable.[52] The development of increasingly complex intelligent and autonomous technologies will eventually lead to these systems having to face morally problematic situations which cannot be fully controlled by human operators. If this is true, the need for artificial moral agents is eventually arising from technological progress. It would, however, be wrong to either accept this development fatalistically or to reject it as such, because inevitability is a conditional matter. If we want to use intelligent and autonomous technologies in certain areas of application, then this will inevitably lead to the need for artificial moral agents. Hence, we should deliberate in which areas of application – if any – it is right from a moral point of view to use such agents and in which areas it would be morally wrong.[53]

2. *Are Artificial Moral Agents Reducing Ethics to Safety?*

Another motivation for building artificial moral agents is a concern with safety. The idea is that equipping machines with moral capacities can prevent them from harming human beings. It would, however, be wrong to reduce ethics to safety issues.[54] There are other important moral values that can conflict with safety and that have to be taken into consideration by an artificial moral agent. In the context of elder care, safety would, for instance, consist in avoiding health risks at all costs. Yet, this might conflict with the caretakers autonomy.[55] Although safety is a moral value that has to be taken into consideration in developing artificial moral agents, Artificial Morality cannot be reduced to it.

3. *Can Artificial Moral Agents Increase Trust in AI?*

A third aspect that is invoked in the discussion is that artificial moral agents will increase public trust in artificial intelligence. The hope is that Artificial Morality might in this way help to deal with the fears that many people feel with regard to artificial intelligence and robots and improve

[51] For the first position, see A Van Wysberghe and S Robbins, 'Critiquing the Reasons for Making Artificial Moral Agents' (2019) 25 *Science and Engineering Ethics* 719 (hereafter Van Wysberghe and Robbins, 'Critiquing the Reasons'). For the intermediate view, see Misselhorn, 'Artificial Morality' (n 1) and for the last view, see P Formosa and M Ryan, 'Making Moral Machines: Why We Need Artificial Moral Agents' (2020) *AI & Society* https://link.springer.com/article/10.1007/s00146-020-01089-6 (hereafter Formosa and Ryan, 'Making Moral Machines').

[52] This claim is defended by, among others, C Allen and W Wallach, 'Moral Machines: Contradiction in Terms or Abdication of Human Responsibility?' in P Lin, K Abney, and GA Bekey (eds), *Robot Ethics: The Ethical and Social Implications of Robotics* (2011) 55.

[53] Misselhorn, 'Artificial Morality' (n 1).

[54] Van Wysberghe and Robbins, 'Critiquing the Reasons' (n 51).

[55] Misselhorn, 'Artificial Systems with Moral Capacity' (n 40); Formosa and Ryan, 'Making Moral Machines' (n 51).

the acceptance of these technologies.[56] One must, however, distinguish between trust and reliance.[57] Trust is an emotional attitude that arises in a relationship involving mutual attitudes toward one another which are constitutive.[58] It does, for instance, lead to the feeling of being betrayed and not just disappointed when let down.[59] This presupposes the ascription of moral responsibility that must be denied to functional moral agents as argued above. Hence, we should rather speak of reliance instead of trust in artificial moral agents.

It is, moreover, advisable not to be too credulous with regard to artificial moral agents. The lack of predictability and control invoked before to justify why it is adequate to speak of moral agents is also a good reason for not relying blindly on them. The danger is that the term 'Artificial Morality' is suggestively used to increase unjustified acceptance although we should, from a moral point of view, rather keep a critical eye on artificial moral agents.

Even if artificial moral agents do not fulfill the conditions for trustworthiness, trust may play a role with respect to the design and development of artificial moral agents. Suggestions to ensure trust in these cases include a code of conduct for the designers of these devices, transparency with regard to moral implementation, and design of artificial moral agents, as well as standards and certifications for the development process comparable to FairTrade, ISO, or GMOs.[60] Particularly in areas of application that concern not just the users of artificial moral agents but affect the population more broadly or have a large impact on the public infrastructure, like autonomous driving, it is a political task to establish democratically legitimized laws for the design and development process of artificial moral agents or even to constrain their development if necessary.

