On the Possibility of Moral Machines: A Reply to Robert Sparrow

Dane Leigh Gogoshin, University of Helsinki, Practical Philosophy Department, RADAR Research Group (Finland)
International Conference on Computer Ethics: Philosophical Enquiry (CEPE) 2023, Chicago, IL

Keywords:machine morality, artificial moral agency, ethics of AI

Abstract

Robert Sparrow has recently argued (Sparrow 2020) against the possibility of building moral machines (or AIs). He urges the community to better understand the nature of ethics before undertaking this endeavor. He suggests, following philosopher Raimond Gaita, that ethics is much more complex, more personal, and in many ways more subjective than is generally acknowledged in machine/AI ethics, and he deploys original thought experiments to show this. In this paper, while granting Sparrow's overall contention – that ethics is too complex to *presuppose* the possibility of moral machines – by way of challenging certain of Sparrow's premises and triggering different intuitions from his thought experiments, defend a cautiously optimistic view of the possibility of moral machines/AIs.

In a recent paper, "Why machines cannot be moral" (2020), Robert Sparrow argues that the nature of ethics precludes the possibility of moral machines (or AIs). He concedes, however, that philosophers are still debating what that nature is. This observation, I think, should encourage rather than discourage attempts to build moral machines. Still, I grant Sparrow's overall contention – that ethical life is surely about much more than a fixed, objective set of rules and ties into many dynamic, subjective, relational, and personal features of human life – and I agree that it is naive (if not flippant) to *presuppose* that machines can be moral. My aim in this paper is to provide reasons to be, if not optimistic, at least more open to the possibility of machine morality than Sparrow would have it. I will explore and challenge certain of Sparrow's pessimistic premises and provide a positive basis for my cautious optimism.

My first challenge is to Sparrow's claim concerning ethical dilemmas and expertise. By their very nature, ethical dilemmas lack a clearcut right or wrong solution. There can only be, at best, right or wrong answers with respect to the particular actors involved and concern particular dispositions, personal histories, cares, values, etc. This means that the problem of generating good solutions to ethical dilemmas is not unique to AIs. Sparrow constructs a thought experiment, "Life-support," in order to trigger the intuition that we should not rely on AIs for ethical advice regarding ethical dilemmas. By offering a variation on this theme, I evoke the opposite intuition – that an AI might be the only possible ethical consultant in Life-support type cases.

My second challenge falls out of the first. It is important to distinguish the issue of the nature of ethics (which Sparrow discusses) from the problem of moral motivation (which he does not). In quotidian life, *doing* the right thing is a bigger problem than *knowing* it. Dilemmas are merely a subset of morally relevant actions. The bulk of moral mistakes arise from practical issues like weakness of will, impulsiveness, selfishness, and inattention. Granted, (lack of) knowledge may play a role in these failings; we may not know how to prioritize our values or cares in such a way that the right course of action is salient to us. But this still seems to boil down

to a practical rather than an epistemic problem. It isn't yet clear *that* or *how* machine/AI morality is (particularly) threatened at this practical level.

My third challenge falls out of the second. Sparrow remarks that remorse is an essential component of moral agency. Indeed, the capacity for remorse seems particularly essential to a moral disposition. Imagined guilt at wrongdoing often informs the way we choose to act and when we act badly, remorse serves to punish and reform our behavior. We tend to trust in others' good will only when we observe this capacity in them. But is this the only basis for such trust? At a deeper level, the capacity for remorse may be linked to our capacity to empathize. Our capacity to empathize may, in turn, be at the root of moral knowledge and moral motivation. That this is the way human morality *might* work says nothing of how and whether we might build moral machines. Perhaps, i.e., it is possible to detect moral significance and be moved by it without the capacity to feel or express remorse.

I will support these three challenges and give weight to my optimism by way of the modified Life-support thought experiment mentioned earlier. In the original, Sparrow describes an end-of-life decision that a son, Adam, must make about his father. Adam is torn about what to do and consults an AI ethical advisor-app. The intuition Adam's choice sparks is, of course, one of moral criticism. We don't think that Adam should hand off this kind of decision to an AI. So consider my version of Life 2
support instead.

