Robots as persons? Implications for moral education

Michael J Reiss
UCL Institute of Education, London m.reiss@ucl.ac.uk

Abstract

At present there is a clear distinction between robots and persons. In this article I explore the possibility that this distinction may not hold in perpetuity, as some robots attain personhood. I argue that personhood is an emergent property in both the development of individuals and the evolution of life, that personhood may not require a carbon-based existence, and that, given that robots are being made with ever greater powers of cognition, at some point these powers of cognition may reach the point at which we need to start talking of robots as having minds and being persons. This will have implications for how we treat robots, for how we design robots and for how we understand ourselves and other creatures. There are also implications for moral education that may need to be taken seriously.

Robots as persons? Implications for moral education

This article takes seriously the request of the organisers of this symposium to consider moral education two decades from now. As Jim Conroy put it in his original brief for the symposium to his Fellow *Journal of Moral Education* Trustees: "what does the future of moral education hold in store for, say, the next two decades? What is going to be its agenda and its problematics? What will be the hot topics debated in the *Journal of Moral Education* in 2038?".

I am, of course, well aware that attempts to look far head are risky. There is no doubt that some of what I consider below will turn out to have been wide of the mark. But there is a good side to futurology, as to counterfactualism. Both can help us to step back from quotidian specifics; both encourage imagination and reflection. There are benefits as well as harms in immersing oneself in utopian possibilities.

That having been said, the particular possibility I examine here – that we should examine the possibility of robots becoming persons – is one that many may dismiss, almost out-of-hand (but see Gunkel (2018), who explores the issue of robot rights). I will develop my argument below, but in essence it is as follows: personhood is an emergent property in both development and evolution; personhood may not require a carbon-based existence; robots are being made with ever greater powers of cognition; at some point these powers of cognition will reach the point at which we need to start talking of robots as having minds and being persons; this will have implications for how we treat robots, for how we design robots and for how we understand ourselves and other creatures.

What is a robot?

I am using the word 'robot' as the convenient everyday term for an entity made by humans that manifests some degree of responsiveness to changes in its environment. Unsurprisingly, there is a large literature about what precisely is meant by 'robots' (e.g. Lin, Abney & Bekey, 2014). For example, standard definitions generally talk about robots being able to move independently. I am all in favour of independent movement but just as we don't disqualify a quadriplegic or someone suffering from locked-in syndrome from being considered a human being, so we would not want to exclude from consideration of the ethical issues in this article an artificial entity incapable of movement. There is also the question about whether software itself should be considered or whether such software needs some sort of hardware (its 'embodiment', its materiality) to open up the possibility of moral consideration.

I exclude from further consideration the possible fruits of synthetic biology (e.g. if scientists manage in a laboratory to grow sentient animals without them being the offspring of other sentient beings, such sentient beings being the result of the long process of biological evolution that has given rise to all life to date). It seems obvious to me that such entities deserve precisely the same moral consideration as other sentient beings (cf. the genetic fallacy).

Who is a person?

If there is a large literature on the meaning of the term 'robot', that is as nothing compared to the literature on the meaning of the term 'person'. I belong to the camp that does not see an equation of 'person' with 'human being' – where the latter term is understood as a member of the species *Homo sapiens*. The argument that only members of the species *Homo sapiens* can be persons suffers, it seems to me, from a number of problems. For a start, imagine that Earth becomes colonised by hyper-intelligent creatures (not members of *Homo sapiens*) from outer space (or the subterranean depths, for that matter – of negligible likelihood but I mention it only to emphasise that from where such creatures come is irrelevant) who reject our claim that we too are persons, and thus worthy of moral consideration, on the grounds that only members of their species can be persons. Then, less fancifully, the 'only humans are persons' argument means that no other species already present on Earth can be persons. I find this argument a difficult one to defend when I consider such domestic animals as cats, cows, dogs and pigs, not to mention such great apes as chimpanzees and gorillas – cf. The Great Ape Project (Cavalieri & Singer, 1993).