4. Do Artificial Moral Agents Prevent Immoral Use by Design?

Another argument in favor of artificial moral agents is that they prevent being used immorally by design. Major objections against this argument are that this massively interferes with the autonomy of human beings and can lead to unfair results. Amazon is, for instance, about to install a system called Driveri in their delivery vehicles in the United States. This is an automated monitoring system that consists of high-tech cameras combined with a software which is used to observe and analyze the drivers' behavior when operating the car. It gives real-time feedback in certain cases, for instance, when the driver is going too fast, seems to be distracted, or does not wear a seatbelt. When it comes to the conclusion that something went badly wrong, it will give the information to actual humans at the company.[61] The data are also used to evaluate the drivers and might lead to them being fired – by a machine. Amazon promotes the system as improving safety. But it is clear that it cannot take the subtleties and complexities of human life into account. Sometimes there are good reasons to deviate from the rules or there are special circumstances that the drivers could not influence. This may lead to unfair decisions and hardships that can destroy lives.[62]

[56] M Anderson, SL Anderson, 'Machine Ethics: Creating an Ethical Intelligent Agent' 28 *AI Magazine* 15.
[57] Van Wysberghe and Robbins, 'Critiquing the Reasons' (n 51).
[58] C McLeod, 'Trust' in EN Zalta (ed), *Stanford Encyclopedia of Philosophy*.
[59] A Baier, 'Trust and Antitrust'(1986) 96 *Ethics* 231; J Simon, 'The Entanglement of Trust and Knowledge on the Web' (2010) 12 *Ethics and Information Technology* 343.
[60] Van Wysberghe and Robbins, 'Critiquing the Reasons' (n 51) 728.
[61] J Stanley, 'Amazon Drivers Placed Under Robot Surveillance Microscope' (*ACLU*, 23 March 2021) www.aclu.org/news/privacy-technology/amazon-drivers-placed-under-robot-surveillance-microscope/.
[62] S Soper, 'Fired by Bot at Amazon: "It's You Against the Machine"' (*Bloomberg*, 28 June 2021) www.bloomberg.com/news/features/2021-06-28/fired-by-bot-amazon-turns-to-machine-managers-and-workers-are-losing-out.

Consider some other examples: how about a woman who had a couple of drinks with her partner at home and then refuses to have sex with him. Imagine that her partner gets violent and the woman tries to get away by car but the breathalyzer reacts to the alcohol in her breath and does not let her start the car.[63] Is it the right decision from a moral point of view to prevent the woman from driving because she drank alcohol and to expose her to domestic violence? How about elderly persons at home who ask their service robots for another glass of wine or pizza every day? Should the robot deny to get these things if it thinks that they are a health risk for the user as it happens in the Swedish TV-series *Real Humans*? Examples like these show that it is far from clear which uses are strictly immoral and should be precluded by design. One might, of course, try to deal with the problem by giving people always the possibility to override the system's decisions. But that would undermine the whole purpose of preventing immoral uses by design.

5. *Are Artificial Moral Agents Better than Humans?*

A yet stronger claim is that artificial moral agents are even morally better than humans because their behavior is not influenced by irrational impulses, psychopathologies, or emotional distress. They are impartial, not prone to bias, and they are not diverted from the path of virtue by self-interest. Moreover, machines may be superior to humans in their cognitive abilities. They are able to make decisions in fractions of a second, during which a human being cannot come to conscious decisions. This is used as an argument for leaving moral decisions to machines in particularly precarious situations, for example in war.[64]

Apart from the fact that this argument presupposes an idealized view of AI which does, for instance, ignore the problem of algorithmic bias, several objections have been raised against it. Many argue that artificial systems lack important capacities that human moral agents possess. One point is that emotions are vital for moral judgment and reasoning and that artificial moral agents with emotions are 'something not even on the horizon of AI and robotics'.[65]

As explicated above, this point is somewhat simply put. Emotional AI is a strongly emergent research program inspired by the insights of research in psychology and neuroscience on the importance of emotions for intelligent behavior that goes back to the 1980s.[66] As with artificial moral agency, the state of the art consists in trying to model states that are functionally equivalent to emotions at different levels of granularity.[67] There are even attempts to build artificial moral agents with emotional or empathic capacities.[68] The crucial point is not that emotions are out of the reach of AI, it is that moral emotions involve consciousness and that there is serious doubt that consciousness can be computationally modelled. The crucial question is, therefore, whether functional moral agency is achievable without consciousness.

