Other minds, other intelligences: the problem of attributing agency to machines

Other Minds, Other Intelligences: The Problem of Attributing Agency to Machines

Sven Nyholm, Eindhoven University of Technology

In his characteristically thought-provoking and readable article[1], John Harris relates the philosophical "problem of other minds" to the interaction between human beings, with our reasonably intelligent minds, and entities with artificial intelligence (AI), which might one day become super intelligent. Harris considers what we might call bilateral mind-reading: human beings reading the minds of artificial intelligences, on one hand, and artificial intelligences reading the minds of human beings, on the other hand. Moreover, Harris also considers the social and ethical challenges posed by such bilateral mind-reading, with particular emphasis on potentially super intelligent AI. In this commentary, I respond to some of the main questions Harris raises in his piece. I also contrast and compare some of the main claims Harris makes along the way with other recent interventions in the discussion of the ethics of human-AI interaction.

Whose Minds?

Harris discusses what he sometimes simply calls "super intelligent AI" and at other times refers to as "AI persons" or "beings with Super AI". In what follows, I will primarily talk about "robots" because I think that what are commonly called robots are most likely to be intuitively interpreted by human beings as having minds. Of course, it can also happen that we spontaneously attribute minds to other types of technology with AI (e.g. a computer or a smart phone). But the questions discussed by Harris will be most pressing in relation to different kinds of robots equipped with AI.

By a robot, I mean any machine with some degree of functional autonomy and some degree of AI, equipped with sensors and actuators, which can interact with its environment in a way that allows the machine to perform tasks otherwise typically performed by humans.[2] This is a fairly broad definition. It allows a wide range of machines to count as robots: anything from a self-driving car (a type of machine Harris also discussed in a recent article of his in this journal[3]) to a more humanoid robot like "Sophia" from Hanson Robotics who, in

2017, was awarded honorary citizenship by Saudi Arabia) or robots used in hospitals or care settings, such as "Kaspar," a robot used in experimental care of autistic children[4].

As mentioned above, my focus here is on robots with AI – rather than simply AI or AI persons because I think that these are the types of entities to which people, now and in the future, are most likely to attribute minds. I should note, however, that other machines with AI may be just as likely as robots to attempt to read the minds of human beings. Your computer, smartphone, or the algorithms behind the social media websites you use might be as likely to try to read your mind as any robot.[5] This could be so, for example, because they are programmed to track your interests and possible purchases, as is the case with targeted advertising on social media websites such as Facebook. Human beings, however, are less likely, I take it, to attribute mental states to the algorithms of Facebook than to any robot with which they might interact, whether that robot is Roomba the vacuum cleaning robot or Sophia the humanoid robot.[6]

Is There a Philosophical Problem Here That is Not "Embarrassingly Artificial"?

When Harris discusses the traditional "problem of other minds" at the beginning of his piece, he refers specifically to the problem of whether other people actually have minds at all, and if so, how we can know them. Harris calls this an "embarrassingly artificial" philosophers' problem.[7] This raises the question of whether a problem of other minds when applied to artificial intelligences might not also be an embarrassingly artificial philosophers' problem, rather than a problem with any real-world relevance. Regarding this question, I suspect that mean-spirited critics of Harris' discussion might respond by saying that the primary problem of Harris' focus – namely, bilateral mind-reading between human beings and super intelligence AI persons – is also an embarrassingly artificial philosophers' problem. The reason that mean-spirited critics might respond in this way is that super intelligent AI persons seem unlikely to be created anytime soon.

If I were to come to Harris' defense here, I would say that whether or not super-intelligent AI will ever come into existence, the philosophical questions Harris discusses are, nevertheless, absolutely fascinating ones on which to reflect. Perhaps this is why Harris also discusses in his article what he sometimes calls "the minds of those who never lived," thereby indicating that he is treating what he is discussing as a piece of speculation, rather than something currently facing us in the real world.

However, after offering this defense of Harris, I would quickly move to arguing that there is a much more pressing social and ethical problem here, one that already has real-world relevance and, therefore, deserves our attention. Specifically, there are social and ethical problems created by a common tendency to attribute minds (and, therefore attempts to read them) to, robots and other machines in situations where we lack evidence to say that they have any minds, especially not humanlike minds. With our social minds and conceptual schemes, it is almost impossible for us *not* to interpret a great many things we experience as if they have minds. As Harris notes, Homer already wrote about mind-reading nearly 3000 years ago. And those studying the evolution of human minds and concepts argue that reading the minds of others is a human characteristic that evolved long ago in order for us to adapt to the highly socially interactive types of lives we lead.[8] What this means is that we interact with robots with AI using human minds and concepts that evolved long before any robots and AI ever existed. We interact with AI systems with minds that were specifically adapted to interpret others using mind reading as a common feature. This being the case, it should be no surprise that people tend to attribute minds to – and try to read – the minds of any machines, or other entities, that give even the most minimal evidence of having some sort of minds -- whether humanlike minds or not.

