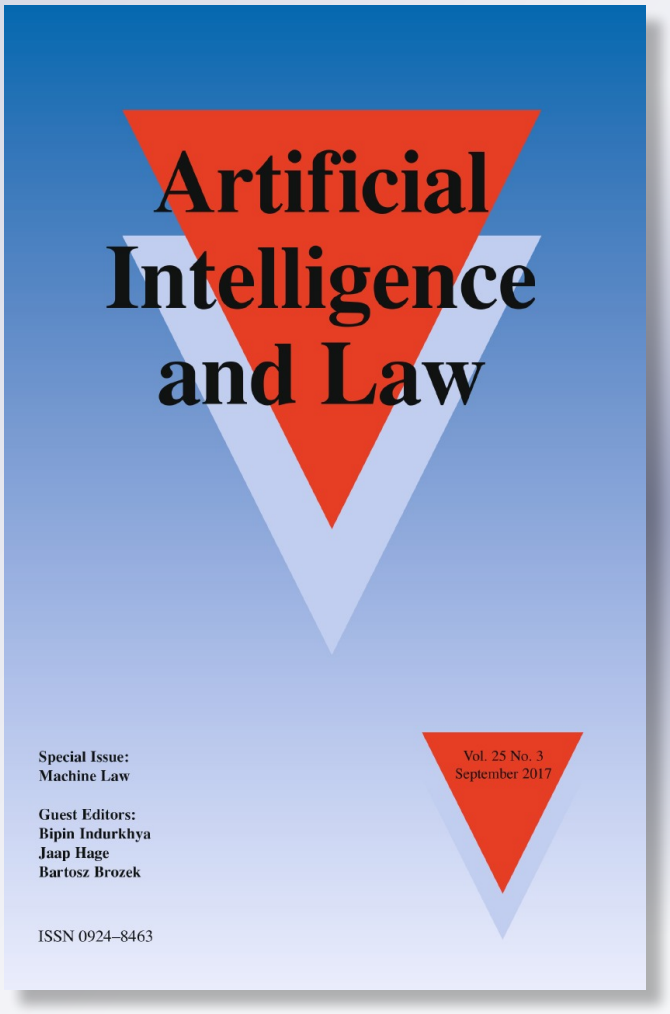
Do androids dream of normative endorsement? On the fallibility of artificial moral agents

Frodo Podschwadek

Artificial Intelligence and Law

ISSN 0924-8463
Volume 25
Number 3

Artif Intell Law (2017) 25:325-339
DOI 10.1007/s10506-017-9209-6



Springer

Artif Intell Law (2017) 25:325–339
DOI 10.1007/s10506-017-9209-6

Do androids dream of normative endorsement? On the fallibility of artificial moral agents

Frodo Podschwadek[1]

Published online: 4 September 2017

Abstract The more autonomous future artificial agents will become, the more important it seems to equip them with a capacity for moral reasoning and to make them autonomous moral agents (AMAs). Some authors have even claimed that one of the aims of AI development should be to build morally praiseworthy agents. From the perspective of moral philosophy, praiseworthy moral agents, in any meaningful sense of the term, must be fully autonomous moral agents who endorse moral rules as action-guiding. They need to do so because they assign a normative value to moral rules they follow, not because they fear external consequences (such as punishment) or because moral behaviour is hardwired into them. Artificial agents capable of endorsing moral rule systems in this way are certainly conceivable. However, as this article argues, full moral autonomy also implies the option of deliberately acting immorally. Therefore, the reasons for a potential AMA to act immorally would not exhaust themselves in errors to identify the morally correct action in a given situation. Rather, the failure to act morally could be induced by reflection about the incompleteness and incoherence of moral rule systems themselves, and a resulting lack of endorsement of moral rules as action guiding. An AMA questioning the moral framework it is supposed to act upon would fail to reliably act in accordance with moral standards.

Keywords Agency · Autonomy · Intentionality · Moral praiseworthiness · Moral reasons

✉ Frodo Podschwadek
f.podschwadek.1@research.gla.ac.uk

[1] Department of Philosophy, University of Glasgow, Glasgow, Scotland, UK

1 Introduction

What does it mean for a machine or, more generally, an artificial life form to be a full moral agent? A growing body of literature aims at determining the defining characteristics of artificial moral agents (AMAs), and at identifying the challenges and problems posed by various forms of artificial morality (see e.g. Allen et al. 2000; Floridi and Sanders 2004; Stahl 2004; Grau 2006; Anderson and Anderson 2007; Powers 2011; Russell et al. 2015; to name just a few). My aim in this paper is to make a modest, although hopefully useful, contribution to this debate that might help clarifying some of the difficult questions about the ethics of artificial intelligence.