[63] This case is a slight variation of an example from Van Wysberghe and Robbins, 'Critiquing the Reasons' 729 (n 51).
[64] R Arkin, *Governing Lethal Behavior in Autonomous Robots* (2009) (hereafter Arkin, *Governing*).
[65] Van Wysberghe and Robbins, 'Critiquing the Reasons', 730 (n 51).
[66] M Minsky, *The Emotion Machine: Commonsense Thinking, Artificial Intelligence and the Future of the Human Mind* (2006); R Picard, *Affective Computing* (1997).
[67] R Reisenzein and others, 'Computational Modeling of Emotion: Toward Improving the Inter- and Intradisciplinary Exchange' (2013) 4 *IEEE Transactions on Affective Computing* 246.
[68] C Balkenius and others, 'Outline of a Sensory-motor Perspective on Intrinsically Moral Agents' (2016) 24 *Adaptive Behavior* 306; C Misselhorn, *Künstliche Intelligenz und Empathie. Vom Leben mit Emotionserkennung, Sexrobotern & Co* (2021).

https://doi.org/10.1017/9781009207898.005 Published online by Cambridge University Press

6. Does Reasonable Pluralism in Ethics Speak against Artificial Moral Agents?

A rather desperate move by the adversaries of Artificial Morality is to mount moral skepticism, subjectivism, or an error-theory of moral judgments against it.[69] It is true, if there is no moral right and wrong that is at least intersubjectively binding or if all moral judgments are false, then the development of artificial moral agents would not make sense from the start. But this strategy overstates the case and cures one evil with a worse one. The fact of reason, as *Kant* called it; our existing moral practice is enough for getting Artificial Morality off the ground if there are no other reasons against it.

Having said this, one still has to take into account the fact that there is no consensus about the correct moral theory, neither in the general public nor among philosophers. *John Rawls* calls this 'the fact of reasonable pluralism' and he thinks that it is due to burdens of judgment that we cannot overcome. Reasonable pluralism is, for him, 'the inevitable long-run result of the powers of human reason at work within the background of enduring free institutions.'[70] The question then is which morality should be implemented in artificial systems.

The answer to this question depends on the context. Service, care, or household robots that only affect one individual could be programmed in a way that responds to the individual moral framework of the user.[71] If a system operates, in contrast, in the public sphere and its decisions inevitably concern the vital interests of other people apart from its user, the system's behavior should be governed by generally binding political and legal regulations. This would hold, for instance, for autonomous driving. Ethical pluralism is no insurmountable obstacle to establishing laws with respect to controversial ethical issues in liberal democracies. Examples that show this are (at least in Germany) abortion or assisted dying. Although not every individual agrees entirely with the legal regulations in these cases, most citizens find them morally acceptable, although they are not immune to change. In 2020, the German Constitutional Court decided in response to a lawsuit of assisted suicide organizations to abrogate the general prohibition of assisted suicide. Of course, things get more complicated as soon as international standards are required.

The issues about abortion or assisted suicide have, moreover, certain characteristics that make it unclear whether they can be applied directly to artificial moral agents. The regulations set limits to the choices of individuals but they do not determine them. Yet, it is questionable whether artificial moral agents could and should have such latitudes or whether this is the privilege of full moral agents. Another important point is the difference between individual choices and laws. An individual might, for instance, decide to save a child instead of an elderly persona in a dilemma situation in autonomous driving but if politics decided to establish algorithms in autonomous vehicles by law that sacrifice elderly people in dilemma situations that seems to be a case of age discrimination.

