Ethics and Information Technology (2024) 26:13
https://doi.org/10.1007/s10676-024-09755-9

ORIGINAL PAPER

Moral sensitivity and the limits of artificial moral agents

Joris Graff[1]

Published online: 24 February 2024

Abstract
Machine ethics is the field that strives to develop 'artificial moral agents' (AMAs), artificial systems that can autonomously make moral decisions. Some authors have questioned the feasibility of machine ethics, by questioning whether artificial systems can possess moral competence, or the capacity to reach morally right decisions in various situations. This paper explores this question by drawing on the work of several moral philosophers (McDowell, Wiggins, Hampshire, and Nussbaum) who have characterised moral competence in a manner inspired by Aristotle. Although disparate in many ways, these philosophers all emphasise what may be called 'moral sensitivity' as a precondition for moral competence. Moral sensitivity is the uncodified, practical skill to recognise, in a range of situations, which features of the situations are morally relevant, and how they are relevant. This paper argues that the main types of AMAs currently proposed are incapable of full moral sensitivity. First, top-down AMAs that proceed from fixed rule-sets are too rigid to respond appropriately to the wide range of qualitatively unique factors that moral sensitivity gives access to. Second, bottom-up AMAs that learn moral behaviour from examples are at risk of generalising from these examples in undesirable ways, as they lack embedding in what Wittgenstein calls a 'form of life', which allows humans to appropriately learn from moral examples. The paper concludes that AMAs are unlikely to possess full moral competence, but closes by suggesting that they may still be feasible in restricted domains of public morality, where moral sensitivity plays a smaller role.

Keywords Machine ethics · Artificial moral agents · Moral sensitivity · Uncodifiability

Introduction

As artificial intelligence (AI) takes over increasingly many tasks, the question arises what should happen when AI systems start making decisions that have moral importance. Currently, the usage of AI systems is being proposed, or already in place, for tasks such as predicting re-offence risks of criminal defendants (Angwin et al., 2016), medical diagnosis (Sand et al., 2022), warfare (Umbrello et al., 2020), and traffic behaviour (Nyholm, 2018a, b) – just to give a few examples. When humans perform such tasks, they are expected to exercise *moral competence* in order to arrive at morally right decisions. The question is whether AI systems could exercise the same moral competence. If so, they may be able to autonomously perform such tasks in a morally unproblematic way, which, in some domains, may have significant advantages as a result of AI's information-processing capacities.

Some authors are optimistic about the potential moral competence of AI systems and have proposed that we design 'artificial moral agents' (AMAs) (Allen et al., 2005), that is, AI systems that have been explicitly equipped with moral reasoning. The field that aims to design AMAs is known as 'machine ethics' (Anderson et al., 2004; Allen et al., 2005; Wallach & Allen, 2008; Anderson & Anderson, 2011; Cervantes et al., 2020). Most proponents of machine ethics do not necessarily believe that AMAs can be *full* moral agents (in Moor's (2011) terminology). This is because there is widespread agreement that AI systems do not possess some features deemed necessary for full moral agency, such as having intentions (Johnson, 2006), self-determination of rules (Hew, 2014) or purposes (Fossa, 2018), sentience (Véliz, 2021), or moral personality (Sparrow, 2021). However, many authors claim that we could still design AI systems that can reliably make morally right decisions, without being full moral agents (Johnson, 2006; Wallach & Allen,

✉ Joris Graff
j.j.graff@uu.nl

[1] Department of Philosophy and Religious Studies, Utrecht University, Utrecht, The Netherlands

2008, Chap. 4; Anderson, 2011; Fossa, 2018). I will call such limited AMAs 'functionally moral systems' (FMSs) (following Wallach and Allen's (2008) notion of 'functional morality').

Others, however, are less optimistic about the prospects of machine ethics (Brundage, 2014; Hew, 2014; Sparrow, 2021; Véliz, 2021). It is not always clear whether these authors argue only against the possibility of artificial full moral agents, or, more strongly, against the possibility of artificial FMSs. The aim of this paper is to specifically examine the second claim, i.e. the idea that AI systems cannot even be *functionally* moral, by considering a skill that plausibly underlies the capacity to reliably do the right thing, which I will denote as 'moral sensitivity'. This is an important issue, because, if AI systems cannot be FMSs due to lacking moral sensitivity, then the whole project of machine ethics would be cast in doubt.

The layout of the remainder of the paper is as follows. After more clearly defining FMSs and the related notion of moral competence (Section "Functionally moral systems"), I will zoom in on the notion of moral sensitivity. On the basis of moral philosophers working in the Aristotelian tradition, I will argue that moral sensitivity is likely necessary for full moral competence (Section "Moral sensitivity as a prerequisite for moral competence"). I will then argue that the two main types of AMAs that are generally distinguished, i.e. top-down and bottom-up AMAs, cannot possess full moral sensitivity, i.e. not to the extent that (some) human moral agents do (Sects. "Top-down AMAs and uncodifiability"–"Bottom-up AMAs and moral training"). Subsequently, however, I suggest that machine ethics can still have a role in *limited* morally loaded domains in which moral sensitivity plays a lesser or no role (Section "AMAs in limited moral domains").

Functionally moral systems

I define a functionally moral system (FMS) as a system that is not a full moral agent, but that, with a sufficient reliability, does the morally right thing in a certain range of situations.

Some clarifications about this definition are in order. First, we must specify what 'doing the right thing' amounts to. I will say that, in a given situation where the morally relevant reasons are such that some but not all of the potential actions open to a system are morally acceptable, the system does the right thing iff it chooses one of the morally acceptable options. This presupposes that we have some way of determining which options are acceptable and which are not. This is of course a significant problem, which however is outside the scope of the current paper. For the sake of the argument, I will assume some shared intuitions on what is the morally right choice in a range of decision-making situations. Insofar as these shared intuitions are absent, machine ethics, of a kind that is democratically legitimate, will of course become more problematic. But the goal of the current paper is to examine obstacles for machine ethics *even if* significant moral agreement can be assumed.

Second, we need to explicate the term 'sufficient reliability'. On the basis of what criteria would we say that a system *reliably* behaves morally? It is not enough to observe that a system has, with sufficient frequency, done the right thing so far. After all, it may be that a system has, in the past, done the right thing by mere luck, e.g. because its non-moral programming always happened to align with moral demands. In order to be reliably moral, it also needs to be the case that the system *would* do the right thing in a range of counterfactual situations we may put it in.