Suppose that an AI, Hal, spends enough time with (being used by, observing, recording, processing, analyzing) its human companion, Adam, such that it can accurately predict most of Adam's choices and behavior. When asked how Adam would feel about something with moral implications, Hal can guess Adam's actual response with high accuracy. Naturally, when Adam must make an end-of-life decision on behalf of his father, Adam becomes terribly distressed and hits a wall of indecision which he just cannot get past. Yet the clock is ticking; he must make a decision and soon. Adam thinks to flip a coin, but that seems flippant, so he thinks to ask someone. He ponders on who knew his father best and realizes that it is he alone. The logical next question is who knows Adam best? Adam doesn't really think of Hal as a person, but Hal, in deducing (from loads of data he's collected on Adam) Adam's need, offers its own consulting services. Upon reflection, Adam agrees that Hal is best positioned to help him make the decision. Adam realizes that there is no perfect answer and that he's going to have to live with the decision either way. Whether it's Hal or another person (if there were a viable human candidate), the responsibility for the outcome is his alone. But Hal is uniquely positioned to predict what Adam should do – in light of (its analysis of) Adam's values, cares, character, and personal history, and his father's values.

This version of Life-support, unlike the original, does not, I think, inspire moral criticism. Moreover, the responsibility for the decision stays with Adam. But this is true irrespective of Hal's status as an AI. Yet this does not tell us that an AI can be an ethical expert. Still, if Hal can give Adam sound advice about something of significant ethical importance, it tells us something significant about what a good ethical consultant might be. But this, on its own, does not tell us what it is to be ethical. After all, Hal doesn't necessarily possess any of its own values and cares, against which it might measure Adam's values or cares. Hal can, at best, help Adam make a decision which best reflects *Adam's* values. It seems reasonable to expect of a human ethical advisor that they have their own set of moral values. However, in line with what Sparrow observes, an advisor is the *right* ethical advisor for *you* because they have moral standing in *your* eyes. Moral standing, it could be said, derives from what values one holds and whether and to what extent one has lived by those values.

In this light, the best ethical consultant is one, first and foremost, whose values the consultee shares. So far, Hal could still qualify. To be helpful in difficult cases, the consultant ought to shed light on moral

considerations the consultee hadn't been aware of or sufficiently sensitive to beforehand, so should potentially possess additional moral values or a different weighting of moral values to the consultee's. The consultee (Adam, in this case) attends to the consultant's moral considerations (can be influenced by them) because, in Adam's eyes, the consultant is respectable and trustworthy (due to the latter's moral standing). Could/should Adam view *Hal* as someone with moral standing, if not on the basis previously provided, on another (i.e., could one take the moral values of a stranger with a different cultural outlook seriously?)? Is moral standing necessary to attend further to or to further moral considerations? I might see an important moral message printed on a tea towel and, despite knowing nothing about who put it there or why (probably just to sell the tea towel), I might ponder its meaning and conclude that it is something I should take seriously and try to apply in my own relationships.

With these considerations in place, Hal remains in the race as an ethical advisor; what about as a moral agent? Provided that Hal is able to abide by its own ethical advice – something which seems likelier for machines than for creatures of flesh and blood – then I don't yet see a reason to categorically exclude Hal-type AIs from the possible set of future moral agents.

Still, you might think that machines can never author their own values and that moral agency 4 requires this capacity. But this is part of a well known philosophical problem for human beings as well, whose values originate at least in part, from without. The standard take (Frankfurt 1971; Watson 1975) on whether we really possess our own moral identity is whether we act in a way that accords with our (reflectively endorsed) values. But we can still ask where these values originate and whether we can ever meaningfully claim them. In the end, we are left with uncertainty about the moral self. There is a bigger issue, here, though and it comes down to the nature of morality. Since the jury is still out on that, though, I do not yet see any obvious red lights (theoretically speaking[1]) on the road to attempting to build moral machines and AIs.

[1] Though I see the practical red flags pointed out in Bryson (2018) and Wynsberghe & Robbins (2019).

References

Bryson, J. J. (2018). Patiency Is Not a Virtue: the Design of Intelligent Systems and Systems of Ethics. *Ethics Inf. Technol. 20*, 15–26. doi:10.1007/s10676-018-9448-6

Frankfurt, H. (1971). Freedom of the Will and the Concept of a Person. *The Journal of Philosophy, 68*(1): 5–20. doi:10.2307/2024717

Sparrow, R. (2020). Sparrow, R. (2020). Why machines cannot be moral. *AI & Soc 36*, 685–693 (2021). https://doi.org/10.1007/s00146-020-01132-6

Wynsberghe, A.R., & Robbins, S. (2019). Critiquing the Reasons for Making Artificial Moral Agents. *Science and Engineering Ethics, 25,* 719 - 735.

Watson, G. (1975). Free Agency. *The Journal of Philosophy, 72*(8): 205–220. doi:10.2307/2024703 6