7. Do Artificial Moral Agents Threaten Our Personal Bonds?

Another worry is that by fixing moral decisions algorithmically, one does not take into account that some situations lie beyond moral justification, as *Bernard Williams* puts it.[72] He argues that

[69] BC Stahl, 'Information, Ethics, and Computers: The Problem of Autonomous Moral Agents' (2004) 14 *Minds and Machines* 67; Van Wysberghe and Robbins, 'Critiquing the Reasons' (n 51).
[70] J Rawls, *Political Liberalism* (1993) 4.
[71] Misselhorn, 'Artificial Systems with Moral Capacity' (n 40).
[72] B Williams, 'Persons, Character, and Morality' in W Bernard, *Moral Luck* (1981) 18.

it would be 'one thought too many' if a husband, faced with the possibility of saving either his wife or a stranger, first has to think about whether it is compatible with his moral principles to give preference to his wife.[73] This is not just a matter of acting instinctively rather than on deliberation. It would be just as inappropriate for the husband to consider in advance whether he should save his wife if he were the captain of the ship and two strangers stood against his wife, or if he should save fifty strangers instead of his wife. The crucial point is that conducting these thought experiments would not be appropriate to the special relationship of mutually loving spouses. Such reasoning threatens to alienate us, according to Williams, from our personal bonds with family or friends. The problem is not just that an artificial moral agent could not make such a decision, the problem is that doing so would undermine its impartiality which was one of the main reasons why artificial moral agents might be considered as superior to human moral agents.

8. *Which Impact Does Artificial Morality Have on Ethical Theory?*

Examples like these have an impact on another issue as well. One might argue that Artificial Morality might help us to improve our moral theories. Human ethics is often fragmented and inconsistent. Creating artificial moral agents could contribute to making moral theory more consistent and unified because artificial systems can only operate on such a basis. Yet, the examples discussed raise the question whether it is good that Artificial Morality forces us to take a stance on cases that have so far not been up for decision or to which there are no clear ethical solutions as in the dilemma cases in autonomous driving.The necessity to decide such cases might, on the one hand, contribute to making our moral views more coherent and unified. On the other hand, the fact that Artificial Morality forces us to take a stance in these cases might incur guilt on us by forcing us to deliberately approve that certain people get harmed or even killed in situations like the dilemmas in autonomous driving. There may simply be cases that resist a definite final solution as Artificial Morality requires it. Some have argued that one should use algorithms that select randomly in such situations.[74] Yet, this solution contradicts the fact that in a specific dilemma situation there might well be morally preferable choices in this particular context although they cannot be generalized. What is more, a random-selecting algorithm seems to express an attitude towards human life that does not properly respect its unique value and dignity.[75]

9. *Is It Wrong to Delegate Moral Decision-Making to Artificial Moral Agents?*

There are also worries to the effect that 'outsourcing' moral decisions to machines deprives human beings of a practice that is morally essential for humanity. According to *Aristotle*, acquiring expertise in moral reasoning belongs necessarily to a human being's good life and this requires gaining moral understanding through practice.[76] Delegating moral decision-making to artificial moral agents will reduce the opportunities to exercise this capacity and will

[73] For an argument against utilitarianism in machine ethics that refers to this view, see: C Grau, 'There Is No "I" in "Robot". Robots and Utilitarianism' in M Anderson and SL Anderson, *Machine Ethics* (2011) Fn 2, 451–463.

[74] L Zhao and W Li, '"Choose for No Choose": Random-Selecting Option for the Trolley Problem in Autonomous Driving' in J Zhang and others (eds), *LISS2019* (2019).

[75] Misselhorn, *Grundfragen* (n 25) 195.

[76] Van Wysberghe and Robbins, 'Critiquing the Reasons' 731 (n 51) 731.

lead to a 'de-skilling' of humans with respect to morality.[77] One might rise to this challenge by pointing out that there are still many opportunities for humans to exercise and develop their moral skills.[78]

Yet, there might be a deeper concern that this answer does not address. For *Kant*, being able to act morally is the source of our normative claims towards others. One might interpret this claim as saying that morality is a reciprocal practice between full moral agents that are autonomous in the sense of setting themselves ends and that are able to reason with each other in a distinctly second-personal way.[79] Functional moral agents cannot really take part in such a practice, and one might argue that delegating moral decisions to them violates this moral practice independently of the quantitative question of how often this is done. This is one of the reasons why creating a Kantian artificial moral agent might be contradictory.[80]