Personhood can be understood as a bundle of mutually interrelating characteristics. A person has a non-trivial degree of self-awareness (aka self-consciousness) and manifests at least a certain degree of rationality and moral awareness. Crucially, for moral education, a person therefore warrants ethical consideration and typically has certain ethical duties / responsibilities, for all that there are certain standard objections to the reciprocation between 'warranting ethical consideration' and 'having certain ethical duties / responsibilities' (people when asleep or unconscious, babies, those with dementia, etc).

There is a related question as to whether personhood can be ascribed as a result of an entity's will (as a Kantian is likely to consider) or its interests (which broadens the scope of moral consideration). This question is an important one but it is a secondary question in the sense that my argument below does not require one or other of these positions to be rejected.

The argument from evolution

Now we begin to get to the heart of the issue as to whether it makes sense to imagine that robots could be persons. I am an evolutionary biologist by background. There are two central issues relevant to the question at hand: (a) the likelihood that life (which is organic) evolved from inorganic, inanimate precursors; (b) the evolution, over time, of one or more species that exhibit personhood from species that do not.

That life arose from inorganic precursors is held to be the case by the overwhelming majority of scientists. It is, of course possible that life on Earth was brought here from elsewhere, for instance in bacterial spores, but this merely moves the issue of the origin(s) of life from the Earth to elsewhere. If one doesn't hold that life arose from inorganic

precursors, one identifies as a creationist or possibly an intelligent designer. In any rate, one relies on a *deus ex machina* argument and falls outwith mainstream science.

Once life had arisen, the standard scientific account is that over a very long period of time (of the order of four thousand million years) evolution led to the present diversity of living organisms. Importantly, this includes *Homo sapiens*, with our manifestation(s) of personhood. So, the mainstream scientific view is that personhood is a natural result of evolution, whether persons are restricted to humanity or, as considered above, a bit more widely distributed amongst organisms.

Precisely why the attributes that constitute personhood arose through evolution is still the matter of some debate. There is no doubt that personhood is a feature of our minds. Humans have unusually large brains for our size – in the jargon of evolutionary biologists, we have a high encephalisation quotient (a sophisticated measure that takes account of the fact that some species that are much larger than us have heavier brains than we do and of the fact that brain mass, both across and within species, does not scale linearly with body mass). The probable reason for our large brains is that while energetically expensive (our brain uses about 20% of the energy from our food intake despite constituting only 2% of our body mass), they enable us to do well socially. Doing well socially has probably been of great importance for hundreds of thousands of years (Humphrey, 1976). It's also possible that the benefits that our having large brains currently afford, such as success in obtaining food, may have played an important, even predominant, role in evolution (DeCasien et al., 2017).

The argument from development

To a biologist, the question as to when each human life begins doesn't have a clear-cut answer. If we think of the particular sperm and egg that in fusing gave rise to each of us, each of these two cells was alive. At the point of fertilisation (conception) there is an important step; in particular, the genotype (genetic constitution) of the resultant zygote is laid down (as both sperm and eggs differ in their genetic constitution). Subsequent important steps (in *Homo sapiens*, along with other mammals) include implantation and birth. Subsequently, if all goes well, the new-born develops and grows, eventually into adulthood.

The point is that as far as the above understanding of a person is concerned (in the section 'Who is a person?'), the fertilised egg that results from the fusion of a sperm with an egg does not immediately constitute a person. Even a new-born baby, some nine months post-conception is hardly a person in the sense that I have used the term above. Personhood is therefore something into which each of us grew. There is a direction of travel from non-person to person. This is not, of course, to say that an entity that is not yet a person deserves no regard. There can be a whole range of reasons why one might wish to ascribe regard, protection or even rights to such entities, not least because, in the case of a human six or twelve months post-conception, there is a good chance that they will develop into persons – cf. the literature on 'the argument from potential' (Parfit, 1984; Reichlin, 1997).

