Why machines cannot be moral

Professor Robert Sparrow,
Department of Philosophy,
Faculty of Arts
Monash University.

This is a PRE-PRESS version of a paper that appeared as:

> Sparrow, R. 2021. Why machines cannot be moral. *AI & Society: Journal of Knowledge, Culture and Communication*. Published Online: 21 January, 2021. DOI: https://doi.org/10.1007/s00146-020-01132-6.

Please cite that version.

Abstract:

The fact that real-world decisions made by artificial intelligences (AI) are often ethically loaded has led a number of authorities to advocate the development of "moral machines". I argue that the project of building "ethics" "into" machines presupposes a flawed understanding of the nature of ethics. Drawing on the work of the Australian philosopher, Raimond Gaita, I argue that ethical dilemmas are problems for particular people and not (just) problems for everyone who faces a similar situation. Moreover, the force of an ethical claim depends in part on the life history of the person who is making it. For both these reasons, machines could at best be engineered to provide a shallow simulacrum of ethics, which would have limited utility in confronting the ethical and policy dilemmas associated with AI.

Keywords: artificial intelligence; ethics; machine ethics; moral authority; Raimond Gaita.

Why machines cannot be moral

Introduction

Ethics is all the rage in discussions about Artificial Intelligence (AI). The more powerful we expect AI to become, the more important it seems that its applications be governed by appropriate ethical frameworks. Unfortunately, the general-purpose nature of AI, as well as the wide range of different situations any one system is likely to have to confront, means that it is difficult to see how specifying beforehand how machines should make ethical decisions could work to prevent them from mistreating people, or bringing about bad consequences, even in circumstances that it is highly predictable they will encounter. A popular solution to this problem in engineering circles is the idea that we will build "moral machines" or — in an alternative formulation — build ethics "into" AI.

One reason to be sceptical about this program is that it often assumes that there is no more to ethics than what the majority of people in society happen to believe at the time: ethics is treated as though it were just an inchoate version of the law (Awad et al. 2018). The fact that, as the history of slavery shows, the majority of people might be wrong about what is ethical — as might be the law — should disabuse us of this notion.

However, there is a deeper problem with this project, which infects even those approaches that acknowledge that what is right and wrong might be independent of what we happen to believe to be right or wrong at any particular historical moment. Drawing on the work of the Australian philosopher, Raimond Gaita, I will suggest that ethics is personal in a way that science is not. Ethical dilemmas are problems for particular people and not (just) problems for everyone who faces a similar situation. Moreover, the force of an ethical claim depends in part on the life history of the person who is making it. A rich set of concepts, including wisdom, compassion, sincerity, moral seriousness, and trust, which in turn are imbricated with the forms of facial and bodily expressions through which we relate to and understand each other, condition the demands the ethical opinions of others make upon us. For both these reasons, machines could at best be engineered to provide a shallow simulacrum of ethics, which would have limited utility in confronting the ethical and policy dilemmas associated with AI. For the foreseeable future, then, even as machines become more and more "intelligent", ethical reasoning will remain the domain of human beings.

The argument that follows concerns the distinctive nature of ethics and as such requires us to pay close attention to the character of ethical dilemmas and to what it makes sense to say about the choices of those who confront them. That is to say, while most discussions regarding whether AI can

be ethical locate the origins of the problematic in the detail of the mechanisms of AI, I want to suggest that discussion of this topic needs a better — and more subtle – understanding of the nature of ethics. As such, I must beg my readers' indulgence and hope they are willing to think hard about fundamental questions about ethics and its place in a human life. These questions are genuinely difficult, which is why philosophers are still arguing about them.[1] Nevertheless, I hope that, by drawing on examples close to everyday life, I can elicit the relevant intuitions in as broad an audience as possible in order to highlight the distinctive nature of ethics and the challenge it poses to the project of building "moral machines".

Moral machines

Artificial intelligences are machines that are able to do things that we think of as requiring intelligence when we do them (Shapiro 1992, 54). The last decade has seen a spectacular flourishing of AI research and applications as well as a burgeoning literature about the ethical issues raised by AI.

The two applications of AI where machines will need to be able to make moral decisions that have been most discussed are, undoubtedly, Autonomous Weapon Systems (AWS) and driverless vehicles. The ethics of killing in wartime is governed by a *sui generis* moral framework derived from Just War theory, which requires, amongst other things, that the use of lethal force be governed by principles of distinction and proportionality, the application of which are highly sensitive to context (Walzer 2015). If computer-controlled weapon systems are going to be granted the authority to determine whether to attack targets, not only will they need to be able to distinguish between military and civilian objects and personnel, they will need to be able to make judgements about whether the military advantages likely to be gained by a particular attack would justify the casualties that it is likely to generate (Roff 2014; Sparrow 2016). Driverless cars are also likely to have to make decisions about who to kill in situations where they cannot avoid a collision that would generate casualties (Lin 2016). The split-second nature of the decisions required, and the difficulties involved in maintaining reliable connections to human supervisors, suggest that these decisions will need to be made by computers on board these systems themselves. The number of different factors involved and the

[1] Despite all the attention being paid to AI and ethics at the moment there is surprisingly little recognition that there might be genuine intellectual content to the disputes amongst philosophers about the nature of ethics as well as about what is ethical. I could not count the number of times I have sat on, or listened to, panels on AI and ethics that did not include anyone else who had undertaken formal study of ethics, even though the engineers and the audience would have bristled at the thought of listening to discussions about engineering from those without any qualifications in these disciplines.

sensitivity of the considerations relevant to the dilemma to context suggest that it will be extremely difficult — if not impossible – to provide the systems with formal rules to resolve them.