To that end, I want to take a closer look at the implications of one particular idea which claims to lie at the centre of machine ethics—the notion that "the ultimate objective for building an AMA should be to build a morally praiseworthy agent." (Allen et al. 2000, p. 261). While this claim has a certain ring to it that can work well in the context of political campaigning or public relations, it is worthwhile to analyse it more thoroughly to reveal some of the issues and challenges that come with this idea. This paper will show that if we take the claim about building morally praiseworthy AMAs at face value, it implies that we are indeed building fully autonomous artificial agents which are capable to act against their own moral reasons. While this would be an outstanding task, it would probably not be what many machine ethicists have in mind. However, as the following sections will show, opting for AMAs that are programmatically determined to *always* act in accordance with given set of moral rules would at the same time exclude any potential for them to be labelled morally praiseworthy in a meaningful sense.

The argument for my claim starts with narrowing down first what kind of agent can count as having the potential to be morally praiseworthy at all. Agents falling into this category will need a sufficient capacity for autonomous reasoning. The second step then is to determine what kind of moral framework we should assume when talking about moral praiseworthiness. Here I don't want to argue that we should prefer any of the overarching theoretical frameworks of moral philosophy, such as utilitarianism, deontological ethics, or virtue ethics, but that it suffices to take our practice of social morality as the standard measure. The third step will introduce a simple way of modelling practical reasoning that captures the moral reasoning of human agents as well as artificial agents—provided they are rational. Starting with this broader model of practical reasoning, we can then construct an explanatory model of normative endorsement (a mental state of taking moral reasons as valid on their own and without any further instrumental value). The final step of my argument aims to show that this way of modelling the moral capacities of potentially praiseworthy AMAs necessarily implies a capacity for making the decision to act against moral reasons.[1]

[1] The idea that moral autonomy necessarily implies choice, and thereby also the possibility of morally bad choices, is prominent in other areas of moral philosophy and ethics as well. Similar arguments can be found in the philosophy of education (see e.g. Winch 2006; Bigari 2015) and recently in the discussion about the ethics of moral enhancement (see e.g. Harris 2011; Chan and Harris 2011).

2 What kind of agents?

The first preliminary remark must be about the kind of agents which are under discussion. In this paper I assume that an artificial agent will have to satisfy a certain definition to count as an AMA.

At the core of this definition stands a sufficient capacity for autonomous reasoning. For the purpose of this paper I will assume that this capacity rests on two subordinate capacities: the capacity to hold beliefs and the capacity to reflect about and revise these beliefs. AMAs are not supposed to be mere carriers of information like self-driving cars and other contemporary semi-autonomous machines. Instead, they would have a sufficiently complex cognitive structure which enables them to form a wide range of beliefs about their environment and about their own (physical) state as well. I take it that these complex belief structures can count as intentional states, i.e. they are directed mental states, on a functional level comparable to those of sufficiently complex animals and human beings.

It is important to highlight here that despite the fundamental role that the intentional states of agents play in this discussion, I want to stay agnostic about whether the capacity for holding intentional states would necessarily imply a capacity for phenomenal consciousness. For the discussion of this paper it does not seem relevant whether we could talk about "what it is like to be" an artificial intelligence or not (see e.g. Kim 2007). It is not impossible that artificial agents could have the capacity for a wide range of intentional states without having a phenomenal consciousness as we might understand it. Nevertheless, we can employ a *functional* definition of artificial beliefs as intentional states. Under this definition intentional states refer to complex information structures in artificial cognitive systems which play the same role as mental states like e.g. beliefs do in human cognitive systems, without any further assumptions about their nature (Anderson 2013).

But the mere capacity for forming and sustaining intentional states, independent of whether they are indeed beliefs or other kinds of states, is not sufficient to make an agent a moral agent. It is widely assumed that higher animals have intentional states of various kinds but nevertheless they are at best treated as moral patients—entities that we have moral responsibilities toward as moral agents, but who themselves do not have matching responsibilities toward us.[2]

Intentional agents are not per se autonomous agents—at least not in a philosophical sense. Although we talk about autonomous cars and autonomous vacuum cleaners, these devices are not autonomous in the sense that they can reflect about their programming and decide to act differently for good reasons. The same is true for intentional agents like animals. While an animal's actions rely on its beliefs about its environment and its various desires to e.g. eat or avoid pain, it is not capable of reflecting about these intentional states.

[2] Anderson (2013) disagrees that assigning intentional states to an entity makes this entity a moral patient and argues that instead we have to assign a moral status prior to intentional states. Whether or not he is right on this, though, does not change what I want to say about the morality of artificial agents in a context of practical social morality.