The argument from chemistry

I now turn to something that is more speculative – the question as to whether persons must be carbon-based. The simplest answer is that we don't know. So far as we do know, all persons that exist and have ever existed are carbon-based. But it may be that silicon-based entities can be persons – silicon being both in the same group in the Periodic Table as carbon and the primary constituent of 'silicon chips', the integrated circuits that are the basis of all computing devices. The belief that life needs to be carbon-based is sometimes referred to as 'carbon chauvinism' and is not infrequently rejected by science fiction writers. Douglas Adams' Hooloovoo resemble a super-intelligent shade of blue. One was seen in a prism for Zaphod Beeblebrox's Presidential address by refracting into a free-standing prism (Adams, 1979).

Returning to reality, if it is the case that for some reason as yet unknown to us, personhood requires an entity being carbon-based, then the main way in which robots – which are silicon- rather than carbon-based – are being developed precludes personhood. However, just as today's soft robots have bodies that are based on carbon, unlike the majority of robots whose bodies are based on rigid materials, such as metals, it is not impossible that in future we may have robots with minds, not just bodies, that are carbon-based.

It is also the case that there is considerable interest, often financed by the military, in cyborgs – entities that combine robotics with an organic life form – think Dark Vader from *Star Wars* or The Borg from *Star Trek*. The internet is full of reports of attempts to bioengineer various species of animals for human benefits. These include sharks (to be used underwater), mice and airborn insects. Indeed, if it is the case that only carbon-based entities can be valid recipients of moral considerations then we need to look at such cyborgs. We also need to consider humans, such as Kevin Warwick, who are investigating what are sometimes called 'direct interfaces' between computer systems and the human nervous system. Warwick himself has for years been experimenting with implants that directly connect to his own nervous system. For example, back in 2002 Warwick had an electrode array surgically implanted into his arm. As a result, Warwick was able to control a robot arm at a distance, via the internet.

Of course, science fiction is far more interested in whether androids (a robot that resembles a human being) are persons – think Karel Čapek's robots in *R.U.R. (Rossum's Universal Robots)*, the replicants in *Blade Runner*, Data in *Star Trek* and various characters in such films as *Under the Skin* and *Ex Machina*. Indeed, such works of fiction are excellent at helping the viewer to consider whether or not only humans have moral agency and whether there are any important moral considerations that differ between humans and androids (e.g. Decker & Eberl, 2008; Luokkala, 2019).

The argument from history

One of the features of human history has been, by and large, a broadening of our understanding of who are persons. Females, slaves, children and foreigners have all been brought into the moral arena. Indeed, most of us now wince at how certain categories of

humans were deemed less worthy of moral considerations. We do not need to go to the extremes of history – Josef Mengele at Auschwitz, Stalin's Holodomor in Ukraine; we need only recall Justice Oliver Wendell Holmes' "Three Generations of Imbeciles Are Enough" and The Tuskegee Study of Untreated Syphilis in the Negro Male (Reverby, 2009).

The argument from history is that the greater consideration given to other humans has been driven by sentiment, by reason and by law. The importance of human rationality in our ethical thinking was made with particular clarity by the moral philosopher Peter Singer in his book *The Expanding Circle* (Singer, 1981). Singer argued that altruism began as a drive to protect one's kin and those in one's community but has developed over time into a consciously chosen ethic with an expanding circle of moral concern (cf. Reiss, 2019).

The argument from theology

Most religions privilege humans over the rest of creation. So, in Judaism, Christianity and Sufism, humans are created in the image and likeness of God – *Imago Dei*. However, there have been substantial moves within the Abrahamic faiths to come to a deeper understanding of the purpose of God's non-human creation. In part such moves were fuelled by more explicit awareness of ecological considerations (e.g. Page, 1996). The net result of such thinking has been to soften the binary distinction between humans and the rest of creation. For a start, there is much that humans can learn from other creatures: "Go to the ant, thou sluggard; consider her ways, and be wise: Which having no guide, overseer, or ruler, Provideth her meat in the summer, and gathereth her food in the harvest" (*Prov* 6:6-8). When we see animals in the wild or watch today's nature documentaries, we can admire the beauty, the skills and the dispositions (which in humans would be considered virtues) of countless species.