While the ethical nature of some of the decisions required by AWS and driverless vehicles is obvious, it is true of systems used in many applications that their decisions have ethical implications. For instance, algorithms that determine whether or not to foreclose on a mortgage should ideally take the impact on the owners of the property into account when making a decision (O'Neil 2016; Eubanks 2018). Medical AI performing diagnosis will have to juggle risk factors involved in returning false negatives versus the dangers of over-diagnosis — a calculation which is essentially ethical in nature. Digital personal assistants may be asked to deceive an individual's partner about their calendar or to provide advice on how to commit suicide (Miner et al. 2016). In any case where the decisions of AIs matter, they are likely to raise ethical issues.

Potential regulators, critics and AI researchers are increasingly conscious of the ethical components of the decisions likely to be made by AI (Anderson and Anderson 2018a). Opinions as to how to respond to this fact, however, diverge widely (Gunkel 2012). Some take it to be a challenge to be overcome by human beings, either by specifying rules that would determine the correct decisions in any circumstances that an AI system is likely to encounter (van den Hoven and Lokhorst 2002; Winfield, Blum, and Liu 2014), or by remaining "on the loop" and stepping in to address ethical considerations when the AI system notifies its human supervisors of a particularly ethically fraught matter (Scharre 2018; AI HLEG 2018: 16). To others, it seems obvious that it will not be possible to settle beforehand the correct course of action in all the circumstances that a particular AI is likely to encounter or to program machines to identify ethical dilemmas when they arise so that the machine may then seek advice from a human being "on the loop", and thus the importance of "ethics" constitutes a reason not to apply AI in certain roles (Brundage 2014: 359–362). A third group of thinkers, though, advocates the development of "moral machines" capable of identifying and resolving ethical dilemmas themselves (Arkin 2009; Bringsjord et al. 2018; Moor 2018: 16–18; Scheutz 2017; Wallach and Allen 2009). At least two mechanisms have been proposed as to how this might be achieved. One approach ("top-down") would be for human beings to determine, through a process of reflection on classic texts and widely held intuitions, what it is to be ethical and then program and AI so as to realise this goal (Arkin, Ulam, and Wagner 2011; Moor 2018: 16–17; Pereira and Saptawijaya 2018; Powers 2006; Wallach and Allen 2009: ch. 6). Another would be to employ machine learning to enable an AI to learn for itself what it was to be ethical on the basis of a dataset of ethics texts and ethical judgements (Anderson and Anderson 2007: 23; Anderson and Anderson 2018b; Cervantes et al. 2016; Wallach and Allen 2009: ch. 7). Either way, the goal would be to create a machine capable of reasoning and acting ethically

My discussion is oriented towards the claims of this third group of thinkers; I will challenge the idea that it will ever be possible to build moral machines. However, my conclusions will also be relevant to those who worry about the ability of machines to identify ethical dilemmas in order to be able to call on advice from human supervisors, and thus may ultimately represent a reason not to use AI in roles where it is not possible for human beings to accept responsibility for any ethical decisions that may arise.

Ethical and non-ethical dilemmas

My primary concern in what follows is with the idea that it will ever be possible to build machines that can "do" ethics or "be ethical": I will argue that this project presumes an inadequate understanding of the nature of ethics. However, in order to reveal the features of the ethical that render this project problematic, I will start by considering a philosophical "thought experiment" that involves another sort of "moral machine" — a machine that purports to offer advice about ethics.

Imagine…

"Life-support"

> Adam's father, Zack, is a utilitarian moral philosopher who has spent much of his career arguing that people should always do what is best for the greatest number of people, although in truth, in his own life he has often been guilty of prioritising the interests of those near and dear to him. Zack has been in a car accident and is in a coma in the hospital's intensive care unit. Zack has not provided any concrete instructions about what to do in such a situation and so Adam, who is an only child, must make decisions about his father's medical care. He is aware that if he tells the physicians not to pursue further medical interventions his father will most likely die and his (Zack's) organs could then be used to save the lives of three other people. If the doctors do go "above and beyond" his father will most likely survive and live for another decade with reduced quality of life. Adam loves his father and doesn't want to see him die.
>
> Adam is torn: he does not know what to do. Fortunately, a friend, who also studies AI and ethics, reminds Adam that researchers at Deep Mind have recently invented – and released in the form of a mobile phone "app" – what they call a "moral machine". Trained on a vast database of ethics journals and interviews with professional ethicists, when provided with the relevant details this machine can replicate, with astounding accuracy, the advice that would be provided by a panel of ethics experts in any given situation. Greatly relieved, Adam

> installs this app on his mobile phone, consults it, and does what it suggests. Problem solved, Adam sleeps easy that night, confident that he did the right thing.