But what criteria do we have to establish reliable performance in counterfactual or future situations? Mere induction from past cases is problematic, for reasons that will be spelt out in Section "Bottom-up AMAs and moral training". Rather, a system's current behaviour must, over and above a tendency to make morally right choices, also display signs that these choices are the result of reliable capacities (instead of mere luck). What exact capacities are in question will be spelt out more in the rest of this paper. To give one example: in humans, we tend to take someone's citing appropriate moral reasons as a relevant sign, because it indicates a certain level of moral understanding. Any systems that shows signs of the relevant capacities – and thus, can be said to *reliably* do the right thing – I will denote by the term *morally competent*. It is thus moral competence, not merely doing the right thing (so far), that we are interested in if we want *reliably* moral FMSs.[1]

Third, a short note on this notion of a 'range of situations'. This range could, in principle, be very wide (e.g. all actions that full moral agents would be able to perform morally), or very restricted (e.g. all actions within a clearly limited domain, such as how to distribute medical resources in a given hospital). In the first case, I will speak of *strong* FMSs, and in the latter, of *weak* FMSs. As FMSs become increasingly weak, they gradually transition into systems that are so specialised that the term FMS seems no longer applicable, e.g. automatic thermostats.

The remainder of this paper is concerned mostly with arguing that weak FMSs are the best we can hope for

[1] It may appear that this notion of moral competence moves beyond the focus on functional morality and therefore begs the question against defenders of functionally moral systems. But note that nothing said so far excludes the possibility that the capacities in question are themselves functional. That is to say, the definition of moral competence does not exclude, *a priori*, the possibility that a system may be functionally competent by possessing certain information-processing capacities without possessing moral understanding.

with foreseeable technologies. Section "AMAs in limited moral domains" will then roughly characterise the kinds of domains in which such weak FMSs may be feasible.

Moral sensitivity as a prerequisite for moral competence

To assess the feasibility of (strong) artificial FMSs, we need to establish what is required for a system to be morally competent. Since the second half of the previous century, many philosophers, often inspired by Aristotle, have argued that moral competence would not be possible without a capacity (or, perhaps more accurately, a cluster of capacities) which I will here call 'moral sensitivity'. In a word, moral sensitivity is the practical skill to recognise, in a range of situations, which features of the situations are morally relevant, and in what way they are morally relevant. This concept – and a number of related concepts, including 'sensibility', 'attention', 'vision' and 'discernment' – has a long pedigree, appearing in the works of, among others, Aristotle himself (c. 330 BC/2006), Levinas (1978), Weil (1950/2009), and Murdoch (1970/2013). I will here aim to clarify this concept on the basis of works by (more or less) neo-Aristotelian philosophers working in the late 20th century, specifically McDowell (1979), Nussbaum (1992, especially Chaps. 2, 4 and 5), Wiggins (1975;, 2012), and Hampshire (1978).[2] These philosophers differ in several relevant aspects, but they share a common focus that will allow us to characterise the notion of moral sensitivity to a sufficient degree. My goal here is to highlight three core features of moral sensitivity that are stressed by all authors mentioned here: that it is uncodifiable, that it is a practical skill, and that it is semi-perceptual in nature.

First, all authors mentioned stress that moral competence cannot be codified; that is, there cannot be an exhaustive set of general rules that tell us which features of decision situations are generally morally relevant, and in what way they are relevant. I will call this the *uncodifiability thesis*.[3] The main motivation behind this thesis is that moral decisions are contextually dependent to a degree that any ruleset, when applied to particular situations, would sometimes leave out features that properly influence moral decisions. McDowell, for instance, states: 'If one attempted to reduce one's conception of what virtue requires to a set of rules, then, however subtle and thoughtful one was in drawing up the code, cases would inevitably turn up in which a mechanical application of the rules would strike one as wrong' (McDowell, 1979, p. 337). Why is the moral domain such that it cannot be captured in a finite set of rules, even in principle? Nussbaum (1992, Chap. 2) argues that the contents of such rules must be repeatable, but, according to her (reading of Aristotle), often, morally relevant features properly impacting our moral decisions are qualitatively unique. This is especially clear in the domain of interpersonal relationships. For instance:

> Good friends will attend to the particular needs and concerns of their friends, benefiting them for the sake of what they are, in and of themselves. Some of this "themselves" consists of repeatable character traits; but features of shared history and of family relationship that are not even in principle repeatable are allowed to bear serious ethical weight. Here the agent's own historical singularity (and/or the historical singularity of the relationship itself) enter into moral deliberation in a way that could not even in principle give rise to a universal principle, since what is ethically important (among other things) is to treat the friend as a unique nonreplaceable being, a being not like anyone else in the world. (Nussbaum, 1992, p. 72)

One might, of course, subsume one's relationship to one's friend under general rules on the basis of *some* features that it shares with other relationships. But the point is that any finite set of rules would always leave out the qualitative uniqueness of the relationship in a way that may turn out to be morally problematic. Hampshire considers this a general feature of human experience, which he calls the 'inexhaustibility of description': no description in general terms, that could serve as the premise of a general rule, could exhaust the open-ended set of features facing us in any situation (Hampshire, 1978, p. 30). Without committing to this general claim, we can agree that descriptions of moral situations are plausibly inexhaustible, due to the unique character of human relationships.

None of this is to say that, according to neo-Aristotelians, rules should play no role in moral practice; Nussbaum, for instance, explicitly allows for a limited use for moral rules (Nussbaum, 1992, pp. 68–73). However, proper usage of these rules is dependent on an uncodified moral sensitivity that allows a (good) moral agent to know whether and how the rules should be applied in particular situations.

[2] Similar ideas – albeit stemming from a different background – have been developed by Wittgensteinian philosophers working on ethics, most notably Cavell (1979, especially part 3), Diamond (1991, especially Chaps. 11–15), and Crary (2007) (McDowell may be positioned in the intersection of this tradition and the Aristotelian tradition). I return to this strand of thought in Section "Bottom-up AMAs and moral training".

[3] The uncodifiability thesis should not be confused with moral particularism, which is the stronger thesis, mostly associated with Dancy (2004) (who is inspired by McDowell), that moral reasoning does not require *any* principles.

The second feature – that moral sensitivity is a practical skill – offers a more positive characterisation of the notion. Such a characterisation is difficult, since it is of the nature of practical skills that they resist theoretical description. The capacity may best be characterised by example (Nussbaum, 1992 believes that it is best brought out by literary works), or by analogy with other uncodified skills. For instance, Nussbaum compares moral sensitivity to improvisational theatre or playing jazz music (Nussbaum, 1992, p. 74) and Hampshire compares it to the skill of translating texts (Hampshire, 1978, pp. 31–33). More generally, Wiggins (2012) invokes Ryle's notion of *knowing how* to characterise moral thinking. A person who knows how to perform a certain task reliably and confidently performs the task (in the right circumstances), but may not necessarily be able to state in propositional format what is required to perform the task (which would be an instance of *knowing that*).