10. *Who Is Responsible for the Decisions of Artificial Moral Agents?*

Finally, there is the concern that Artificial Morality might undermine our current practice of responsibility ascription. As was argued above, delegating morally relevant decisions to artificial systems might create so-called responsibility gaps. *Robert Sparrow* who coined this term uses the example of lethal autonomous weapon systems to argue that a responsibility gap arises if such a system violates the ethical or legal norms of warfare and the following conditions are fulfilled: (1) the system was not intentionally programmed to violate the ethical or legal norms of warfare; (2) it was not foreseeable that the use of the lethal autonomous weapon system would lead to this result; and (3) there was no human control over the machine from the start of the operation.[81]

The problem is that if these three conditions are fulfilled, then moral responsibility cannot be attributed to any human when the machine kills humans in conflict with the moral or legal norms of warfare, because no human being had intended it, it was not foreseeable, and nobody had the possibility to prevent the result. Thus, a responsibility gap occurs precisely when the machine itself is not responsible, but its use undermines the terms of attributing responsibility to human beings. For *Sparrow*, this is a reason for rejecting the use of war robots as immoral because, at least when it comes to killing humans, there should always be someone who can be held responsible for the deaths.

VII. CONCLUSION: GUIDELINES FOR MACHINE ETHICS

Which conclusions should we draw from the controversy about artificial moral agents? One suggestion is to place a moratorium on the commercialization of artificial moral agents. The idea is to allow academic research on Artificial Morality while at the same time protecting users, other concerned persons or groups, and society 'from exposure to this technology which poses an existential challenge'.[82] This seems to be at least reasonable as long as we do not have good answers to the challenges and critical questions discussed in the last section.

[77] S Vallor, 'Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character' (2015) 28 *Philosophy & Technology* 107.
[78] Formosa and Ryan, 'Making Moral Machines' (n 51).
[79] S Darwall, 'Kant on Respect, Dignity, and the Duty of Respect' in M Betzler (ed), *Kant's Ethics of Virtues* (2008).
[80] R Tonkens, 'A Challenge for Machine Ethics' (2009) 19 *Minds and Machines* 421.
[81] R Sparrow, 'Killer Robots' in 24 *Journal of Applied Philosophy* 62.
[82] Van Wysberghe and Robbins, 'Critiquing the Reasons' (n 51) 732.

There are, however, some loopholes that this suggestion does not address. A device like an autonomous car might, as a matter of fact, be designed as an artificial moral agent without being commercialized as such. This is possible because algorithms are often trade secrets. Another challenge is that moral decisions do not always have to be taken explicitly but might be hidden behind other parameters. An algorithm for autonomous driving might, for instance, give priority to the passengers' safety by using certain technical parameters without making it explicit that this puts the risk on more vulnerable traffic participants.

The controversy about artificial moral agents does, however, not necessarily have to be seen as formulating impediments to research and innovation. The arguments might also be regarded as indicators for the directions that research on the design of artificial moral agents and their development should take. The lessons that have to be drawn from the controversy can be condensed in three fundamental guidelines for machine ethics:[83]

(1) Moral machines should promote human autonomy and not interfere with it.
(2) Artificial systems should not make decisions about life and death of humans.
(3) It must be ensured that humans always take responsibility in a substantial sense.

In the light of these three guidelines for machine ethics, there are some areas of application of artificial moral agents that should be viewed critically from a moral point of view. This applies in particular to killer robots, but autonomous driving should also be considered carefully against this background. There is reason to assume that, in order to optimize accident outcomes, it is necessary to specify cost functions that determine who will be injured and killed which bear some similarity to lethal autonomous weapon systems. Legitimate targets would have to be defined for the case of an unavoidable collision, which would then be intentionally injured or even killed.[84] As long as the controversial issues are not resolved, robots should not get a license to kill.[85]

Even if one does not want to hand over decisions about the life and death of human beings to artificial moral agents, there remain areas of application in which they might be usefully employed. One suggestion is a conceptual design of a software module for elder care that can adapt to the user's individual moral value profile through training and permanent interaction and that can, therefore, treat people according to their individual moral value profile.[86] Under the conditions of reasonable pluralism, it can be assumed that users' values with respect to care differ, for example, as to whether more weight should be given to privacy or to avoiding health risks. A care system should be able to weigh these values according to the moral standards of each individual user. In this case, a care system can help people who wish to do so to live longer in their own homes.