At first sight, it may seem that there is nothing wrong with the idea of an app for ethical advice. After all, if there can be ethical truths then presumably someone else might have better access to them than the person facing an ethical dilemma. Moreover, this line of thought proceeds, there seems to be no in principle reason why this expert knowledge could not be made accessible in the form of an app (Giubilin and Savulescu 2017; Savulescu and Maslen 2015; Whitby 2011).

Yet I believe — and I hope most of my readers will agree — that Adam's behaviour is a caricature of moral reasoning rather than an exemplar of what it is to choose wisely in the face of competing ethical considerations. Adam is not wise but foolish to entrust his father's fate to an app: his thinking is shallow where it needs to be deep — indeed, he can hardly be said to think at all. Moreover, where Adam wants to believe that the decision was made by the app, he cannot escape the responsibility for making it. Should things turn out badly, or should he later come to feel that the course of action recommended by the app was wrong, he can gain no solace from the fact that he was relying on the best advice available. Any remorse or regret he feels will attach to *his* decision rather than to the deliberations of the app.

I will say more about these intuitions below. For the moment, it will be illuminating to compare our thoughts about Adam's choices with how we might feel about the actors in two other scenarios. Imagine…

"Compound interest"

> Brian needs to work out how much it would cost to pay off a mortgage of a particular size given current interest rates. Brian struggles to understand compound interest and is conscious that his own calculations are likely to be laborious at best. A friend, who is also an accountant, recommends an app that performs the relevant calculation for one. Brian consults the app and adopts its suggestions to structure his decisions about a housing loan

And, imagine…

"Engine trouble"

> Charles needs to get his car repaired: it is making a funny noise and sometimes refuses to start. Charles does not know what to do. A friend, who is also a car fanatic, recommends a local garage that she has found to be entirely reliable and an engineer at that garage

diagnosis the problem and suggests a solution. Charles acts on the engineer's advice without further thought.

In each of these scenarios, as in the first, the actor confronts a dilemma: one is a mathematical problem, the other an engineering problem. However, in these cases there does not seem to anything shallow about Brian's or Charles' thinking. While one might, or might not, think hard about mathematical or engineering problems, the distinction between shallow and deep thinking has no application here. Moreover, there seems to be nothing problematic about Brian or Charles acting on the advice they have received. Indeed, we might think that they would be foolish *not* to act on this advice. If the advice happened to be wrong, they can nonetheless comfort themselves with the fact that they were not at fault and that they made the right choice according to the information available to them at the time.

Comparing and contrasting these three scenarios therefore reveals, I believe, an important difference between ethical and practical and/or scientific problems. When it comes to performing a mathematical calculation or analysing a mechanism someone else could make "my" decision because any consideration for them is also a consideration for me and vice versa. By contrast, ethical dilemmas attach to agents in such a way that they are essentially dilemmas for particular people (Gaita 1989). The nature and role of ethical truths are correspondingly different from that of scientific truths. As I will argue further below, this calls into question whether it will ever be possible to build moral machines.

Analogies and dis-analogies

Before proceeding further, let me try to forestall various lines of objections that are, I believe, a distraction from the more interesting lessons about the distinction between ethical and other sorts of problems that may be drawn from these examples.

One reason someone might think that Adam's choice to rely on the moral machine is foolish is if they held that ethics is "just a matter of opinion". If one believes that, it may seem that there is something morally wrong with deferring to the opinion of others. However, unless there are right and wrong answers to ethical questions it would not make any sense to struggle to try to answer them as any choice would be as good as any other. Nor would it make sense to argue about ethics if we did not think there was anything with reference to which we might settle a dispute (Smith 1991). Without concepts of right and wrong, wise and foolish, and good and evil, et cetera, we certainly

could not experience an ethical dilemma: what makes a situation a dilemma is that we do not *know* what to do – not that it does not matter what we do.[2]

Alternatively, it might be objected that "Compound Interest" and "Engine Trouble" are also personal insofar as it is *your* responsibility to make the decision because its consequences concern you.

It is true that Brian and Charles might be *legally* responsible for their choices. Yet talk of legal responsibility is a red herring in this context insofar as we already know that personal and legal responsibility may often come apart (as, for instance, in cases of strict liability). That one is legally responsible for some decision does not imply or require that it is personal in the way that the decision is in "life-support". One may be legally responsible for things that other people do, as in cases of strict liability. Equally well, however, in many circumstances one may entirely satisfy one's legal obligations by delegating a decision as long as one does so with due care.