Notably, technical experts on robotics and AI are beginning to worry that people in general are overly naïve about the minds of robots (or the lack thereof) and about the potential for technology companies to develop products they purport to have minds; but, are instead deceptive devices with nothing resembling minds. For example, a team of engineers at Columbia University recently developed a robotic arm that (they claimed) after 35 hours of training developed a basic form of "self-image".[9] Several technology news outlets enthusiastically reported on this study as an important step towards robotic self-consciousness.[10] This prompted Noel Sharkey, founder of the Foundation for Responsible Robotics and roboticist-turned-robot-ethicist, to appear on Sky News to, as he put it, pour "cold water" on the claim that the Columbia engineers had developed "a near sentient robot arm with self-awareness".[11] Similarly, developers of sex robots and other humanoid robots (like Sophia, mentioned above) are also making claims for their products that some worry will deceive the general public to falsely believe that these robots have more advanced minds than they really do. Here, too, Sharkey has been a vocal critique of the claims that technology companies have made on behalf these "show robots" as he calls them.[12]

Looking at how language reflects thinking about AI, consider the way that people in both academic and public discussion talk about self-driving cars with relation to crashes. Experts and laypeople alike quickly fall into talking about what a car should "decide to do" if it faces a risky scenario in which it cannot avoid getting into a potentially deadly accident. Self-driving cars, it is often said, will sometimes "make life-or-death decisions" and therefore need to be equipped with ethical algorithms that will help them to make the right decisions.[13] This way of talking and thinking about self-driving cars is another illustration of how most people find it hard not to conceive of robots and other machines in anthropomorphizing terms that attribute agency and minds to these machines. Again the question arises whether there is something ethically problematic about talking and thinking about machines in such ways. Some even worry that this might give rise to so-called responsibility gaps, since they think that while robots and other AI systems can make decisions and exercise basic agency, they cannot (yet) be morally responsible for their decisions in the ways that human beings can be.[14]

Self-driving cars might also potentially be viewed as a case of a technology that needs to be able to read the human minds with which they interact. In traffic involving a mixture of human-driven cars, bike riders, and pedestrians, self-driving cars have to be able to calculate the movement of others in order to be able to adjust their positions to avoid accidents. This can be interpreted as a need for self-driving cars to be able to predict what decisions people are *likely* to make based on their outward behavior, so that the self-driving cars can be "one step ahead" and not simply react to how people are already behaving.[15] That can be seen as a form of mind-reading.

In general, my point in this section is this: Harris is certainly right that it is philosophically intriguing to consider how and whether humans and super intelligent "AI persons" might be able to read each other's minds. However, it is a much more pressing ethical and social problem that people and much less advanced robots (and other AI systems) are already trying to read each other's minds. This is a more pressing problem since this might lead to – as may already be the case – people being deceived or harmed, (perhaps without their awareness), or their privacy being invaded by companies with AI systems trying to predict their intentions, emotions, and other states of mind.

Harris and "Robot Rights"

Harris also brings up the controversial issue of robot rights and our responsibilities. He posits:

> What has been almost entirely lost it seems to me, in the debate about possible dangers posed by AL, are real and planned, or at least envisioned, dangers we imagine we will be able to pose to them, the beings with Super AI, and which current debate supposes . . . that we will be justified in posing to them.[16]

Harris then goes on to suggest, a few sections later, that:

> If we create Superintelligent AI, we will neither be able to own them . . . nor enslave them, nor have sex with them without their consent, nor be able to destroy them without just and sufficient cause. We may hope they will think the same of us. . .[17]

Again I would like relate this pair of claims to the case of robots. As I said above, robots are the types of machines with AI to which people will be most likely to attribute minds. But they are also, I now wish to suggest, the types of machines to which people are most likely to be willing to extend moral consideration.[18] Moreover, if we focus on the case of robots with AI (whether it is a more modest form of AI or "super intelligent AI"), then the claim that the issue of whether we are justified in treating machines with AI in any manner we please is not a question that has been "altogether lost" in the discussion. Instead, there is a small, but growing debate about what David Gunkel calls "robot rights": the question of whether, in Gunkel's terms, robots can and should have rights.[19]

This is another issue where I would like to suggest that while Harris' discussion of the moral status of super intelligent AI persons is captivating in the abstract, a more pressing question, with greater real-world relevance, is whether we should extend any kind of moral consideration to robots that actually exist already or that will exist in the near future. Regarding this issue, some commentators, such as the roboticist Joanna Bryson, are outspoken in their view that robots ought not to be given moral consideration. Bryson argues that like any piece of technology, a robot is a tool and should only be treated as such. Besides, a robot will be a tool somebody owns – something we can buy and sell. For that reason, Bryson

argues, "robots should be slaves."[20] Bryson adds that, because robots should be slaves or tools we can buy or sell, there is a moral imperative not to create any robots that would appear to merit serious moral consideration.