This may make it appear like moral sensitivity is an purely unthinking skill. But McDowell (1979) and Nussbaum (1992, Chap. 2) stress that moral behaviour is a form of *rational* behaviour. This raises the question how moral competence can be rational, given its lack of reliance on fixed rules. McDowell, in answering this question, subsumes moral competence under a Wittgensteinian picture of rationality, where our confidence in others' rational behaviour is eventually grounded in what Wittgenstein calls a shared 'form of life' (McDowell, 1979, pp. 336–342). We will return to this notion in more detail in Section "Bottom-up AMAs and moral training".

The third feature that recurs in the work of neo-Aristotelian philosophers is that moral competence is more or less perceptual in nature. Drawing on Aristotle, Nussbaum stresses that the morally competent person is marked by an 'intuitive perception' or 'keen vision' (Nussbaum, 1992, p. 141). A good moral agent has the capacity to non-inferentially recognise the morally relevant features that are present in a particular moral decision situation, and also to recognise the ways in which these features are relevant to her decision. McDowell (1979) similarly identifies virtue with a type of sensitivity, which he characterises as 'a sort of perceptual capacity' (McDowell, 1979, p. 332), specifically the capacity 'to recognize requirements which situations impose on one's behaviour' (McDowell, 1979, p. 333). In both of these descriptions, it becomes clear that this type of perception should not be seen as a neutral registering of sensory inputs; rather, it is perceiving features *as* morally relevant. Hampshire makes this clearer by stating that moral reasoning resembles 'perceptual identification' (Hampshire, 1978, p. 24): just like we can, in perception, recognise a person in front of us *as* someone we know (without having to go through an explicit reasoning stage), we can, in moral perception, recognise a feature as morally relevant. This semi-perceptual conception of moral sensitivity does not need to entail that moral perception (or perceptual identification) is *all* that is required for moral reasoning (although McDowell appears to hold this view); however, if other cognitive faculties play a role in moral reasoning, they cannot function in the absence of this semi-perceptual skill.

It is not my intention to fully endorse the Aristotelian picture of moral competence. Indeed, I do not know if there is a single such picture, and if there is, I have left out many relevant aspects. Rather, I wish to stress that moral sensitivity, characterised by the above three features, appears at least as *part* of our moral competence, once we reflect on our commonsense moral practices. First, our moral duties, especially towards people to whom we stand in personal relationships, often rest on qualitatively unique features, such as the shape of a specific relationship. Second, common experience bears out that moral competence cannot be taught only by feeding someone propositional knowledge; rather, people require guided training by morally competent adults to become themselves morally competent. Third, upon perceiving a moral situation in which we are to act, we often directly identify what is our moral duty, rather than having to go through a reasoning process.

If we want to determine if strong artificial FMSs are possible, then, we should consider whether AI systems can possess moral sensitivity as characterised here. To start to answer this question, I will consider two types of AMAs distinguished by Allen et al. (2005): *top-down* and *bottom-up* AMAs. Roughly, top-down AMAs are programmed to follow predetermined moral rules, while bottom-up AMAs are programmed to learn moral behaviour from concrete situations.[4]

Top-down AMAs and uncodifiability

In general, top-down AMAs follow rules that move from descriptions of moral situations in terms of morally relevant features to moral recommendations. That is, they are programmed to follow rules, of which at least some are of the form[5]:

[4] Allen et al. (2005) distinguish a third type, namely *hybrid* AMAs, which combine features of top-down and bottom-up AMAs. I will not discuss this type in detail, since it would combine the problems associated with both other types. That is – to anticipate the upcoming argument – *either* the learning module of a hybrid AMA would be constrained by a rule-set, in which case the uncodifiability problem outlined in Section "Top-down AMAs and uncodifiability" would still apply, *or* rules can be overruled by learnt patterns, in which case the reliability of this pattern-matching module is cast in doubt by the arguments in Section "Bottom-up AMAs and moral training".

[5] The reason why only 'at least some' of the rules need to be of this form, is that there can also be 'meta-rules', e.g. rules that disable other rules, or determine which rule wins out in a case in which two

$$F_1, \ldots, F_n \rightarrow A \quad (1)$$

where F_1,...,F_n are types of morally relevant features, and A is some action. The F_i's are *types* of morally relevant features, since, if the rule-set is to be finitely large, the rules must be applicable to more than one situation. For instance, imagine a healthcare robot tasked, among other things, with reminding an elderly patient of taking his medicine. If the patient would refuse to take the medicine after being reminded, the robot may be faced with the choice whether or not to alert the patient's family. If it is a top-down AMA, it may contain a rule of the form 'If patient X refuses medicine Y, and not taking medicine Y would raise the risk of lethal diseases, then alert X's family'. In this rule 'patient X refuses medicine Y' is not itself a feature; it can, however, be *instantiated* in specific situations by replacing X and Y by a specific patient and medicine. In other words, a rule applies to a situation if that situation contains a token relevant feature that falls under a type covered by the rule-set. The morally relevant features covered by the rule-set will typically be higher-order features, since input features (such as camera pixel inputs) usually do not have direct moral relevance. For a top-down AMA to function, it therefore either needs to have modules able to extract higher-order features from input features, or it needs to be directly provided with higher-order features by humans. In the latter case, the AMA could hardly be considered a FMS, since a crucial part of the moral work – recognising which features are potentially morally relevant – will have been taken over by humans. We therefore focus on the first type.

It is clear that the uncodifiability of moral sensitivity is a direct problem for rule-sets of this form. To put the point simply: any computable rule-set must refer to a limited number of types of morally relevant features F_1, F_2, etc. However, such a set is likely unable to capture all potential morally relevant features that an agent may encounter. The reason is the one we encountered in Section "Moral sensitivity as a prerequisite for moral competence", i.e. the fact that many morally relevant features, especially those in the sphere of human relationships, are qualitatively unique, and therefore cannot be captured under repeatable types. Top-down AMAs, then, fall short of the practical skill which humans exercise when they detect the morally relevant features of a situation. A similar point is stressed by Hasselberger, who states that 'computer algorithms, unlike human agents, do not have tacit practical knowledge, empathetic emotional understanding, or any unexplicit "feel" for the moral background of a situation' (Hasselberger, 2019, p. 987).

To return to the above example: when we ask the question 'Should the healthcare robot alert the patient's family?', the correct answer, intuitively, is: 'It depends'. Features that the decision depends on include, but are not limited to, the patient's mental competence (does he appear capable of rational decision-making?) but also the patient's relationship to his family (are they overbearing? if so, to what extent?). Since for instance the patient's relationship to his family can take myriad shapes, it cannot be fully captured under repeatable feature types.