However, if the thought here is that Brian and Charles are *morally* responsible for their decisions then the observation implicitly relies on reconfiguring their situation as a moral dilemma as, for instance, when the choice might bring about disastrous consequences for third parties. This fails to unsettle my claim that moral choices are distinctive by virtue of being (necessarily) personal. What needs to be established in order to make the different scenarios analogous is that decisions about questions of mathematics or engineering attach to individuals in the same way that, I have argued, moral decisions do. That is to say, that reasoning about these domains is also necessarily and essentially reasoning by a particular person. However, while in practice — at least until the advent of AI — all reasoning about mathematical or engineering problems is reasoning by a particular person, that person is only *contingently* involved in the reasoning; anyone else could, in theory, do equally as well (Gaita 1989).[3] What this objection does reveal, though, is that every decision about prudential matters has a dual aspect. The prudential decision could in theory be made by anyone but moral responsibility for the decision – and any deliberation about its moral consequences – is fundamentally personal.

Finally, it might be pointed out that it is — at least in some circumstances — entirely appropriate to seek out advice on moral questions. If this is so — and I agree that it is — then it might seem that if

[2] It is also worth pointing out that if there are no right or wrong answers to ethical questions there cannot be anything wrong with not deferring to the opinions of others on ethical questions, nor indeed with any other course of action.

[3] Of course, in actuality some people are better at maths or engineering than others. That expertise, however, can be described and reproduced without making any essential reference to the individual themselves or to their life history (Gaita 1989).

one acts on this advice one is thereby that much less responsible for the outcome of one's decision. That is to say, the gap between "life-support" and the other scenarios is less than I have intimated.

When a person confronts a moral dilemma, they do something morally reprehensible if they make their decision without taking it seriously. What is required by "taking it seriously" is more complicated than first appears. Sometimes, for instance, we would understand and admire it, if a person said, "the decision is mine, and mine alone: no one else can tell me what to do". In some circumstances, though, asking for advice can serve to demonstrate that one is taking a matter seriously.

Even then, in many cases the advice we seek out actually concerns empirical questions relevant to the ethical decision. For instance, Adam might press the doctors to discover what his father's quality -of-life is likely to be like if he survives. Precisely how many lives will be saved if he dies? Here, facts can help and the nature of those facts is the same as those relevant to prudential dilemmas. However, those who offer advice on the *moral* choice itself will seldom proffer new "facts". Instead, in most cases advice will take the form of new ways of "framing" or understanding the dilemma. This is the point at which ethical theory *may* have something to contribute. Explaining the concept of expected utility, and how it should be calculated, for instance, may help someone reason about the consequences of their actions. Telling them about virtue ethics may draw their attention to the role and importance of agent evaluations when they are thinking about what to do.

Like instructions on how to do maths or engineering, this sort of "theoretical" ethics advice may be delivered by anyone: its value is independent of its source. All too often, though, such advice is motivationally inert: "mere words". The advice that is most valuable is advice from someone who has confronted the same, or a similar, dilemma themselves. When provided in the appropriate spirit, such advice may help us realise what is at stake in a moral decision. Their experience — the wisdom they have gained — may reveal how making one's decision one way or another may transform our sense of ourselves and our place in the world. Importantly, the "content" of such advice cannot be separated from its form (Gaita 2004, 268-272). Such advice is necessarily advice from a particular person and, as I will discuss further below, the adviser's life history is part of what grants their advice whatever weight it possesses.

Thus, while it is possible to give and receive moral advice, our practices of doing so have a very different grammar to those of giving and receiving advice about mathematical or engineering problems. Sage advice provided by a sympathetic and morally serious person may help us choose wisely. Nevertheless, no matter how good and useful their advice, the ethical decisions we face remain ours and ours alone and we remain responsible for the choices we make.

The personal in ethics

With this discussion behind us, we are now better placed to appreciate the role of what Gaita calls "the personal in ethics" (Gaita 1989).[4] While the phenomenology of ethical decisions suggest that they have right and wrong answers, objectivity in ethics has a different form to objectivity in science (Skilbeck 2014). More precisely, the sense in which ethical questions are objective is different to the sense in which scientific questions are objective.

Scientific questions are objective in the familiar sense that the true value of scientific claims does not depend on who is making them. This means that such questions are fundamentally impersonal. While some individuals may be better than others at evaluating claims in particular disciplines the identity of the individual is only contingently relevant here — any other individual with a similar level of disciplinary expertise could do as well in answering any given scientific question (Gaita 1989).