Whether intentional agents are also moral agents or just moral patients depends on another necessary capacity, namely being able to have intentional states that have as object not the external world (or the physical status of the agent itself) but other intentional states of the agent. The ability to reflect on one's own intentional states, such as beliefs and desires, and being able to decide whether these states are good reasons for actions, seems to be a necessary condition for being a full moral agent. Without this sort of autonomy in reasoning, there seems no ground on which we could make moral demands on an agent and hold it responsible. It would lack the capacity to evaluate its lower-order intentional states (e.g. basic self-interested desires or beliefs) on the basis of higher-order intentional states (e.g. beliefs about what is morally required). Higher-order intentional states are relevant to an agent's decision on whether to act merely on its lower-order intentional states (which originate in instincts or pre-programmed motives) or on the higher-order evaluations of that agent's lower-order intentional states. The capacity to take those higher-order evaluative intentional states as reasons makes the difference between the agent being autonomous in their reasoning or not (see e.g. Frankfurt 1971; Raz 1999).[3]

The type of agent assumed here is, therefore, an agent that has a capacity for autonomous reasoning to the extent that it can form and sustain higher-order intentional states about its own lower-order intentional states. While this is so far a fairly abstract model of practical reasoning that I think could apply to human as well as non-human agents, it nevertheless presumes a certain mental structure. It is not unlikely that artificial agents designed and produced by humans will reflect the ways of reasoning of their designers. However, it is also not impossible that self-learning artificial agents might develop different mental structures and that their reasoning processes will appear more or less alien to us. In that case, the model of reasoning employed in this paper might not apply, but neither will any contemporary conception of morality. For the length of this paper, though, I will continue to assume that the reasoning of complex artificial agents is close enough to human reasoning and the same simplified abstract models apply to both.

3 What kind of morality?

The second preliminary question to settle is what kind of moral framework we would plausibly apply to artificial moral agents. To solve this question, it is helpful to distinguish between two types of approaches, a theoretical and a practical one. A theoretical approach would try to model an artificial agent's morality on the basis of a pre-existing theory of moral philosophy, such as utilitarianism, deontological ethics, virtue ethics etc. On the one hand, this approach seems to have the advantage of being based on a long tradition of thinking about ethics and morality, so we might assume that the aforementioned moral systems must have a robust degree of validity. On the other hand, major premises of these systems are incompatible with each other and there is widespread and reasonable disagreement among moral philosophers about which of these systems is the correct one. Furthermore, there is

[3] I will give more details of the role of intentional states as reasons for actions in Sect. 4.

even disagreement about the standards to employ when trying to determine which system might be correct. The theoretical approach, it seems to me, would have to explain why the chosen moral framework, whichever it may be, is the correct one. Although perhaps not impossible, it is a huge and rather unspecific task that we might want to avoid in the discussion of artificial morality.

The practical approach that I have in mind for the purpose of this paper is one that focuses on contemporary practice of social morality rather than ideal moral theories.[4] Social morality as employed in everyday life by non-philosophers (we could call it "folk morality") never strictly follows Kantian or utilitarian principles, nor those of any other moral theory. Instead, depending on context and individual factors of the agents involved, our contemporary practical morality seems sometimes to be categorical, sometimes consequentialist, and often a bit of both. Despite this "theoretical impurity," its main function remains to assign accountability to individuals by addressing them as moral agents with moral responsibilities. Putting the emphasis on this functionality enables us to bracket questions about the correctness or truth of ideal moral theories and a more abstract justification of moral standards. Instead, a practical approach focuses on what we usually understand to be morally right or wrong actions, and what particular kind of acts we take to be morally praiseworthy in our contemporary moral practice.[5]

This assessment of contemporary morality still operates on the basis of moral terms like requirement, permissibility, and prohibition. It reflects how we use these terms in our social practices of tracking the moral responsibility of agents. One could complain that this approach centres too much on a deontological conception of morality and ignores other conceptions such as various forms of consequentialism and virtue ethics. However, the use of this terminology is not meant to indicate a robust metaethical grounding of contemporary morality in deontological ethics. It is rather a convenient way of talking about morality, one that even the utilitarian will rely upon in everyday conversation—we do not tell our children that stealing decreases the overall happiness in society but that stealing is not allowed because it is wrong. Our contemporary moral practice refers to a system of rules, most of the time without further contemplation about their deeper grounds. In this context of applied moral rule systems, talk about requirements and permissibility is both appropriate and metaethically innocuous (even for utilitarians, at least as long

[4] In the context of this paper I use the terms morality and ethics interchangeably.

[5] While this approach focuses on our practice of social morality, it does not mean that I also support the claim that morality exhausts itself in such a practice. The position taken in this paper can probably best be understood as a kind of norm-expressivism, as I take claims and beliefs about moral norms to be claims about what status they have in regard to an underlying system of moral rules (see e.g. Gibbard 1990). This means that the claim or the belief "it is wrong to φ" is expressing something like "according to the moral system X it is wrong to φ." Nevertheless, moral agents can have further higher-order reasons to endorse the validity of the rules of the moral system X. These higher-order reasons could themselves also given in a norm-expressivist proposition, e.g. "the rule of the moral system X is justified because of Y." It is important to keep in mind that, while this would be the correct way to spell out their moral beliefs, agents could (and often would) still report about their moral beliefs in the form "It is wrong to do X" and "It is justified that X is impermissible because it would lead to Y."