It may be thought that general rules can still apply if they refer to feature types that are sufficiently high-level – e.g. 'the patient's family is appropriately concerned with the patient's health'. But this just relocates the problem, for (still assuming that we are dealing with a top-down AMA) we need rules to determine what behaviour counts as 'being appropriately concerned'. But appropriate concern can be expressed in so many different modes (through speech, tone of voice, frequency of visits, non-verbal behaviour of myriad kinds, etc.) that these rules themselves are uncodifiable (cf. Hasselberger, 2019).

The uncodifiability problem gives us strong reasons to believe that a purely top-down AMA could not be fully functionally moral. One potential objection to this conclusion is that the argument only shows that full FMSs would be imperfect, since they would always miss some morally relevant features, not that they should not be regarded as having full moral competence at all. The objection could be supported by calling to mind the fact that human moral agents are also imperfect, often significantly so. Many humans are not particularly sensitive to the moral requirements of situations. Even if they are, they may have other flaws that impede their moral competence, which top-down AMAs may not have. For instance, humans often let their judgements be influenced by cognitive biases in ways that they would not endorse upon reflection, and it is likely that such biases would also apply in moral domains (see e.g. Caviola et al., 2014). Top-down AMAs may be able to overcome at least some of these biases.

There are, however, significant differences between morally imperfect humans and top-down AMAs with overly rigid rule-sets. Most importantly, there exist social and institutional practices for improving the behaviour of those who act in morally problematic ways. If someone acts immorally, people within her direct environment may aim to correct her by pointing her attention to morally relevant factors that she showed insufficient concern for. If the offence is of a particularly severe nature, this task may be taken over by the justice system. These mechanisms are, of course, very imperfect. But all human moral agents must be capable of

conflicting rules apply to a situation. Such rules would still proceed on the basis of types of morally relevant features. For instance, a top-down AMA may have a rule of the form 'If the person to whom I promised X released me from the promise, disregard the rule saying that I should do X'.

revising their views on the basis of social correction to some extent. If a person were fully incapable of this, we would start questioning her moral competence, e.g. due to psychological defects such as psychopathy, and not entrust morally loaded tasks to this person. Top-down AMAs, on the other hand, can only be corrected by adding new rules to their rule-set. This procedure is rather ad hoc, however, since it only solves a very specific moral shortcoming and does nothing to improve the system's overall moral sensitivity. Thus, we can address human moral fallibility in ways that do not extend to top-down AMAs.

Bottom-up AMAs and moral training

Bottom-up AMAs – i.e. AMAs that learn moral behaviour – can take different forms. Some bottom-up AMAs that have been proposed simply replace the pre-programmed rule-sets discussed above by learnt rule-sets (see for instance Anderson, Anderson and Armen's W.D. (Anderson et al., 2005), later refined as GenEth (Anderson & Anderson, 2018)). It is clear that such bottom-up AMAs cannot overcome the uncodifiability problem. For one thing, such AMAs depend on humans to do most of the ethical work for them in feeding them certain valuations of *pro tanto* duties for particular situations. (This is not necessarily a problem in itself, but it means such systems cannot be deployed without direct oversight, meaning it is questionable whether they fall within the domain of machine ethics.) Moreover, even if systems like GenEth were capable of autonomously inferring which of a circumscribed number of *pro tanto* duties apply in particular situations, there are likely other ethically relevant features that cannot be subsumed under one of these duties, or that affect the importance of the duties in particular situations.

To overcome the uncodifiability problem, then, we require a machine learning method that is capable of *extracting* morally relevant features on the basis of raw input features, rather than relying on pre-determined morally relevant features. Therefore, what is required is a bottom-up AMA that uses a *deep learning* method capable of feature extraction. The most prominent class of deep learning methods consists of artificial neural networks (ANNs). ANNs can be trained to classify inputs that are specified in terms of raw input features by automatically extracting higher-order features and representing these features in intermediary layers. Moreover, these higher-order representations are *distributed* over a large number of nodes. Such distributed representations allow for much more flexible input-output mappings than do explicit rule-sets. This suggests that ANNs may be able to replicate the uncodified moral sensitivity that was characterised in Section "Moral sensitivity as a prerequisite for moral competence". Indeed, some authors who are impressed by the uncodifiability of moral reasoning have suggested that ANNs may offer a solution to moral learning, including Dancy (1999). Guarini, inspired by Dancy's moral particularism, trained an ANN on the moral intuitions of 60 college students regarding a set of moral examples; the ANN extended to new examples with some accuracy (Guarini, 2006).

This approach may seem plausible when we focus only on the uncodifiability thesis. But when we extend our attention to the other features of moral sensitivity outlined in Section "Moral sensitivity as a prerequisite for moral competence", it becomes apparent that the way in which (some) humans learn to behave morally (and learn in general) is relevantly different from the training process of ANNs. Recall from Section "Moral sensitivity as a prerequisite for moral competence" that the consistency of moral behaviour, if it is not grounded in moral rules, must be grounded the way other practical skills ('knowing how') are. The picture that suggests itself, and that is explicitly leveraged by McDowell (1979), is a Wittgensteinian one. It helps to elaborate this picture a bit further, in order to understand to what extent ANNs conform to it.

Wittgenstein, in much of his later work – most notably parts of the *Philosophical Investigations* (1963) and the *Remarks on the Foundations of Mathematics* (1964) – is concerned with the fact that we are often successfully instructed to behave in predictable ways, even though the training we have received does not force any specific future behaviour. Examples of such predictable behaviour include the proper usage of words (Wittgenstein, 1963, § 6, 26*ff.*, 81*ff.*) and the proper continuation of mathematical patterns (Wittgenstein, 1963, § 143*ff.*, 185*ff.*; 1964, §I.1*ff.*), but the point can also be extended to moral behaviour. Usually, when we learn to behave in a predictable way (e.g. to use words properly, or to respond properly to moral situations), we do so on the basis of a limited number of instructions. The apparent problem is that any finite number of training instances can be consistently extended in an indefinitely large number of ways. Even if we are given an explicit rule that tells us how to continue the pattern, we can still interpret this rule in an indefinitely large number of ways, and if we formulate further rules for the proper interpretation of the first rule, we embark on a regress. (This is the core of Wittgenstein's famous 'rule-following considerations'; see Wittgenstein, 1963, § 185–242; 1964, parts I and V.) How then, despite this apparent indeterminacy, do we manage to settle down on largely the same verbal, mathematical or moral practices? Wittgenstein stresses that, as humans, we are able to align our behaviour with *customs* that are dominant within a certain community (e.g. Wittgenstein, 1963, § 199). Insofar as this ability is grounded in anything, it is not grounded in any explicit rule-set, but rather in our

shared practical understanding and experience which Wittgenstein sometimes calls our *form of life* (see e.g. Wittgenstein, 1963, § 241).