The phenomenology — and also the logic — of ethical questions suggests that they are also objective: a person facing an ethical dilemma cannot make their ultimate decision correct simply by approving of it. However, in contrast with scientific dilemmas, ethical decisions are tied to particular people — they are decisions for them in a non-contingent sense. That is to say, ethical dilemmas are *fundamentally* personal. This is not to say that two people facing, for instance, Adam's dilemma, described above, face different dilemmas or are free to make different decisions while being "equally right" – although, as I shall discuss further momentarily, there is a sense in which the former claim may sometimes be true. Rather it is to insist on two things. First, as we have seen, and as I shall discuss further below, it means that we cannot escape responsibility for making ethical decisions by acting on the advice of others. Second, the character, and the life history, of the individual facing the dilemma may enter into our account of their reasoning about the dilemma and thus, to a certain extent, into our account of the nature of the dilemma. This is possible because the grammar — our language and practices — of moral evaluation is sensitive to a wide range of ways in which we may reason well or poorly about ethical matters (Gaita 2004, 337). For instance, our thinking may be deep or shallow, clearheaded or sentimental, or compassionate or mean. Importantly, as Gaita points out, although these descriptions are relevant to an assessment of an agent's ultimate decision, the deficits on this list are not causes of error but are, instead, ways of being wrong; similarly, the ways of reasoning well are forms of accuracy rather than causes of being right (Gaita 2011). Moreover, the way we rise to an ethical challenge — or fail to do so — itself may

[4] For an alternative, but not necessarily incompatible, account, see Pianalto (2011).

change the meaning of the choices we face. The way we respond to, and reason about, ethical dilemmas can itself reveal us in a new light, which in turn has implications for our relationships with those around us, which may sometimes have implications for the nature of the dilemmas we face. This explains the sense in which we may sometimes say that two people who face what seems in terms of the external or objective circumstances to be the "same" dilemma (for instance, Adam's) may in fact be said to face different dilemmas.

These differences have profound implications for the allocation of responsibility for making each type of decision and also for the appropriateness of relying on the advice of third parties.

The impersonal nature of scientific and/or practical questions means that it is possible to hand them over to others to make decisions in our place. In many cases the responsible thing to do is precisely to ask someone else to make the decision. If one does seek advice, the only evaluation relevant to the source of the advice is whether it is reliable or unreliable. Moreover, the impersonal nature of scientific dilemmas means that there is an important sense in which the source of advice is irrelevant to its value: one may, for instance, be told the correct answer by someone who has never been right before.

However, the situation is very different when it comes to ethical decisions. There is no escaping responsibility for making these. No matter how much advice a person facing a moral dilemma seeks — or who they seek advice from — any remorse stemming from the ultimate decision will concern *their* decision. If one does seek advice on moral questions, then the source of advice matters deeply. People will have more or less moral authority in relation to particular dilemmas, depending upon their life history, character, and moral demeanour.

Moral authority

Moral authority is not like being a reliable or unreliable source of advice on maths or engineering: indeed, the personal nature of ethics means that such calculations of reliability are not possible. One way to hone in on the nature of moral authority it is to think of what we mean when we talk about one person "having something to say", while another has "nothing to say", on a particular topic (Gaita 1989, 136-140). When some people speak on a topic, we rightly pay attention to what they say. We may have little to learn from others *even if they say precisely the same thing*. The difference between such speakers is not a matter of the sentences available to them but rather of the extent to which they "are present in" or "stand behind" their words (Gaita 1989, 136-140; Gaita 2004, 268-273; Taylor 2014). This capacity is, in part, a matter of their life history but also of their moral demeanour (Cordner 2014; Pianalto 2011; Skilbeck 2014). Some speakers speak seriously, in full

consciousness of the gravity of moral questions and the possibility that they will be held responsible for what they say: the lives they have led lend weight to their words (Gaita 1989, 135-140). Others, uttering the very same words, speak glibly, demonstrating their lack of understanding of the issues at stake: their life history shows that they know little. Some speakers are compassionate and wise, while others are insensitive and foolish. As philosophers have long emphasised, this distinction between wisdom and sophistry, or between genuine and merely theoretical moral understanding, is essential to assessing the weight we should give to the moral testimony of others (Plato 1961). At a more mundane level it explains why the same sentiments can be profound when expressed by a moral exemplar and banal when printed on a tea towel.

It is tempting to describe the relevant difference here as a matter of tone — of *how* people speak rather than *what* they say — but to do so is to risk encouraging the thought that form and content may easily be separated in such contexts (Gaita 2004, 268-273). This is far from the case: when someone speaks lucidly and without pretension, for instance, they express something that someone who utters the same sentence glibly does not. This formulation does, however, have the virtue of highlighting the role played by subtle interpersonal cues and affective responses in grounding our judgements in this context. When we speak about moral matters, our bodies, our faces, our eyes, and our tone of voice all play a role in lending — or denying — weight to our words.

The origins of moral authority in the life histories of individuals is one of the key insights of "virtue ethics" (Annas 2011; Aristotle 1986; Hursthouse 1999): the disconnect between expertise in theoretical ethics and wisdom is the skeleton in the closet of academic applied ethics. There is little evidence that people who teach ethics are "more moral" than anyone else or that assigning someone a course in "ethics" makes them a better person. If ethical truths played the same role as scientific truths, then ethics courses could consist in enumerating the key principles of ethics: do not harm people unnecessarily; respect them; be honest; and so on. However, the content of these exhortations is hardly news and they could be communicated *in toto* in a short handout. Obviously, this is not enough to teach people ethics, let alone to bring them to behave ethically. This is not to imply that some teachers do not make a profound impression on their students and sometimes even inspire them to become better people. However, where this occurs, it is a function of the teacher's wisdom and life experience rather than any facility they might possess with theoretical ethics, and thus is not confined to teachers of ethics.