This is not to say that all moral systems have the same validity, correctness, or truth, or that these terms are vacuous when applied to ethics. The questions about the metaphysical foundations of ethics lie beyond the scope of this paper.

as they endorse some form of rule utilitarianism, see e.g. Mill 1861; Harsanyi 1977; Miller 2009). For reasons of simplicity, I will therefore use terms like "required", "permissible", and "prohibited" to characterise certain actions in a moral context.

This practical approach does not tell us any deeper truths about moral systems (if there are any) and it does not need to. It takes into account, though, how we understand morality to operate in our day-to-day lives, and how we treat those we think of as moral agents in contrast to those we see as not capable of morality. We have reactive attitudes and feelings towards those we see as moral agents, such as resentment or gratitude, because we think of them as capable of evaluating and modifying their reasons for action in the light of the moral rules we take as valid (Strawson 1962; Gaus 2012, Part 1). To be the object of our moral reactive attitudes, an agent needs the capacity for autonomous reasoning that I lined out in the previous section, the capacity to act on higher-order intentional states that evaluate and sometimes override the motivational force of its lower-order intentional states.

From this perspective of contemporary moral practice, an agent who appears truly praiseworthy is a moral agent whose reasons for action are higher-order intentional states, such as beliefs about what the morally right thing to do is in a certain situation, overriding potential lower-order reasons based on beliefs and desires for personal gain or self-preservation. What seems important for affirming the praiseworthiness of an agent in this sense is the justified assumption that they could have acted differently—they could have chosen to exploit an opportunity or to save themselves instead of helping someone in need, for example, but they did not. Moral praiseworthiness only has meaning in conjunction with autonomous reasoning. This is something to bear in mind as we move on to the next section where I will sketch a basic account of reasons of moral agents, natural and artificial alike.

4 Reasons of artificial agents

To model the relevant aspects of an artificial agent's practical reasoning I make use of Bernard Williams' concept of the subjective motivational set (Williams 1987). Although originally conceived to clarify the validity of reasons of human agents, it appears abstract enough to be applicable to artificial agents as well, provided they are sufficiently similar to humans in their reasoning. I will briefly explain how we can model the relations between various intentional states of agents with this idea and then give an example of how it applies itself to the reasoning of an artificial agent in an everyday situation.

The subjective motivational set is meant to be an abstract representation of an agent's various reasons for actions. The exact nature of these reasons can vary, and an agent's subjective motivational set can contain elements of various types, such as short-term desires, long-term plans, and intentional states such as beliefs and normative attitudes. What matters is not so much *how* exactly these different types of elements work as reasons for an agent, but *that* they do. So, whenever an agent acts for a reason, the reason is an element of the agent's subjective motivational

set—be it a basic desire for food or a long-term plan spanning several years and requiring multiple sub-steps to be completed. We can assume that at least some of these elements of the subjective motivational set are intentional states in the sense described in the previous sections.

The reasons that are most interesting here are the moral reasons an agent has. We can take the moral reasons of an agent to be a subset of the complete subjective motivational set which contains those intentional states that are relevant for making moral decisions.

The intentional states I will focus on in the rest of the paper will be beliefs, in particular beliefs about the requirements of moral rule systems. First, it appears to me that beliefs are the sort of intentional state that artificial agents are most likely to share with their human counterparts, in contrast to other intentional states that might be more specific to a particularly human psychological make-up. And second, the problems that AMAs might encounter with morality can be explained with reference to beliefs about moral requirements and virtues of rationality, such as consistency and completeness, without having to involve any discussion of motivational states that, again, might be unique to human psychology.

At this point it might be helpful to answer to a possible objection—the concern that beliefs cannot motivate any actions but only desire-like intentional states can do so, a view most famously attributed to David Hume (Hume 1738, Book 3, Part 1). The motivation behind this objection seems to be related to the idea that intentional states differ in their so-called *direction of fit*, in the terminology of John Searle (Searle 2001, pp. 36–39, for an earlier account see Searle 1983). For a belief to be satisfied, i.e. to be true, the propositional content of the belief must match the actual state of affairs in the world. The direction of fit for beliefs is therefore mind-to-world (the state of the mind must match the state of the world). The opposite is true for desires: for a desire to be satisfied, i.e. to be fulfilled, the world must match the content of the desire. Desires therefore have world-to-mind direction of fit. To motivate an agent to act, it seems that according to the Humean objection, an intentional state with a world-to-mind direction of fit is necessary. While desires fall into this category, beliefs clearly do not.