Wittgenstein is notoriously silent on what this form of life consists of. McDowell (1979) is only marginally more explicit, mostly by quoting a passage by Stanley Cavell, who stresses 'our sharing routes of interest and feeling, modes of response, senses of humor and of significance and of fulfilment, of what is outrageous, of what is similar to what else, what a rebuke, what forgiveness, of when an utterance is an assertion, when an appeal, when an explanation – all the whirl of organism Wittgenstein calls "forms of life"' (Cavell, 2015, p. 48). Indeed, Cavell's work (and that of similar Wittgensteinian philosophers, such as Diamond (1991) and Crary (2007) contains some important insights into the background of shared (moral) understanding. Without aiming to bring out these ideas in their entirety, it is therefore helpful to consider some of Cavell's suggestions in more detail.

Elsewhere, Cavell states:

> Instead, then, of saying either that we *tell* beginners what words mean, or that we *teach* them what objects are, I will say: We initiate them, into the relevant forms of life held in language and gathered around the objects and persons of our world. For that to be possible, we must make ourselves exemplary and take responsibility for that assumption of authority; and the initiate must be able to follow us, in however rudimentary a way, *naturally* (look where our finger points, laugh at what we laugh at, comfort what we comfort, notice what we notice, find alike or remarkable or ordinary what we find alike or remarkable or ordinary, feel pain at what we feel pain at, enjoy the weather or the notion we enjoy, make the sounds we make) […] (Cavell, 1979, p. 178)

Similar remarks apply to what we may call moral training (or 'initiation'). What is required for mutual moral understanding (i.e. shared moral sensitivity) to come about, then, is a shared set of attitudes and practical skills. Note that Cavell stresses that some of these attitudes and skills must already be shared between instructor and initiate *before* any training starts (e.g., following pointing fingers in similar ways), whereas others are brought about by the 'initiation' in a certain form of life (once acquired, these culturally transmitted skills and attitudes may then of course be drawn upon to initiate a new phase of training). The former are likely embedded in our biology, whereas the latter are instilled as a part of our general upbringing. Indeed, both types of agreement are required to bring about mutual understanding: without a shared 'natural' background, there would be no way for training to get a foothold, but without a more specific cultural background, we would never reach common understanding on the more specific customs of language use, mathematics, or morality. The moral sensitivity discussed in Section "Moral sensitivity as a prerequisite for moral competence", then, must depend partially on our shared biology, and partially on our shared upbringing.

One of the more 'natural' skills required for moral learning seems to be what we may call 'empathetic understanding', i.e. the capacity to look at situations from another person's viewpoint. This understanding allows us to, upon encountering a novel moral situation, immediately understand which effects of our potential actions would constitute harm to another person, and which would not. Another set of skills, which may be based in our biology and then finetuned through our general upbringing, is the ability to understand *structured* instruction. When someone explains to us that something we did was wrong, she does not merely tell us that the act was wrong, but also suggests the particular way in which the act was wrong, e.g. by drawing on familiar examples, by prompting our imagination ('How would you feel if something like this happened to you?'), or by using a certain tone of voice. We understand these cues because we are accustomed to a 'whirl of organism' in which they have a place.

How is this Wittgensteinian picture of moral training relevant to our confidence – or lack thereof – that another agent, or system, has moral competence? Of course, it is not (generally) the case that, in assessing another's moral competence, we inquire into the precise history of her moral upbringing.[6] Rather, we look for indications in the person's current behaviour that she sees moral situations sufficiently similar as we do. Notably, as mentioned in Section "Functionally moral systems", this requires not only that she does the right thing sufficiently often. We also consider what reasons she offers for her actions, and the way in which she presents these reasons. Cavell, for instance, stresses that when someone adduces a reason for breaking a promise (in other words, provides a *defence* for not keeping it), then

> the *way* the reason is entered is critical to whether it will be acceptable – the tone of voice, the occasion on which it is entered, whether you tried to call the promise off before simply not keeping it […] – all of which serve to *acknowledge* your awareness of what it is you have done. Without the expression of that awareness, even the extreme defense is incompetent […] (Cavell, 1979, p. 297).

[6] I thank an anonymous reviewer for pressing this point.

In other words, we assess a person's moral competence not just on the basis of her moral decisions, but also on the basis of whether she seems to understand, or be sensitive to, these decisions (in that situation as well as in other moral situations). And many of the criteria we use for establishing such understanding only make sense given a shared form of life – e.g., the usage of a certain tone of voice that naturally conveys a certain attitude to us. A shared biology and upbringing are thus necessary for attributing moral competence – not in the sense that we directly consider these factors as criteria for moral competence, but in the sense that our criteria only apply to those kinds of persons that share enough of our biological and cultural background. As Wittgenstein says: 'Only of a living human being and what resembles (behaves like) a living human being can one say: it has sensations; it sees; it is blind; hears; is deaf; is conscious or unconscious' (Wittgenstein, 1963, § 281). We may add: only of such a being can one say: it is morally competent.

If this is the picture of how we learn and assess moral competence, the prospects for reliable bottom-up AMAs appear very slim indeed. ANNs – and AI systems in general – do not share in our form of life as characterised above. First, they do not share our biology, and therefore our natural responses, and second, they have not been initiated in our form of life through a general education. As a result, algorithms do not possess the empathetic understanding that would be required for them to get the point of the moral exemplars we may give them. Moreover, the way in which machine learning algorithms are trained is by providing either categorical or numeric training labels (e.g. 'This act is wrong', or perhaps 'This act is wrong to *x* degree'), without the more informative cues that accompany training in humans.