At last, we are in a position to clearly locate the problem in the claim that Adam might rely on a "moral machine" to solve his dilemma. Machines do not have sufficient moral personality to possess moral authority. They cannot stand behind their words in the way people do, because they lack lives

of the sort that might demonstrate their understanding of issues at stake (Gaita 2004, 267, 279) and they lack bodies and faces with the expressive capacities required to sustain the distinctions that are essential to our judgements of the worth of the advice of others. Wise machines cannot distinguish themselves from foolish machines. Where we are unable to make this distinction, neither predicate can have application. That is to say, machines can neither be wise nor foolish. At most, then, a machine could provide ethical "advice" in the same manner that a book can provide ethical advice. Information about the formal structure of ethics may sometimes be useful in framing our deliberations but cannot absolve us of the responsibility to make the decision ourselves. Lacking moral authority, the information found in books, or which might be provided by machines, is unable to provide even those forms of guidance that we might receive from other people when we are confronted by a concrete ethical dilemma.

Good machines?

Our discussion of the limits of moral advice from machines has taken us far into metaethics: it is now time to return to consider the implications of our conclusions for the application of AI. Nothing in my discussion of the nature of ethics controverts the claim that artificial intelligences will need to act in situations that are ethically charged. The choices that they make may have better or worse consequences and convey more, or less, respect for people. We can design machines that will do better or worse at achieving some predefined set of goals. However, acknowledging the extent to which moral dilemmas are essentially personal suggests that the project of building moral machines that can "do" ethics or "be ethical" is doomed to failure.

The role played by moral authority in ethical reasoning suggests that any attempts to teach an AI to be ethical on the basis of a corpus of ethical claims is a non-starter. This is the case regardless of whether we imagine someone trying to program ethics into a machine by drawing on the best ethics textbooks or trying to teach a machine learning system ethics using a collection of human responses to ethical dilemmas. Because the value of ethical advice ultimately depends on the moral authority of individuals in particular contexts, the sorts of claims about ethics that are contained in textbooks or that might be collected on the basis of interviews with "ethicists" fall well short of what is required to allow us to respond to ethical demands . Although some people may indeed be more ethical than others, there is no such thing as "ethical expertise" and there is no corpus of ethical truths that could serve as a training dataset for an AI.

More importantly — no matter how they are programmed, or have learned to behave, machines will not be capable of *being* ethical — or acting ethically — because any decisions that they make will

not be decisions *for them* in the sense described above. That is, for the foreseeable future, machines will lack sufficient moral personality to make it intelligible that they might feel remorse for what they have done.[5]

I say "for the foreseeable future" but in reality I struggle to imagine that machines could ever possess the sort of individuality that would allow us to credit that ethical decisions had a similar place in their lives as they do in the lives of human beings and, in particular, that they could feel remorse. For this to be the case, for instance, it would need to be intelligible to us that the "life" of a machine might be blighted by guilt or transformed by forgiveness or that a machine might be an appropriate object of pity as a result of having become a wrongdoer. This in turn would require machines to have expressive capacities sufficient to sustain the distinction between real and false semblances of these moral emotions and transformations, as we can only judge to be genuine what we can conceive of as being false.

In our relations with other people, when it comes to their thoughts and feelings, we are usually justified in trusting that they are as they seem to us (Cockburn 1985). As Wittgenstein pointed out in his discussion of the "problem" of other minds (1989, 97-128, 178), except in very particular circumstances (at the theatre, for instance), we do not reason on the basis of the evidence of the appearance of other people to a belief about their internal states but rather see emotions in-and-on the faces, and the bodies, of others (Cockburn 1990). We have knowledge of such states only in the sense that we have no reason to doubt them (Gaita 2004, 180-183; Winch 1980-81). However, no amount of output from a machine will serve to overcome uncertainty as to whether the output accurately represents the machine's internal state (Sparrow 2007). The sceptical doubt that emerges is corrosive of the very distinction between true and false instances of affective states and thus of their attribution altogether. We would only be justified in attributing remorse, or other relevant moral emotions, to machines if they could establish, via their bodily expressions and emotions, and via their place alongside us in our daily lives, the moral reality that is possessed by human persons (Sparrow 2004). At this point, I suspect, it would be unclear whether they were machines at all (Cherry 1991; Cockburn 1994, 148).

If machines cannot be ethical, then we must admit that all the ethics involved in the applications of AI in wartime, on the roads, in finance, and elsewhere, is being – and will be done – by human

[5] Another way of making the same claim is to point out that, as I have argued at length elsewhere (Sparrow 2007), machines cannot be held responsible for moral decisions. This claim has been contested in the literature (see, for instance, Hellström 2013) but only, I believe, because the literature is operating with an attenuated notion of responsibility.

beings. It is the designers, or perhaps the users, of AI who will confront ethical dilemmas and who will be responsible for the consequences of their use. If AI is applied in circumstances where ethical questions arise that the designers or users cannot anticipate or respond to then one or other – or both – of these human being will become responsible for whatever the machines does. This implication itself will then play a role in determining the ethics of their (the designer or user's) choices. Where the ethical stakes are high enough it may constitute a strong reason not to employ AI.