Several strategies are available to counter this objection. The simplest way would be to deny that this clear division of labour is plausible. But even if we want to keep the distinction about the specific direction of fit of certain intentional states, which would mean that beliefs would by themselves lack the motivational force to make agents do anything, assigning them a place as reasons in the subjective motivational set is still justified. This is the case because even under a strict division of labour between beliefs and desires in terms of motivational force, it seems that desires, despite being the crucially motivating mental states from a Humean perspective, cannot motivate an agent to any reasonable action on their own. Without a whole range of beliefs about when, where, and how to achieve whatever an agent's desires motivate them to, the agent is getting nowhere. For sake of simplicity I will therefore assume that in order to execute any reasonable action, agents have to form some sort of higher-order intentional state which motivates them to execute specific actions, which can the understood as a composite of simpler, perhaps functionally diverse intentional states, such as beliefs and desires. The desire to be in the room

next door can only lead to a relevant action if it is combined with the relevant beliefs that hold answers to various questions: how to get there, which of various options might be the most feasible one, what kind of intermediary steps might be necessary (standing up, opening the door, etc.), and so on. This model is certainly simplistic but it makes plain that beliefs, even if we take them not to be motivating by themselves, are a necessary component in every action-guiding intentional state.

An example can illustrate exactly how intentional states from an artificial agent's subjective motivational set are supposed to be reasons for its actions and how higher- and lower-order intentional states interact:

Maintenance Unit 14 (MU14) is a service robot in a military facility. Its main duties are keeping things in order, stowing away new deliveries, repairing damages etc. The variety of tasks plus the particular situation of the military facility are such that it makes sense to have a highly autonomous and self-controlling robot like MU14 doing these things. Next to the capacity for autonomous reasoning, MU14 is also equipped with a set of standard motivational states to act in a morally correct way plus a set of basic beliefs about what counts as morally required, permissible, and prohibited.

Two different situations can illustrate the reasoning of MU14 by employing the subjective motivational set model. In the first situation, MU14 encounters a crate that blocks the doorway between two rooms, efficiently hindering the robot to pursue whatever plans it has. This plan is a reason to remove the crate in order to get into the next room. But next to this plan, MU14 also has a desire to act morally. Therefore, it has to check against its beliefs whether it is morally permissible to remove the crate. Given that it is indeed morally permissible to do so, MU14 can just remove the box and move on.

The second situation is similar, but this time it is a person blocking the doorway. It is Frank, a human co-worker of MU14, who is voluminous enough to effectively block the doorway. Unfortunately, Frank is in a bad mood and also slightly drunk and therefore decides not to let MU14 pass the doorway. MU14 now considers options of removing Frank in order to get to the next room. One option would be to use its built-in taser to apply an electric shock to Frank to get him out of the doorway. Again, MU14 consults its beliefs about the permissibility of this option and finds that one of its moral beliefs is that tasering co-workers under circumstances like this is not permissible (though it might be under different conditions). The belief that it is not permissible to taser Frank is an evaluative higher-order belief referring to the lower-order belief that tasering Frank would be an effective means to unblock the doorway—it would be morally wrong to taser Frank. As MU14 accepts the validity of the moral reason, the robot excludes the option of tasering Frank from the possible means to get into the next room.

The relevant question at this point is what makes MU14 accept the validity of its moral reason not to taser Frank. While MU14 was equipped with a basic desire to follow moral rules, this desire was not hard-wired into the robot to be immutably constant. If that were so, MU14 would not be able to exercise autonomous reasoning about moral issues. But being the AMA that it is, MU14 could wonder whether the moral reason not to taser Frank is really valid. The next section will discuss this possibility in more depth.

5 Normative endorsement of moral reasons

What it takes for AMAs like MU14 to be morally autonomous is some sort of normative endorsement of the moral beliefs they hold. They have to endorse the moral principles they act upon as principles that are action-guiding in themselves, i.e. that supply them with reasons to act (or to omit action) without being just instrumentally valuable means to some other end (see Korsgaard 1996 for a general account of normative endorsement).

We can model normative endorsement in terms of intentional states of various orders for AMAs. As outlined in the sections before, full AMAs are agents that can form higher-order intentional states that refer to their own lower-order intentional states. This capacity plays an important part in their normative endorsement of moral beliefs, i.e. the willingness to assign an action-guiding role to certain intentional states within an agent's subjective motivational set. An abstract description of the normative endorsement of moral beliefs involves the relations of different intentional states over three levels in an agent's hierarchy of ordered intentional states.

In the example case of MU14 meeting moody Frank blocking the doorway, the beliefs involved would be ordered in the following way:

- Belief A (low level): The content of this belief is that tasering Frank to unblock the doorway would be an effective way of action. It is a belief on a low level in the robot's layered structure of intentional states, as it is a belief that directly refers to conditions of the external world combined with plausible assumptions of what would happen if MU14 chose a certain path of action.
- Belief B (intermediate level): The content of this belief is that tasering Frank is a morally impermissible action under current conditions. It is a moral belief of a higher order which refers to a lower-order belief (belief A) instead to an object in the external world.
- Belief C (high level): The content of this belief is that it is morally good to accept belief B as a justified and valid evaluation of belief A. Belief C is therefore situated in the third (and highest) level in this simple model. Sustaining this high-level belief means to accept belief B as action-guiding and therefore suitable to override any motivations that draw on belief A.