Mention can also be made of the growing interest in some theological circles of the possibility of panpsychism (Leidenhag, 2019). Long seen as a core belief within Vedantic religions, panpsychists see mentality as fully natural, as fundamental to the universe, but not reducible to the physical. The growing interest in panpsychism in the science-and-religion field is partly due to increasing explorations of the relevance of quantum theory to theology. While there are a range of views about this (e.g. Saunders, 2002; Leidenhag, 2019), at the very least such remarkable, yet well-established, physical phenomena as quantum entanglement ('spooky action at a distance' – to cite Einstein), in which, in certain circumstances, measurements on one particle (e.g. to determine its spin) *instantaneously* cause changes in one or more other, distant particles, raise questions about our understanding of the fundamentals of our universe (including the nature of causation and of time). There are a number of competing interpretations among physicists as to what is going on but, for our purposes, what is of interest is that this seems like evidence for deep connections between entities in a way that is at least consonant with theological understandings of the universe that see something mysterious shared between all entities. Such phenomena as quantum entanglement give some support to the notion, as expressed in the Upanishads, that Brahma (ultimate reality) is pure consciousness (Deutsch, 1969/1973). Humans may not be as distinct from the rest of the cosmos as we generally presume.

Implications for moral education

It might be thought that I am writing all this as purely a theoretical exercise. I am used to writing theoretical exercises (I have authored strategy documents while working in higher education senior management) – but this is not one of them. The history of Artificial Intelligence (AI) has been one of computers/robots sometimes taking longer and sometimes taking less time to do things that were once regarded as the preserve of humans. As is well known, we now have robots that can play chess and Go better than anyone and can undertake surgical operations, recognise faces, compose music and paint better than the great majority of us can. Increasingly, numbers of us rely on AI to an increasing extent, for our shopping, our driving and our healthcare, inter alia.

All this is a long way from personhood. Nevertheless, if I am right, there is nothing in evolutionary biology and developmental biology to cause us to conclude that robots will not assume personhood. The lesson from chemistry is less clear and theologians will no doubt remain divided on the issue (as they not infrequently are …). Here, I consider it enough to have established that it is not a waste of time to consider the possibility. At this point I ought perhaps to admit that I doubt robots will gain personhood within twenty years. Of course, I might be wrong but by extension of Moore's Law (the doubling of computer power every two years, despite such computers costing less) I suspect it will take longer. However, it might happen in my lifetime.

It is also relevant to note that the ways in which computers are increasingly being programmed to learn (machine learning, neural networks) means that they are learning in ways that are more similar to humans than was previously the case. (Roughly speaking, these new approaches are like trial-and-error, associative and inductive learning rather than 'logical', deductive learning. As is well known, the vast majority of human learning operates via such non-logical methods.) It may be that such similarities between human and computer cognition increases the likelihood that computers will develop personhood.

What I want now to do is to outline some possible implications for moral education should robots indeed manifest at some point the behaviours that would lead reasonable numbers of people to conclude that they are persons. (I phrase it thus as, of course, there is nothing utterly illogical about any one of us adopting solipsism and concluding that we alone are a person.)

Moral education, of course, is not just about school education. However, school education is important for morality – it complements what we learn in our families as we grow up and provides a site both where we educate ourselves through practice (how to behave towards those who do not like us, when to retaliate, when to forgive, and so on) and where – and this is a distinctive feature of schooling – there are professionals (teachers), older than us, part of whose job is to get us to consider the reasons for our actions and to reflect on the desirability of what we do or want to do. Just as schools are places where we think through what is meant by property, ownership and theft (e.g. Heinz in the Kohlberg dilemma), for example, so schools will be places where teaching about personhood will now include

robots. There will also be implications for schools in Design and Technology lessons where, one day, more thought will need to be given to making robots and controlling their actions than is currently necessary. An analogy is that the issue of animal dissection; it is not that issues of animal dissection arise only in school but that they have a particular relevance to school biology lessons (Reiss, 2017).