Others, like Gunkel – and, along with him, writers like Mark Coeckelbergh and John Danaher – claim that we should take the prospect of robot rights, or moral consideration for robots, seriously. According to the "social-relational" perspective favored by both Gunkel and Coeckelbergh, the most pressing moral question is not whether robots have minds or whether they can suffer, nor whether they can talk or think.[21] Rather, the most pressing moral question, which can help to determine whether we should extend moral consideration to robots, is instead the question of how we interact with robots and what role(s) they play in our communities or in the relationships we form with the robots around us. According to Gunkel and Coeckelbergh, if we introduce robots into our homes or, perhaps, into our workplaces (which might be hospitals or care home settings), this can give us moral reason to treat these robots with moral consideration, because this will be appropriate given the social-relational settings in which we interact with these robots. We should not, to use Bryson's terminology, have "slaves" in our homes, whether they are humans or robots with AI.

Danaher, in turn, also argues that we should not place too much weight on mind-reading when we consider what kinds of relationships and interaction we should have with robots.[22] Rather, we should use "behavioral" or "performative" criteria in determining whether it makes sense to treat robots and other AI systems with moral consideration. On this view, if a robot is able to behave in a way that is similar to a being with moral status – or in a way that is similar to a friend or companion – then we ought morally to treat that robot in a way that gives it moral consideration, or that treats it like a friend or companion. In defense of this view, Danaher argues that behavior and performances are also all we have to go on in the case of other humans (and animals) to whom we attribute moral status. To be sure, we infer mental states from the ways people behave and based on what they say. But the only thing we can be sure of is how they behave and perform around us. Therefore, we should also treat robots with AI based on that basis and on what kinds of minds they have. Or so Danaher argues.

Others, like Kate Darling, take an even more pragmatic perspective on how we should interact with robots. If we are uncomfortable with the prospect of treating robots badly – as some people have been in experimental studies Darling herself has conducted – then we have reason to treat the robots well, for our own sake, Darling suggests.[23] Or if our treating robots

brutally might lead us to treat people brutally, this would be another reason to treat robots in a way that suggests moral consideration – for the sake of the humans we wish to avoid treating brutally. Similarly, Kathleen Richardson, who is the leader of the campaign against sex robots, also argues that we should avoid creating and interacting with sex robots that reinforce negative stereotypes about women, or that in other ways promote the objectification of human sex partners.[24] This is another pragmatic argument against treating robots in certain ways, based on the consequences this might have for us as humans, whether or not the robots themselves merit moral consideration for their own sake.

I bring up these various recent suggestions about the moral status (or lack thereof) of robots in order to make a point (similar to the one I made about mind-reading above), that it seems more pressing to focus on how we ought morally to interact with robots already among us – or perhaps soon to be among us – whether or not they possess any kind of super intelligence. How people treat robots of different kinds is already a controversial matter, and one worth careful consideration; i.e. whether or not the robots among us have any kinds of minds, let alone minds at all resembling our own.

I agree with Harris, however, that our tendency to engage in mind-reading is highly relevant to this issue of how we should, or should not, treat robots and other machines with AI. In my view, this tendency is highly relevant, not only because robots with AI might eventually come to have minds that deserve respect. It is also relevant because this mind-reading tendency of ours shapes the types of interactions with robots that come naturally to us and thereby affects what we can reasonably expect from people. For example, Bryson's suggestion that we should view all machines as tools and all robots as servants is psychologically difficult for humans to put into actual practice. People will find themselves resisting treating robots that appear to have intelligence as mere tools that might as well be regarded as slaves.[25] Social minds like ours will feel too strong a pull to anthropomorphize many robots in order for it to be realistic to expect us to regard all robots as mere tools – nor should we.[26]

---

[1] Harris J, 'Reading the Minds of Those Who Never Lived. Enhanced Beings: The Social and Ethical Challenges Posed by Super Intelligent AI and Reasonably Intelligent Humans', *Cambridge Quarterly of Healthcare Ethics . . .*

[2] Gunkel D, Robot Rights. Cambridge, Massachusetts: MIT Press, 2018

[3] Harris J, Who Owns My Autonomous Vehicle? Ethics and Responsibility in Artificial and Human Intelligence. The Cambridge Quarterly of Health Care Ethics, 2018;27(4):599-609

[4] Robins B, Dautenhahn K, & Dubowski, J. Does appearance matter in the interaction of children with autism with a humanoid robot? Interaction Studies 2006;7(3):479–52.