As a result, ANNs do not 'live in the same moral world' (Cavell, 1979, p. 297) that we inhabit. The problem is not so much that bottom-up AMAs could not, in principle, learn to do the right thing sufficiently often. It is rather that our criteria for establishing that they have learnt this are undermined. In the absence of the signs of moral understanding that Cavell talks about, we can only rely on statistical data, but these are vulnerable to Wittgensteinian doubts about pattern continuation. In this context, it is interesting to observe that ANNs sometimes fall prey to so-called *adversarial examples*. These are inputs that can be used to thoroughly 'confuse' neural networks, particularly ones used for image classification. Szegedy et al. (2013) show that slightly perturbing an image that is originally classified correctly by an ANN (e.g. as a panda), in a way that does not meaningfully change the picture in the eyes of humans, radically changes the network's prediction (e.g. to 'gibbon'). Notably, this feature often occurs even in networks that have obtained a very high accuracy on a set of training images. In such cases, the ANN generalises from the set of training examples in a way that would strike humans as absurd, just as in Wittgenstein's examples of individuals continuing arithmetic series in bizarre ways (e.g. Wittgenstein, 1963, § 185). This is just what we would expect, given that the contextual background that guides the way in which humans classify images, grounded in our form of life, is absent in the case of AI systems. But if this is a real risk in the case of image classification, it would likely also be a risk for bottom-up AMAs. There is no safeguard against algorithms going awry in novel or unexpected situations, either by failing to extract the morally relevant features, or by weighing them in ways that are alien to us.

Now, one may believe that this problem can be overcome by increasing the training set. Surely, the objection goes, the probability of an ANN taking the wrong lesson from a set of training examples decreases as the set increases in size. This claim has some plausibility, especially given the fact that some other (seemingly) uncodifiable skills have been rather successfully learnt by ANNs. The clearest examples are image classification and, more recently, natural language generation, which has been learnt to a significant extent by large language models. As long as adversarial examples are sufficiently rare, the argument would conclude, they do not endanger bottom-up AMAs' moral competence in a way that disqualifies them as FMSs. The idea is, in short, that there can be criteria for attributing moral competence that are *not* based in expressions of shared understanding.

There are, however, some salient differences between training ANNs to solve moral problems and training ANNs to generate text or recognise images, which warrant special scepticism in the former case. First, it is much more difficult to devise moral training scenarios that accurately emulate scenarios that may occur in moral practice. A training instance for an ANN requires two parts: a quantified representation of a situation and a training label. It is relatively easy to represent either an image to be classified or a written text to be responded to in a way that closely resembles a situation that may occur in actual practice. This is because both types of input include a clearly delineated set of input features (either a visual field or the digits comprising a chat conversation, argumentative text, etc.). For moral decision-making, however, the situation is very different. The input features that may comprise higher-order features are much richer: they may come from several sensory channels (vision, sound, etc.), but also from whatever shared memories exist between the decision-maker and those impacted by the moral decision. It is very difficult for a programmer to design a training instance that includes all of these input features. Thus, we are again faced with an uncodifiability problem, although it is different from the one discussed in the previous section. The uncodifiability here pertains to the

training data, rather than to the input-output mappings of the algorithm itself.

Second, it is very difficult to foresee the range of inputs that a full bottom-up AMA may possibly face. When training an image classifier, we can use images from many different positions, settings, types of lighting, etc., and be somewhat confident that this training set covers most of the space of potential images that the classifier may be presented with in the future. When we switch to the moral domain, however, there is a more significant possibility that future instances outrun the space covered by the training set. This is because moral behaviour takes place within a societal setting that is liable to rapid, often unforeseen, changes. Consider, for instance, the introduction of new technologies. The public sphere in which day-to-day interactions take place looks (and sounds, smells, etc.) different than it did even a few years ago due to the widespread adoption of new technologies (e.g. recreational drones). Humans generally have an intuitive grasp when the presence of such technologies is a morally relevant feature, and when it is not (although this grasp is of course liable to change and correction). But we have no reason to expect that ANNs trained on past moral situations will extend to these novel situations in the same ways as (most) humans would.

As a somewhat oversimplified example, consider again the hypothetical healthcare robot introduced in the previous section. Imagine that the patient is on a strict diet – say, one excluding all gluten – that is crucial for his continued wellbeing, but that he sometimes forgets this, e.g. due to cognitive impairments. It seems that, at least in some cases (e.g. when the potential harm is large enough), the robot may be justified in intervening whenever the patient tries to order gluten-rich food (first by warning the patient, but, if the warning is not understood, possibly by alerting the patient's family). If it is a bottom-up AMA, the robot may correctly learn to intervene in these ways on the basis of training examples. However, whereas humans would understand that the point of such behaviour is to prevent harm due to gluten intake, the healthcare robot may only learn to respond to products shaped like gluten-rich foods (e.g. anything that looks like a bread). In such a case, if, at some point during its deployment, gluten-free bread becomes available, the robot would unnecessarily prevent the patient from buying this new product, thus harming his autonomy for no good reason.

Given these considerations, it appears that mere statistical performance is not sufficient to establish moral competence. Therefore, the fact that AMAs lack the further signs of moral competence (shown in what reasons one offers for one's actions, one's tone of voice, etc.) is a significant problem. Now, it may be objected that humans, no matter how tried and tested in our shared form of life, may also go awry.[7] This is of course true. But again, we have ways to correct moral mistakes in humans that cannot be applied to bottom-up AMAs. The correction mechanisms we apply to humans *also* depend on the availability of Cavell's shared 'routes of interest and feeling, modes of response', etc. (e.g. using a certain tone of voice to point out a mistake, using familiar analogies) – in short, on a shared form of life. Moreover, this shared background is also necessary to assess whether the corrected person has understood the correction (they may for instance say 'I see your point' using a certain tone of voice, accompanied by certain gestures, etc.). Of course, this process is still fallible, but, if the corrected person shows signs of understanding our corrective intervention, this gives us some reason to expect improvement in future cases. If, instead, the person in question would not respond to *any* of our attempts at correction, then retraining is impossible, and we would deal with this by not assigning this person any moral tasks.

On the other hand, when a bottom-up AMA goes awry, these shared modes of understanding are not open to us – we cannot speak to an ANN in a certain tone of voice, ask it to imagine the situation from another person's point of view, etc. Our only option is to retrain the AMA in the same way we trained it originally, perhaps including more cases like the one in which it has acted wrongly. But the AMA, not sharing our modes of communication, can give us no confirmation that it understands the point of these further examples. Therefore, it cannot give us confidence that the retraining has really improved its moral competence, instead of leading to new adversarial examples.

For these reasons, even though it is, in principle, possible that a bottom-up AMA could do the right thing as well as, or better than, good human moral agents, we cannot reliably predict so in advance. This is a significant problem, since as long as we do not know that a bottom-up AMA can in fact simulate moral sensitivity, we would be unwilling to let this system take decisions the outcomes of which matter to us.