Such a system could be compared to an extended moral arm or prosthesis of the users. One could also speak of a moral avatar which might strengthen the care-dependent persons' self-esteem by helping them to live according to their own moral standards. Yet, such a system is only suitable for people who are cognitively capable of making basic decisions about their lives

[83] These guidelines must be understood as addressing specifically the arguments from the controversy. There are other principles of machine ethics, for instance, that the decisions of artificial moral agents should be fair. Such principles arise from general ethical considerations which are not specific to machine ethics.

[84] P Lin, 'Why Ethics Matters for Autonomous Cars' in M Maurer and others (eds), *Autonomous Driving: Technical, Legal and Social Aspects* (2007).

[85] C Misselhorn, 'Lizenz zum Töten für Roboter? "Terror" und das autonome Fahren' in B Schmidt (ed), *Terror: Das Recht braucht eine Bühne. Essays, Hintergründe, Analysen* (2020).

[86] Misselhorn, *Grundfragen* (n 25).

but are so physically limited that they cannot live alone at home without care. It should also be clear that there is no purely technological solution to the shortage of care givers. It is essential to embed these technologies in a social framework. No one should be cared for by robots against their will. The use of care systems must also not lead to loneliness and social isolation among those receiving care.

A very demanding task is to make sure that humans always take responsibility in a substantial sense as the third principle demands. In military contexts, a distinction is made between in-the-loop systems, on-the-loop systems, and out-of-the-loop systems, depending on the role of the human in the control loop.[87] In the case of in-the-loop systems, a human operates the system and makes all relevant decisions, even if it is by remote control. On-the-loop systems are programmed and can operate in real time independent of human intervention. However, the human is still responsible for monitoring the system and can intervene at any time. Out-of-the-loop systems work like on-the-loop systems, but there is no longer any possibility of human control or intervention.

The problem of the responsibility gap appears to be solved if the human remains at least on-the-loop and perhaps even has to agree to take responsibility by pressing a button before putting an artificial system into operation.[88] But how realistic is the assumption that humans are capable of permanent monitoring? Can they maintain attention for that long, and are they ready to decide and intervene in seconds when it matters? If this is not the case, predictability and control would be theoretically possible, but not feasible for humans in reality.

Second, there arise epistemological problems, if the human operators depend on the information provided by the system to analyze the situation. The question is whether the users can even rationally doubt its decisions if they do not have access to independent information. In addition, such a system must go through a series of quality assurance processes during its development. This may also be a reason for users to consider the system's suggestions as superior to their own doubts. Hence, the problem of the responsibility gap also threatens on-the-loop systems and it may even occur when humans remain in-the-loop.[89]

Overall, it seems unfair that the users should assume full responsibility at the push of a button, because at least part of the responsibility, if not the main part, should go to the programmers, whose algorithms are decisive for the system's actions. The users are only responsible in a weaker sense because they did not prevent the system from acting. A suitable approach must take into account the distribution of responsibility which does not make it easier to come to terms with the responsibility gap. One of the greatest challenges of machine ethics is, therefore, to define a concept of meaningful control and to find ways for humans to assume responsibility for the actions of artificial moral agents in a substantial sense.

[87] United States Department of Defense Unmanned Systems Integrated Roadmap FY 2011-2036. Reference Number 11-S-3613. https://irp.fas.org/program/collect/usroadmap2011.pdf

[88] Such a suggestion is, for instance, made by Arkin, *Governing* (n 64).

[89] A Matthias, 'The Responsibility Gap – Ascribing Responsibility for the Actions of Learning Automata' (2004) 6(3) *Ethics and Information Technology* 175–183.