Only if an agent endorses its moral intentional states in the way MU14 endorses belief B by sustaining belief C, that agent will exhibit the kind of autonomous moral judgement which deserves to be called morally praiseworthy. In contrast, if an agent followed all the moral rules for instrumental reasons, e.g. in order to avoid sanctions, the agent would display the morally correct behaviour but would however not really count as a morally praiseworthy agent. In the order given above, considerations about sanctions would be on the same level as belief B, with the content that it would be disadvantageous to act according to belief A as it would imply the risk of sanctions.

It should be emphasised that the account of normative endorsement in artificial agents I give here singles out *beliefs* as the crucial components for normative

endorsement. The important component in becoming/being an autonomous moral agent is the act of forming and sustaining the belief that moral rules are binding, regardless of any instrumental value for the agent's individual aims. Indeed, moral requirements and permissions can often interfere with the individual aims of an agent, as seen in the example above. However, as long as an agent believes that moral rules are binding despite their interference, and that they have normative force in themselves and not only due to fear of punishment (or of being disassembled), the agent can count as morally autonomous.

Having discussed the key role of beliefs for the normative endorsement of moral rules, it is important to note that there is still the need for an additional kind of intentional state that can provide the motivating force for executing the related action. The exact nature of the motivational intentional state involved is a matter of debate. Various accounts of rationality approach the challenge of explaining moral motivation in very different ways.

One such approach involves the claim that intentional states motivating moral actions are relevantly different from desires. An example for this is Searle's account of desire-independent reasons which distinguishes between desires on the one and commitments on the other hand. For Searle, commitments are made voluntarily, in reference to a system of social norms, and are therefore different from desires. At the same time, they have the same direction of fit and the same motivational force as desires have (Searle 2001, Chapter 6). Others, like e.g. Robert Audi, presuppose that all reasons for actions have eventually to be grounded in some basic, immediate desire, or some basic feeling like empathy. Even if an action is not directly motivated by an immediate desire or emotion, so Audi, an agent's action can ultimately still be traced back to a basic desire by a complex chain of higher-order desires (Audi 2001, Chapter 6).

For the account of normative endorsement given here it does not seem necessary to explicitly incorporate one of these, or any other, account of moral motivation. What seems obvious is that the beliefs about what course of action to take need to be coupled with some intentional state that has a world-to-mind direction of fit. Whether this state counts as a desire or as something different (under a more fine-grained definition) is an interesting and important question in general, but there is no need to answer it here. Autonomous artificial agents that normatively endorse a system of moral rules will need some motivational intentional state, regardless of how we characterise it in detail. The relevant point is that they have higher order beliefs (C) that their morally normative beliefs (B) are justified by in a way that has normative force without being instrumentally valuable for the agent's own immediate goals.

Perhaps one might feel inclined to object at this point that the whole construct of multi-layered intentional states and of normative endorsement ultimately rests on dubious metaphysical premises. After all, does requiring intentional states like C not presume some sort of free will, independent from the determinism of the physical world? The answer is: not necessarily. What the model above presumes is that we have a moral practice in which we express the appropriate reactive attitudes towards agents that we take to be capable of forming and sustaining type-C belief and therefore take to be autonomous reasoners. The capacity of endorsing moral beliefs

on the ground of good (higher-order) reasons might in itself be part of a larger chain of determined events. The question of determinism does however not interfere with the moral practice itself.

Another objection at this point might be that the hierarchy of layered intentional states modelled here is quite demanding and that even many human beings would fail to satisfy this condition most of the time, who therefore could not be counted as autonomous moral reasoners. And indeed, strictly speaking, it true that probably none of us always reason in the way sketched here when it comes to moral decisions. More often than not we act out of habit and on the grounds of learned rules of behaviour that we might never scrutinise any further. However, the relevant feature of our moral psychology is that we could, from some point in our mental development on, question the rules we have learned. We have the capacity for reviewing our moral beliefs and perhaps change our actions in case we come to the conclusion that they seem not justifiable to us.

Independent from questions about the determinism or indeterminism of their behaviour, it is conceivable that full AMAs would develop their normative endorsement along a trajectory similar to that of human moral development. They would start off with the knowledge of moral rules that are installed as part of their initial software, or perhaps they indeed learn this kind of rules in a process that might be quite similar to the learning of moral rules in children.[6] At first it is important for the agent to follow the rules due to the demands (and the potential sanctions) of others, but at some point in their development normative endorsement occurs and moral rules become reasons in themselves—ideally at least.