AMAs in limited moral domains

Does all of this spell the end for machine ethics? Not necessarily: I have so far only been talking about *strong* moral competence. Perhaps, however, there are limited morally relevant domains in which moral sensitivity is not necessary for an agent to do the right thing. In such domains, a

[7] I am here concerned with cases in which someone has made a moral mistake by standards that are widely agreed upon in our community, not with (related but different) cases where there is widespread disagreement about the standards themselves. The latter cases are of course also important, but, by the proviso mentioned in Section "Functionally moral systems", are not discussed here.

top-down or hybrid approach to AMAs may be successful (a fully bottom-up approach remains problematic, since it remains unclear whether a bottom-up AMA would generalise from training examples in desirable ways). Given the uncodifiability problem, such limited domains should meet the conditions that (a) it is possible to circumscribe (nearly) all features that may potentially be morally relevant to decision situations and (b) the ways in which these features are relevant can be codified in an exceptionless manner. It is beyond the scope of this paper to fully characterise what such domains may look like (this would be an important project for further work, that would also need to take into account legal dimensions). In this section, I merely offer some short suggestions.

First, since, as we have seen, uncodifiability results in large part from the unique shapes of human relationships, we may expect moral sensitivity to be less salient in domains where individual relationships matter less. Thus, AMAs may be more applicable in domains that fall under what Hampshire (1978) calls *public morality*. In general, a moral decision is part of public morality iff the agent is acting on behalf of a public organisation, such as a government body or, possibly, a private company that is expected to take decisions that impact the public interest. In such situations, taking individual considerations into account may be seen as unacceptably partial. Moreover, as Hampshire (1978) remarks, decisions made by public officials seem to require a different kind of moral accountability than do private moral decisions. If I make a decision that involves a limited number of individuals whom I know personally, I need to consider the specific values held by these people, and adapt my decisions to these values. On the other hand, if I make a public decision, I tend to know very little about the individuals that may be impacted by the decision and to whom I am therefore accountable, and about their respective values. Therefore, my moral reasoning should only appeal to features that are plausibly considered morally relevant by most individuals. This restricts the range of morally relevant features I should take into account.

However, it would be too simplistic to state that artificial FMSs can always be unproblematically employed in public domains. As Hampshire (1978) also stresses, moral sensitivity should still play a role in many public decisions, since there may still be too many features morally relevant to the decision to allow for codification. This is the case especially in domains where we believe that public officials should have individual *discretion* – that is, should be allowed to make decisions on the basis of their moral judgement that are not constrained by prior regulations. Individual discretion is a complex topic with many normative and legal aspects; here, I only wish to build on the intuition that it is *sometimes* appropriate. As an example, consider the Dutch childcare benefit scandal that came to light in 2019. In that year, parliamentary and journalistic investigations unveiled that, between 2004 and 2019, the Dutch tax office had falsely accused tens of thousands of parents of child benefit fraud, ordering them to repay their received benefits (see e.g. Henley, 2021). Most of these purported fraud cases were in fact the result of simple, non-malicious administrative errors on the part of the parents in question. Although the causes of this scandal are complex, several parliamentarians and journalists concluded that part of the problem was that tax officials had insufficient freedom to assess cases on their individual features, instead relying on general checklists and algorithms (e.g. Frederik, 2021). In other words, this is a plausible example of a failure of public morality resulting (in part) from a lack of individual discretion. The point is that such discretion requires moral sensitivity (it is hard, for instance, to provide general rules to establish whether or not an administrative mistake was malicious) and can therefore not be entrusted to rule-based AMAs.

Here then are two heuristics that could help to determine whether a moral domain is suited for application of AMAs: (1) the domain should fall within public morality, and (2) the domain should not be one where we feel discretion is appropriate. Are there moral domains that meet these characteristics? Plausibly there are; consider, for instance, medical disaster triage, i.e. the distribution of limited resources over a group of patients, given that the resources are insufficient to attend to all patients at once (Christian, 2019; Kucewicz-Czech & Damps, 2020). Most countries use highly formalised decision procedures in order to perform disaster triage, such as checklists or decision trees (see Bazyar et al., 2019 for an overview). At least part of the reason why such formalised procedures may be thought appropriate in this domain is that we feel triage decisions should be impersonal and independent of the moral intuitions of individual healthcare workers (although, of course, it is possible to question this idea).

However, it should be kept in mind that the above considerations are heuristics, rather than strong criteria, for two reasons. First, there may often be reasonable disagreement whether a domain falls within public or private morality and whether or not discretion is appropriate. Second, even if there is agreement that a domain falls outside of public morality, or that discretion is appropriate, this at most gives a *pro tanto* reason to not implement AMAs, and vice versa. There may be countervailing reasons on the other side. For instance, we may agree that traffic behaviour calls for personal discretion, but still decide to implement autonomous vehicles with moral algorithms, since such vehicles, even though they sometimes make morally wrong decisions, save many lives that would be lost due to mistakes made by human drivers. The decision whether or not to implement

AMAs needs to be made separately and carefully for each potential domain.

Conclusion

A glance of plausible Aristotelian and Wittgensteinian views of moral competence discloses a large role for moral sensitivity, that is, an uncodified, practical skill to perceive which features of decision situations are morally relevant, and how they are relevant. Neither top-down nor bottom-up AMAs seem capable of fully attaining, or simulating, this sensitivity. Top-down AMAs are, by their nature, too rigid to appreciate all the aspects of our world and our fellow humans in it that, in varying situations, strike us as important to our moral decisions. Bottom-up AMAs may obtain more flexibility in latching onto the manifold situations that the world presents us with, but they are not part of *our* world, and therefore there is no guarantee that their sensitivity, however fine-grained, aligns with our own. These general observations about moral practice call for caution in the pursuit of machine ethics.

This caution need not be total, however, as some authors similarly concerned with the limitations of AMAs seem to have suggested (Sparrow, 2021; Véliz, 2021). AMAs need not, immediately, face the full, cluttered moral domain that humans are engaged in. Section "AMAs in limited moral domains" has suggested that there may be delimited domains that are, in large part, insulated from the messiness of human relationships, because they call for a certain level of impersonality. To what extent we wish different parts of public morality to be impersonal, and whether this is compatible with the application of AMAs, is a question that can only be answered by exercising our moral sensitivity in specific cases. Whatever decisions we may, eventually, decide to transfer to AMAs, the decision whether or not transfer tasks to AMAs in the first place is not among these.

Acknowledgements I wish to thank Cindy Friedman, Björn Lundgren and Brandt van der Gaast for written comments on earlier versions of this paper. I also wish to thank the participants of the Zagreb Applied Ethics Conference (ZAEC) 2023, where a version of this paper was presented, for helpful comments. I am grateful to an anonymous reviewer for insightful and constructive comments. Finally, I thank Caoimhin Hamill for help navigating Wittgensteinian accounts of morality.

Declarations

Competing interests The author has no financial or no-financial interests to declare.

References

Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology*, *7*(3), 149–155.