[5] Frischmann B & Selinger E, Re-engineering Humanity, Cambridge: Cambridge University Press, 2018
[6] I mention Roomba here because some people treat Roomba in anthropomorphizing ways, such as giving it names and displaying gratitude to and concern for Roomba. See Scheutz M, The Inherent Dangers in Unidirectional Emotional Bonds between Humans and Social Robots. In Lin P, Abney K, & Jenkins R eds. Robot Ethics: The Social and Ethical Implications of Robotics. Cambridge, MA: MIT Press, 2012:205-221
[7] See note 1, Harris
[8] See, for instance, Dennett D, From Bacteria to Bach and Back Again, Cambridge MA: Harvard University Press 2017 and Heyes C, Cognitive Gadgets, Oxford: Oxford University Press, 2018
[9] Kwiatkowski R & Lipson H, Task-agnostic Self-modeling Machines. Science Robotics 4(26),2019: eaau9354
[10] Some samples: http://blogs.discovermagazine.com/d-brief/2019/01/30/self-aware-robot-arm-learning/#.XGAAjTF3E2x , https://www.forbes.com/sites/bridaineparnell/2019/01/31/robot-know-thyself-engineers-build-a-robotic-arm-that-can-imagine-its-own-self-image/#4cb306f54ee3
[11] https://twitter.com/RespRobotics/status/1091317102009634817 , and https://techxplore.com/news/2019-01-closer-self-aware-machinesengineers-robot.html
[12] See, for instance, Sharkey N, Mama Mia It's Sophia: A Show Robot or Dangerous Platform to Mislead? Forbes Magazine, November 17, 2018, available online at: https://www.forbes.com/sites/noelsharkey/2018/11/17/mama-mia-its-sophia-a-show-robot-or-dangerous-platform-to-mislead/#5cc17e007ac9
[13] For an overview of the ethics literature on self-driving cars, see Nyholm S, The Ethics of Crashes with Self-driving Cars: A Roadmap, I-II, Philosophy Compass, 2018;13(7):e12507-e12506
[14] E.g. J. Danaher J, Robots, Law and the Retribution Gap, Ethics and Information Technology 2016;18(4):299–309
[15] For a 2015 Ted talk in which Google's Chris Urmson claims that Google cars are able to do this, follow this link: https://www.ted.com/talks/chris_urmson_how_a_driverless_car_sees_the_road
[16] Harris J, 'Reading the Minds of Those Who Never Lived. Enhanced Beings: The Social and Ethical Challenges Posed by Super Intelligent AI and Reasonably Intelligent Humans', *Cambridge Quarterly* Harris, page numbers
[17] Harris J, 'Reading the Minds of Those Who Never Lived. Enhanced Beings: The Social and Ethical Challenges Posed by Super Intelligent AI and Reasonably Intelligent Humans', *Cambridge Quarterly* Harris, page numbers. Incidentally, Lily Frank and I discuss the topic of robot sex and consent in our 'Robot Sex and Consent: Is Consent to Sex between a Robot and a Human Conceivable, Possible, and Desirable?' *Artificial Intelligence and Law*, 2017;25(3):305–323
[18] See, for instance, Nijssen SRR, Müller, BCN, van Baaren RB, & Paulus, M. Saving the Robot or the Human? Robots who Feel Deserve Moral Care, Social Cognition 2019;37(1):41-52
[19] Gunkel D, Robot Rights, Cambridge, MA: MIT Press, 2018
[20] Bryson J, Robots Should be Slaves. In Y Wilks, ed. Close engagements with artificial companions: key social, psychological, ethical and design issues. Natural Language Processing, vol. 8, Amsterdam: John Benjamins Publishing Company, 2010:63-74.
[21] Gunkel D, *Robot Rights*. Coeckelbergh M, Robot rights? Towards a social-relational justification of moral consideration, Ethics and Information Technology 2010;12(3): 209–221
[22] Danaher J, 'The Philosophical Case for Robot Friendships', Journal of Posthuman Studies, in press. See also Danaher's blog post 'Ethical Behaviourism in the Age of the Robot', available at: https://philosophicaldisquisitions.blogspot.com/2017/12/ethical-behaviourism-in-age-of-robot.html
[23] Darling K, Who's Johnny? Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy in Lin P, Abney K, & Jenkins R. eds. Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence, New York: Oxford University Press 2016, 173-91
[24] Richardson K. The Asymmetrical 'Relationship': Parallels between Prostitution and the Development of Sex Robots. SIGCAS Computers & Society 2015;45(3):290-293
[25] Cf. Gunkel G, *Robot Rights*
[26] Many thanks to Tomi Kushner for her feedback and helpful suggestions.