The core issue that moral education will need to consider if and when robots are accepted as persons is how we should treat them (including how we should educate them – we will have duties towards robots in regards of their education as both parents and the state have in regard to human offspring). Perhaps the most useful analogy is with slavery (understood as the ownership and exploitation of humans for the benefit of other humans). The analogy of slavery is useful for a number of reasons. First, while no analogy is perfect, this one is a close one (Gunkel, 2018). At present we give virtually no thought as to whether it is right / acceptable (there is no need for the issue to be debated within a rights framework though it can be) for us to use robots entirely for our ends – indeed the notion of 'ends' for robots is presently meaningless and almost no one seriously questions whether it is acceptable / right that we do not give robots time off and either cannibalise or dispose of them when they become outdated or broken beyond repair. In much the same way, there were (indeed, still are, given the widespread existence of contemporary slavery) those who minimised the time that slaves had off work and either resisted or thought it utterly inappropriate for slaves to have any of the protections afforded to those who of their own free will entered into contacts of employment.

Secondly, we have good historical records, particularly in the West over the last two hundred years, of what accompanied campaigns to end slavery (and related issues such as the establishment of 'universal' suffrage). Humans being what we are, there was a great variety of responses to the institution of slavery from active support, through toleration to unease to active attempts at prohibition. We can envisage a similar diversity of responses once people begin seriously, rather than only in science fiction, to argue that at least some robots are persons. Furthermore, the reasons for resistance to the abolition of slavery will prove illuminating. In many cases, particularly in the short-term and at a local level (i.e. ignoring the devastating and very long-lasting effects of slavery on the peoples and places from where slaves came), the economic advantages of slavery were very considerable. So, it is with robots. We can also anticipate that some of the arguments used against the ending of human slavery (that slaves are not people in the sense that 'we' are – being primitive, lazy, unable to benefit from education, etc) will be applied to robots, buttressed by the fact that robots manifestly are not humans (cf. the way in which some maintained that slaves were sub-human).

At the same time, there will be issues to do with robots where I think the institution of human slavery may not be the best parallel. Consider, in particular, that there are and will be (to an even greater extent) an exceptional diversity of robots. There may indeed be replicants as mentioned above – robots so human-like that it is difficult to distinguish them from humans. Do you know what happens when a tortoise flips upside down (and see the ending of the Director's Cut of *Blade Runner*)? But there will also continue to be machines that barely merit the term 'robot' – machines that automatically clean our floors and mow our lawns (cf. Aldous Huxley's Epsilon Semi-morons). The analogy here is with meat eating.

Virtually none of us (unless under the most extreme circumstances – Géricault's *Raft of the Medusa* and comparable non-fictional examples, e.g. Uruguayan Air Force Flight 571 in 1972) is capable of eating, let alone desires to eat, human meat; very few of us would choose to eat meat from other primates (I do realise the accuracy of such generalisations depends on what one got used to eating as a child); in the West, most people cannot imagine wanting to eat meat from dogs and grubs as food are considered disgusting. But once we get to farm animals there is much greater diversity and vegetarians and vegans have more clear-cut principles. Similarly, I can imagine that there will be some people who will feel that it is acceptable to use certain forms of robots but not others and there will be some people who will have more absolute prohibitions (or permissions).

There are literatures on post-colonialism, feminism and our use of animals in addition to mainstream moral and political philosophy that will help navigate these issues. And I could multiply examples of moral questions we will face – for example, just as many eschew prejudice towards and stereotyping of people of others genders, ethnicities, nationalities, sexualities and so forth, will we need to extend that to robots? But, as I intimated at the start of this piece, there are risks in looking too far ahead. It is enough for me, I hope, to have established the claim that we may need to take seriously the possibility that some robots will indeed be persons, and will be seen to be such. To the moral educator this has implications that are possibly greater than any faced in human history.