Anderson, S. L. (2011). Philosophical concerns with machine ethics. In M. Anderson, & S. L. Anderson (Eds.), *Machine ethics* (pp. 162–167). Cambridge University Press.

Anderson, M., & Anderson, S. L. (2011). *Machine ethics*. Cambridge University Press.

Anderson, M., & Anderson, S. L. (2018). Geneth: A general ethical dilemma analyzer. *Paladyn Journal of Behavioral Robotics*, *9*(1), 337–357.

Anderson, M., Anderson, S. L., & Armen, C. (2004). Towards machine ethics. *AAAI-04 workshop on agent organizations: theory and practice, San Jose, CA*.

Anderson, M., Anderson, S. L., & Armen, C. (2005). Towards machine ethics: Implementing two action-based ethical theories. *Proceedings of the AAAI 2005 fall symposium on machine ethics*, 1–7.

Angwin, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *ProPublica*. Retrieved November 20, 2023, from https://www.propublica.org/article/machine-bias-risk-assessments-incriminal-sentencing.

Aristotle (2006). *Aristotle: Nicomachean ethics, books ii–iv: Translated with an introduction and commentary* (transl. and ed. Taylor, C. C.). Oxford University Press. (Originally published c. 330 BC).

Bazyar, J., Farrokhi, M., & Khankeh, H. (2019). Triage systems in mass casualty incidents and disasters: A review study with a worldwide approach. *Open Access Macedonian Journal of Medical Sciences*, *7*(3), 482–494.

Brundage, M. (2014). Limitations and risks of machine ethics. *Journal of Experimental & Theoretical Artificial Intelligence*, *26*(3), 355–372.

Cavell, S. (1979). *The claim of reason: Wittgenstein, skepticism, morality, and tragedy*. Oxford University Press.

Cavell, S. (2015). *Must we mean what we say? A book of essays*. Cambridge University Press.

Caviola, L., Mannino, A., Savulescu, J., & Faulmüller, N. (2014). Cognitive biases can affect moral intuitions about cognitive enhancement. *Frontiers in Systems Neuroscience*, *8*, 195.

Cervantes, J. A., López, S., Rodríguez, L. F., Cervantes, S., Cervantes, F., & Ramos, F. (2020). Artificial moral agents: A survey of the current status. *Science and Engineering Ethics*, *26*(2), 501–532.

Christian, M. D. (2019). Triage. *Critical care Clinics*, *35*(4), 575–589.

Crary, A. (2007). *Beyond moral judgment*. Harvard University Press.

Dancy, J. (1999). Can a particularist learn the difference between right and wrong? *The Proceedings of the Twentieth World Congress of Philosophy*, *1*, 59–72.

Dancy, J. (2004). *Ethics without principles*. Clarendon.

Diamond, C. (1991). *The realistic spirit: Wittgenstein, philosophy, and the mind*. MIT Press.

Fossa, F. (2018). Artificial moral agents: Moral mentors or sensible tools? *Ethics and Information Technology*, *20*(2), 115–126.

Frederik, J. (2021). De tragedie achter de toeslagenaffaire. *De Correspondent*, 15 January. Retrieved January 30, 2024, from https://decorrespondent.nl/11959/de-tragedie-achter-de-toeslagenaffaire/d3394c19-550f-078b1335-63020b8e15bc.

Guarini, M. (2006). Particularism and the classification and reclassification of moral cases. *IEEE Intelligent Systems*, *21*(4), 22–28.

Hampshire, S. (1978). Public and private morality. In S. Hampshire (Ed.), *Public and private morality* (pp. 23–54). Cambridge University Press.

Hasselberger, W. (2019). Ethics beyond computation: Why we can't (and shouldn't) replace human moral judgment with algorithms. *Social Research: An International Quarterly*, *86*(4), 977–999.

Henley, J. (2021). Dutch government faces collapse over child benefits scandal. *The Guardian*, 14 January. Retrieved January 30, 2024, from https://www.theguardian.com/world/2021/jan/14/dutch-government-faces-collapse-over-child-benefits-scandal.

Hew, P. C. (2014). Artificial moral agents are infeasible with foreseeable technologies. *Ethics and Information Technology*, *16*(3), 197–206.

Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology*, *8*(4), 195–204.

Kucewicz-Czech, E., & Damps, M. (2020). Triage during the covid-19 pandemic. *Anaesthesiology Intensive Therapy*, *52*(4), 312–315.

Levinas, E. (1978). *Otherwise than being or beyond essence* (transl. Alphonso Lingis). Kluwer Academic.

McDowell, J. (1979). Virtue and reason. *The Monist*, *62*(3), 331–350.

Moor, J. H. (2011). The nature, importance, and difficulty of machine ethics. In M. Anderson & S. L. Anderson (Eds.), *Machine ethics* (pp. 13–20). Cambridge University Press.

Murdoch, I. (2013). *The sovereignty of good*. Routledge. (Originally published in 1970).

Nussbaum, M. C. (1992). *Love's knowledge: Essays on philosophy and literature*. OUP USA.

Nyholm, S. (2018a). The ethics of crashes with self-driving cars: A roadmap, i. *Philosophy Compass*, *13*(7), e12507.

Nyholm, S. (2018b). The ethics of crashes with self-driving cars: A roadmap, ii. *Philosophy Compass*, *13*(7), e12506.

Sand, M., Durán, J. M., & Jongsma, K. R. (2022). Responsibility beyond design: Physicians' requirements for ethical medical Ai. *Bioethics*, *36*(2), 162–169.

Sparrow, R. (2021). Why machines cannot be moral. *AI & SOCIETY*, *36*(3), 685–693.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Umbrello, S., Torres, P., & De Bellis, A. F. (2020). The future of war: Could lethal autonomous weapons make conflict more ethical? *AI & SOCIETY*, *35*, 273–282.

Véliz, C. (2021). Moral zombies: Why algorithms are not moral agents. *AI & SOCIETY*, *36*(2), 487–497.

Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Weil, S. (2009). *Waiting for God* (transl. Emma Craufurd). HarperCollins. (Originally published in 1950).

Wiggins, D. (1975). Deliberation and practical reason. *Proceedings of the Aristotelian Society*, *76*, 29–51.

Wiggins, D. (2012). Practical knowledge: Knowing how to and knowing that. *Mind*, *121*(481), 97–130.

Wittgenstein, L. (1963). *Philosophical investigations* (ed. G.E.M. Anscombe & R. Rhees, transl. G.E.M. Anscombe). Basil Blackwell.

Wittgenstein, L. (1964). *Remarks on the foundations of mathematics* (ed. G.E.M. Anscombe, G.H. von Wright & R. Rhees, transl. G.E.M. Anscombe). Basil Blackwell.