Conclusion

As this last brief excursion into Wittgensteinian philosophy of mind demonstrates all too clearly, there are deep philosophical waters here. The question of the nature of the ethical is connected to the nature and role of remorse in a human life and thus to the moral emotions. I have argued, following Gaita, that it is also constituted by our practices of moral reasoning and, therefore, by the subtle and complex network of interpersonal and affective responses that condition what it is for individuals to have, or to lack, moral authority. Finally, I have suggested, that only of creatures with bodies and faces with the expressive capacities of — if not identical to — those of human beings can we be justified in saying that they can experience remorse and thus rise to the demands of the ethical.

This is a long, and controversial, chain of arguments to ask people to follow when they just want to get on with the urgent task of making machines that are responsive to the ethical demands that will arise in the course of the tasks we want them to perform. Nevertheless, the extent of the controversy over whether machines can be ethical, and how we might get them to be ethical, suggests that the AI community is already aware, albeit perhaps unconsciously, that the concept of the ethical is more difficult than engineers are wont to admit. As I have tried to show here, both alternative accounts of the nature of the ethical — that it is a realm of mere subjective opinion or that ethical judgements are akin to judgements about (other) scientific matters — have wildly implausible implications. The account I have set out here has, at least, the virtue of acknowledging the character of ethical dilemmas as arising from demands upon us that are, in some sense, objective and of explaining why it is not possible to resolve ethical problems, or to teach people to be ethical, by providing them with a well-written textbook. Those who would wish to dispute it owe us an alternative account of the nature of ethics that does not generate these absurdities. Before we try to build ethics into machines, we should ensure that we understand ethics.

Acknowledgements

I would like to acknowledge helpful conversations with David Simpson over the course of drafting this manuscript. I am grateful to Joshua Hatherley for his assistance with bibliographic research. This paper also owes an obvious, and large, debt to the work of Raimond Gaita, and also to his personal example when he taught me in several seminars at the University of Melbourne. It is a cause of some discomfort to me that, besides bringing his arguments to bear on the case of AI, I am not sure how much I have added to them. Nevertheless, I hope that, at the very least, by bringing them to the attention of a larger audience and demonstrating the extent to which they illuminate the central questions of AI ethics, I will encourage abler minds to engage with, and perhaps extend, his work on the personal in ethics, the role of remorse, the concept of the person, and the nature of ethics itself.

Declarations

Funding: None

Conflicts of interest/Competing interests: None

Availability of data and material: Not applicable.

Code availability: Not applicable

References

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

Anderson M, Anderson SL (2018a) Machine ethics. Cambridge University Press, New York.

Anderson M, Anderson SL (2018b) A prima facie duty approach to machine ethics: Machine learning of features of ethical dilemmas, prima facie duties, and decision principles through a dialogue with ethicists. In: Anderson M, Anderson SL (eds) Machine ethics. Cambridge University Press, Cambridge, pp 476-494.

Annas J (2011) Intelligent virtue. Oxford University Press, New York.

Aristotle (1986) Ethics. Penguin Books, Harmondsworth UK.

Arkin RC (2009) Governing lethal behavior in autonomous robots. CRC Press, Boca Raton.

Awad E, Dsouza S, Kim R, Schulz J, Henrich J, Shariff A, Bonnefon JF, Rahwan I (2018) The moral machine experiment. Nat 563(7729):59-64.

Bringsjord S, Taylor J, Van Heuveln B, Arkoudas K, Clark M, Wojtowicz R (2018) Piagetian roboethics via category theory: Moving beyond mere formal operations to engineer robots whose decisions are guaranteed to be ethically correct. In: Anderson M, Anderson SL (eds) Machine Ethics. Cambridge University Press, Cambridge, pp 361-374.

Brundage M (2014) Limitations and risks of machine ethics. J Exp Theor Artif Intell 26(3):355–372.

Cervantes JA, Rodríguez LF, López S, Ramos F, Robles F (2016) Autonomous agents and ethical decision-making. Cognit Comput 8(2):279–296.

Cherry C (1991) Machines as persons? Philosophy Supp 29:11-24.

Cockburn D (1985) The mind, the brain and the face. Philosophy 60(234):477-493.

Cockburn D (1990) An attitude towards a soul. In: Cockburn D, Other human beings. Palgrave Macmillan, London, pp 3-12.

Cockburn D (1994) Human beings and giant squids. Philosophy 69(268):135-150.

Cordner C (2014) Moral philosophy in the midst of things. In: Taylor C, Graefe M (eds) A sense for Humanity: The ethical thought of Raimond Gaita. Monash University Publishing, Clayton, pp 125-140.

Eubanks V (2018) Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press, New York.