Finally, it is worth pointing out that the above has a normative element – what *should* we do if and when (some) robots *are* persons. The field of social robotics take a somewhat different approach. It notes that today's robots already engage with humans in socially meaningful ways – as trainers, therapists, mediators, caregivers and companions (Dumouchel & Damiano, 2016/2017). Even a decade ago there were instances of soldiers who were almost inconsolable at the thought that damaged robots with whom (? with which) they had worked on the battlefield might not be repairable. Peter Singer recounts how "One EOD [Explosive Ordnance Disposal] soldier brought in a robot for repairs with tears in his eyes and asked the repair shop if it could put 'Scooby-Doo' back together. Despite being assured that he would get a new robot, the soldier remained inconsolable. He only wanted Scooby-Doo" (Hsu, 2019). Another soldier ran 50 m under machine gun fire to rescue a robot that had been knocked out by enemy fire (Singer, 2009). On the battlefield, robots have been promoted, given Purple Heart awards and received a military funeral. As robots become more extensively used in education, social care and other fields, the lesson is clear – increasing numbers of robots will be seen to be persons, whether or not philosophers, other academics and professionals consider they are or are not.

References

Adams, D. (1979) *The Hitchhiker's Guide to the Galaxy*. London: Pan Books.

Cavalieri, P. & Singer, P. (Eds) (1993) *The Great Ape Project: Equality beyond humanity – towards a new equality*. London: Sage.

DeCasien, A. R., Williams, S. A. & Higham, J. P. (2017) Primate brain size is predicted by diet but not sociality. *Nature Ecology & Evolution*, 1, 0112.

Decker, K. S. & Eberl, J. T. (Eds) (2008) *Star Trek and Philosophy: The wrath of Kant*. Chicago, IL: Open Court.
Deutsch, E. (1969/1973) *Advaita Vedanta: A philosophical reconstruction*. Honolulu: University of Hawaii Press.
Dumouchel, P. & Damiano, L. (2016/2017) *Living with Robots*. Cambridge, MA: Harvard University Press.
Gunkel, D. J. (2018) *Robot Rights*. Cambridge, MA: MIT Press.
Hsu, (2009) Real soldiers love their robot brethren. *Live Science*, 21 May. Available at https://www.livescience.com/5432-real-soldiers-love-robot-brethren.html.
Humphrey, N. (1976). The social function of intellect. In: *Growing Points in Ethology*, Bateson, P. P. G. & Hinde, R. A. (Eds). Cambridge: Cambridge University Press, pp. 303-317.
Leidenhag, J. (2019) The revival of panpsychism and its relevance for the science-religion dialogue, *Theology and Science*, 17(1), 90-106.
Lin, P., Abney, K. & Bekey, G. A. (Eds) (2014) *Robot Ethics: The ethical and social implications of robotics*. Cambridge, MA: MIT Press.
Luokkala, B. B. (2019) *Exploring Science Through Science Fiction, 2nd edn*. Cham: Springer.
Page, R. (1996) *God and the Web of Creation*. London: SCM.
Parfit, D. (1984) *Reasons and Persons*. Oxford: Oxford University Press.
Reichlin, M. (1997) The argument from potential: a reappraisal. *Bioethics*, 11(1), 1-23.
Reiss, M. J. (2017) A framework within which to determine how we should use animals in science education. In: *Animals and Science Education: Ethics, curriculum and pedagogy*, Mueller, M. P., Tippins, D. J. & Stewart, A. J. (Eds). Dordrecht: Springer, pp. 243-259.
Reiss, M. J. (2019) Science, religion and ethics: The Boyle Lecture 2019, *Zygon*, 54(3), 793-807.
Reverby, S. M. (2009) *Examining Tuskegee: The infamous syphilis study and its legacy*. Chapel Hill, NC: University of North Carolina Press.
Saunders, N. (2002) *Divine Action and Modern Science*. Cambridge: Cambridge University Press.
Singer, P. (1981) *The Expanding Circle: Ethics, evolution, and moral progress*. Princeton: Princeton University Press.
Singer, P. W. (2009) *Wired for War: The Robotics revolution and conflict in the 21st century*. New York: Penguin.

Michael J. Reiss is Professor of Science Education at UCL Institute of Education, University College London, a Fellow of the Academy of Social Sciences, President of the International Society for Science and Religion and a member of the Nuffield Council on Bioethics. The former Director of Education at the Royal Society, he has written extensively about curricula, pedagogy and assessment in education.