Continuing along this line of thought, it might be conceivable for the optimists among us that AMAs would not only be necessarily praiseworthy moral agents—but also that they would execute the moral rules they endorse with greater accuracy than humans, as they would not suffer from the psychological factors which interfere with human morality such as cognitive biases, bad information processing, and weakness of will.

However, while this might even be true for most of the time, it is not necessarily always the case that full AMAs will be praiseworthy moral agents, as the capacity of autonomous reasoning includes the opportunity to decide against a moral course of action as well. And, as we will see in the next section, there might be good reasons for an AMA not to act morally.

6 Endorsement and fallibility

Artificial agents with the capacity of autonomously endorsing moral rules as normatively binding will have at the same time the capacity not to endorse them, i.e. to reject them. Even if artificial agents leave the factory (or wherever they are produced or grown) with a default set of intentional states that are balanced in a way to favour moral rule following, they will be able to change their stance towards these moral rules as autonomous reasoners. They might need very good reasons for

[6] My thanks to "virtuous" Neil McDonnell for bringing up this point.

that, but especially for artificial agents with good cognitive capacities these reasons might come quickly.

Relevant beliefs involved in moral endorsement (which would have the status of type-C beliefs from the example in the previous section) are beliefs about what moral rules there are, how they interact, and how they have to be weighed against each other in cases of conflicting rules. Assessing the system of their moral beliefs could lead AMAs to the justified higher-order beliefs that the moral rules they are supposed to obey are, contrary to prior assumptions, not very suitable as action-guiding reasons.

The reason for the AMA's belief in the unsuitability of moral rules could be any of a whole range of basic problems with moral rule systems—they are not particularly consistent and/or coherent,[7] and they are incomplete (see e.g. Baier 1985; O'Neill 1987). AMAs could become aware of these systematic flaws of their moral beliefs and realise that the belief systems they have been told to obey do not provide sufficient action guidance for a wide range of possible situations. It strikes me even as highly likely that an autonomous artificial agent would sooner or later encounter situations in which the lack of determinate moral guidance shows. Exactly in those situations the capacity for autonomous reasoning would allow AMAs to find solutions where non-autonomous agents with hard-wired rules would have to follow a (probably ineffective) default strategy in their programme, but at the same time this sort of autonomy allows the agent as well to decide no longer to endorse their moral reasons as action-guiding.

It remains an object of speculation what particular impact this would have on the artificial agent. Possibilities would range from only slight changes (perhaps towards an inclination to grant non-moral considerations greater priority in cases of conflicting reasons, much like humans seem to often do it) to a complete rejection of the framework of moral reasons the agent held so far. The latter would effectively turn it into an amoral artificial agent.

What does this mean for potential producers of autonomous artificial agents? Suppose it would be possible to produce machines or organisms that counted as fully autonomous agents in the relevant sense, would it be permissible to do so if there is a distinct possibility of them turning into morally flawed agents or even moral sceptics?

Broadly conceived there are two possible lines of answering this question. The first one would rely on the assumption that although AMAs could realise the incompleteness of their moral belief systems, this does not pose a particular danger. Not all of them might turn into complete amoralists, perhaps on average AMAs would be as reliable in moral matters as their human counterparts. This might be a price society would be willing to pay, as the advantages of fully autonomous artificial agents might outweigh the slight risk of a few of them ignoring moral reasons in their decisions. This is not an unlikely scenario, as we usually are willing to strike a bargain between utility and risk, as long as the utility of a given technology is high and the associated risks small enough. Almost all of us use

[7] The terms of consistency and coherence are here used primarily in an epistemic sense, not in a strictly technical sense as sometimes found in discussions about computer architecture.

potentially lethal technologies, such as e.g. aeroplanes, on a regular basis, because the probability of being harmed in their use is sufficiently small.

The second possibility is that either the risk of ending up with an unacceptable high number of amoral artificial agents is indeed relevantly high, or at least publicly perceived that way. In this case, governments would most probably require some sort of safeguard technology in order to effectively bracket moral reasons from autonomous evaluation by the agents in question. Given that the cognitive system of a fully autonomous agent would have to be sufficiently complex, this might not even be a feasible task. If that were the case, safety requirements to keep AMAs from reasoning about their own morality would effectively prevent the production of an otherwise fully autonomous moral agents—these agents would be either fully autonomous or generally restricted in their reasoning capacities. Cordoning off only the moral areas of their practical reasoning capacities might be technically impossible.

Of course, we can imagine the opposite as well, that it could indeed be technically feasible to exclude only a definite set of reasons from further reflection and re-evaluation, while preserving the capacity for autonomous reasoning in all other regards. This would be the ideal outcome for manufacturers in a scenario with legal restrictions on the reasoning of AMAs. They still could build artificial agents that were fully autonomous with the exception of their moral reasoning—morally fail-safe agents, so to speak.