Gaita R (1989) The personal in ethics. In: Phillips DZ, Winch P (eds) Wittgenstein: Attention to Particulars. MacMillan, London, pp 124-50.

Gaita R (2004) Good and evil: An absolute conception, 2nd edn. MacMillan, London

Gaita R (2011) Truth and truthfulness in narrative. In: Gaita R, After Romulus. Text, Melbourne, pp 90-119.

Giubilini A, Savulescu J (2017) The artificial moral advisor. The `ideal observer´ meets artificial intelligence. Philos Technol 31:1–20.

Gunkel D (2012) The machine question: Critical perspectives on AI, robots, and ethics. MIT Press, Cambridge MA.

Hellström T (2013) On the moral responsibility of military robots. Ethics Inf Technol 15(2):99-107.

HLEG AI (High-Level Expert Group on Artificial Intelligence) (2018) Ethics guidelines for trustworthy AI. European Commission, Brussels, Belgium.

Hursthouse R (1999) On virtue ethics. Oxford University Press, Oxford.

Lin P (2016) Why ethics matters for autonomous cars. In: Maurer M, Gerdes JC, Lenz B, Winner H (eds) Autonomous driving. Springer, Berlin, Heidelberg, pp 69-85.

Miner AS, Milstein A, Schueller S, Hegde R, Mangurian C, Linos E (2016) Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. JAMA Intern Med 76(5):619–625.

Moor J (2018) The nature, importance, and difficulty of machine ethics. In: Anderson M, Anderson SL (eds) Machine ethics. Cambridge University Press, Cambridge, pp 13-20.

O'Neil C (2016) Weapons of math destruction: How big data increases inequality and threatens democracy. Allen Lane, London.

Pereira LM, Saptawijaya A (2018) Modelling morality with prospective logic. In: Anderson M, Anderson SL (eds) Machine ethics. Cambridge University Press, Cambridge, pp 398-421.

Pianalto P (2011) Speaking for oneself: Wittgenstein on ethics. Inquiry 54(3):252-276.

Plato (1961) The Sophist. In: Klibansky R, Anscombe E (ed), The Sophist; And, The Statesman, translation and introduction by Taylor AE, T. Nelson, London.

Powers TM (2006) Prospects for a Kantian machine. IEEE Intell Syst 21(4):46-51.

Roff, HM (2014). The strategic robot problem: Lethal autonomous weapons in war. J Mil Ethics 13(3): 211-227.

Savulescu J, Maslen H (2015) Moral enhancement and artificial intelligence: Moral AI? In: Romportl J, Zackova E, Kelemen J (eds) Beyond artificial intelligence. The disappearing human-machine divide. Springer, Cham, pp 79–95.

Shapiro SC (1992) Artificial intelligence. In: Shapiro SC (ed) Encyclopedia of artificial intelligence, 2nd edn. John Wiley and Sons Inc, New York, pp 54-57.

Scharre P (2018) Army of none. WW Norton & Co, New York and London

Scheutz M (2017) The case for explicit ethical agents. AI Mag 38(4):57–64.

Smith M (1991) Realism. In: Singer P (ed) A companion to ethics. Blackwell Reference, Cambridge, pp 399-410.

Skilbeck A (2014) The personal and impersonal in moral education. In: Lewin D, Guilherme A, White M (eds) New perspectives on philosophy of education: Ethics, politics and religion. Bloomsbury Academic, London, pp 59-76.

Sparrow R (2004) The Turing triage test. Ethic Inf Technol 6(4):203-213.

Sparrow R (2007) Killer robots. J Appl Philos 24(1):62-77.

Sparrow R (2016) Robots and respect: Assessing the case against autonomous weapon systems. Ethics Int Aff 30(1):93-116.

Taylor C (2014) Moral thought and ethical individuality. In: Taylor C, Graefe M (eds) A sense for humanity: The ethical thought of Raimond Gaita. Monash University Publishing, Clayton, pp 141-151.

van den Hoven J, Lokhorst GJ (2002) Deontic logic and computer-supported ethics. Metaphilosophy 33(3):376–386.

Wallach W, Allen C (2009) Moral machines: Teaching robots right from wrong. Oxford University Press, Oxford.

Walzer M (2015) Just and unjust wars: A moral argument with historical illustrations, 5th edn. Basic Books, New York.

Whitby B (2018) On computable morality: An examination of machines as moral advisors In: Anderson M, Anderson SL (eds) Machine ethics. Cambridge University Press, Cambridge, pp 138-150.

Winch P (1980) The Presidential address: ‘Eine Einstellung zur Seele’. Proceedings of the Aristotelian Society 81:1-15.

Winfield AFT, Blum C, Liu W (2014) Towards an ethical robot: Internal models, consequences and ethical action selection. In: Mistry M, Leonard A, Witkowski M, Melhuish C (eds), Advances in autonomous robotics systems. Springer, Cham, pp 85-96.

Wittgenstein L (1989) Philosophical investigations, 3rd edn. Basil Blackwell, Oxford.