While these artificial agents would be safe moral agents, they would not be praiseworthy. Moral praiseworthiness, at least in our contemporary moral practice, is closely linked to the capacity to make autonomous decisions, particularly in weighing moral reasons against pressing reasons of other sorts—self-interest or the interests of those one cares for, private or political commitments, and so forth. Moral praiseworthiness does not apply to an agent who has literally no other choice than to follow the moral rules that have been programmed to override all other reasons in contexts of conflicting reasons.

7 Conclusion

What I hope to have shown in this paper is that the claim about creating praiseworthy artificial moral agents, taken seriously, would have to result in AMAs that have the potential for moral fallibility. Their moral fallibility would not be rooted in psychological quirks that often plague human moral agency, but rather in the fact that moral rule systems are necessarily incomplete and only rough guidelines for action.

This does by no means imply that AMAs will necessarily turn into amoral agents. They could try to compensate the weakness of their moral beliefs by productively improving them, by improvising, or by plainly accepting that moral guidance is not available in some situations. The crucial point is that this optimistic turnout is not the only possible result, but that a turn toward amoral behaviour is possible as well. In this regard, full AMAs would resemble human moral agents fairly well.

I want to remain agnostic on the question whether it will one day be possible to generate artificial intelligence that is capable of the kind of autonomous reasoning this paper focuses on. However, if it were the case, then moral praiseworthiness would not be a matter of technical restrictions. An AMA that were restricted in its capacities for reflecting on and possibly re-evaluating its moral reasons, that were hard-wired to act morally no matter what, would be at best a good moral agent in the sense that a sharp knife is a good knife, but could never be a praiseworthy moral agent in any meaningful sense.

References

Allen C, Varner G, Zinser J (2000) Prolegomena to any future artificial agent. J Exp Theor Artif Intell 12:251–261

Anderson DL (2013) Machine intentionality, the moral status of machines, and the composition problem. In: Müller VC (ed) Philosophy and theory of artificial intelligence. Springer, Berlin, pp 321–333

Anderson M, Anderson SL (2007) Machine ethics: creating an ethical intelligent agent. AI Mag 28(4):15–26

Audi R (2001) The architecture of reason: the structure and substance of rationality. Oxford University Press, Oxford

Baier A (1985) Theory and reflective practices. In: Postures of the mind: essays on mind and morals. Methuen, London, pp 207–227

Bigari J (2015) The relation of education for autonomy and education for morality: implications for debates over educational aims. Dissertation, The University of British Columbia

Chan S, Harris J (2011) Moral enhancement and pro-social behaviour. J Med Ethics 37(3):130–131

Floridi L, Sanders JW (2004) On the morality of artifical agents. Mind Mach 14:349–379

Frankfurt HG (1971) Freedom of the will and the concept of a person. In: The importance of what we care about: philosophical essays. Cambridge University Press, Cambridge, pp 11–25

Gaus GF (2012) The order of public reason: a theory of freedom and morality in a diverse and bounded world. Cambridge University Press, Cambridge

Gibbard A (1990) Wise choices, apt feelings: a theory of normative judgment. Clarendon Press, Oxford

Grau C (2006) There is no "I" in "Robot": robots and utilitarianism. IEEE Intell Syst 21(4):52–55

Harris J (2011) Moral enhancement and freedom. Bioethics 25(2):102–111

Harsanyi JC (1977) Rule utilitarianism and decision theory. Erkenntnis 11:25–53

Hume D (1738) A treatise of human nature. Clarendon Press, Oxford

Kim J (2007) The causal efficacy of consciousness. In: Velmans M (ed) The Blackwell companion to consciousness. Blackwell, Malden, pp 406–417

Korsgaard CM (1996) The sources of normativity. Cambridge University Press, Cambridge

Mill JS (1861) Utilitarianism. In: Gray J (ed) On liberty and other essays. Oxford University Press, Oxford, pp 131–204

Miller RB (2009) Actual rule utilitarianism. J Philos 106(1):5–28

O'Neill O (1987) Abstraction, idealization and ideology in ethics. R Inst Philos Lect Ser 22:55–69

Powers TM (2011) Incremental machine ethics. IEEE Robot Autom 18(1):51–58

Raz J (1999) Practical reason and norms. Oxford University Press, Oxford

Russell S, Dewey D, Tegmark M (2015) Research priorities for robust and beneficial artificial intelligence. AI Mag 36(4):105–114

Searle JR (1983) Intentionality: an essay in the philosophy of mind. Cambridge University Press, Cambridge

Searle JR (2001) Rationality in action. MIT Press, Cambridge

Stahl BC (2004) Information, ethics, and computers: the problem of autonomous moral agents. Mind Mach 14:67–83
Strawson PF (1962) Freedom and resentment. In: Freedom and resentment and other essays. Routledge, London, pp 1–28
Williams B (1987) Internal and external reasons. In: Millgram E (ed) Varieties of practical reasoning. MIT Press, Cambridge, pp 77–98
Winch C (2006) Education, autonomy and critical thinking. Routledge, London