
Article:
Sharkey, A. (2017) Can robots be responsible moral agents? And why should we care? Connection Science, 29 (3). pp. 210-216. ISSN 0954-0091

https://doi.org/10.1080/09540091.2017.1313815

Can robots be responsible moral agents? And why should we care?

Amanda Sharkey

Department of Computer Science, University of Sheffield, Sheffield, UK

Department of Computer Science, Regent Court, Portobello Rd, University of Sheffield, S1 4DP Email: a.sharkey@shef.ac.uk

Can robots be moral agents? And why should we care?

Principle: Humans, not robots, are responsible agents. Robots should be designed; operated as far as is practicable to comply with existing laws & fundamental rights & freedoms, including privacy.

This principle highlights the need for humans to accept responsibility for robot behaviour and in that it is commendable. However it raises further questions about legal and moral responsibility. The issues considered here are (i) the reasons for assuming that humans and not robots are responsible agents (ii) whether it is sufficient to design robots to comply with existing laws and human rights and (iii) the implications, for robot deployment, of the assumption that robots are not morally responsible.

Keywords: robot, moral agent, embodiment

Introduction

At first glance, this statement or principle seems convincing. It makes sense to insist that humans and not robots are responsible agents. It usefully reminds us of the limited abilities of robots, and provides a helpful antidote to the strong claims and warnings sometimes made about them. We should not offload blame for mistakes or bad consequences onto robots. Emphasising human responsibility for robot behaviour should help to restrict the possible harmful uses to which robots could be put. It also makes sense to suggest that robots should be designed and operated to comply with existing laws and fundamental rights and freedoms: it is difficult to imagine anyone suggesting otherwise.

But, on further consideration, it becomes apparent that the statement does not give any justification for saying that humans and not robots are responsible agents, nor does it provide any guidance about where and when robots should be used, or the consequences that follow from assuming that robots are not responsible agents. The statement raises a number of issues that deserve further discussion. These include

important questions about legal responsibility that are not discussed here. The issues that will be considered are (a) What are the reasons for assuming that humans and not robots are responsible agents? (b) Is it sufficient to design robots to comply with existing laws and fundamental rights and freedoms? And (c) If robots are not responsible agents, should this limit the roles they are given and the situations in which they are deployed?

(a) What are the reasons for assuming that humans and not robots are responsible agents?

Aside from legal responsibility, it is possible to identify two reasons for this assumption. The first is based on the difference between biological and mechanical machines, and the biological basis of morality. The second is to do with the need for society to accept responsibility for the artefacts that humans have produced. We consider both of these in turn.

(i) Biological machines versus Mechanical machines: The principle states that humans and not robots are responsible agents: a statement that can be interpreted as implying that robots should not be viewed as moral agents. This view is not universally held: some (e.g. Asaro, 2006; Wallach and Allen, 2009) have argued that moral agency should be viewed more as a continuum, and others (e.g. Sullins, 2006) have claimed that robots could be full moral agents (if certain conditions were met). Nonetheless, claims that robots are not moral agents, and a belief that they are unlikely to become so in the near future, can be grounded in arguments about the biological basis for morality.

Patricia Churchland (2011) discusses the basis for morality in living beings, and argues that the basis for caring about others lies in the neurochemistry of attachment and bonding in mammals. She explains that it is grounded in the extension of self-maintenance and avoidance of pain in mammals to their immediate kin. Neuropeptics, oxytocin and arginine vasopressin underlie mammals' extension of self-maintenance and avoidance of pain to their immediate kin. Humans and other mammals feel anxious about their own well-being and that of those to whom they are attached. As well as attachment and empathy for others, humans and other mammals develop more complex social relationships, and are able to understand and predict the actions of others. They also internalise social practices, and experience 'social pain' triggered by separation, exclusion or disapproval. As a consequence, humans have an intrinsic sense of justice. The same is largely the case for non-human mammals. Bekoff and Pierce (2009) provide many examples of evidence of a moral sense of justice in mammals. For example, capuchin monkeys working for treats seemed offended and would refuse to cooperate further when they saw that another monkey was given a more desirable reward for the same work (Brosnan and de Waal, 2003).

Of course there are differences between humans and mammals in terms of morality. Animals are more often described as moral patients than moral agents: but the implication here is that the capacity to be a moral patient is necessary for the development of moral agency. Caring about oneself, and extending that care to others, forms the basis for the development of morality in humans.

By contrast, robots are not concerned about their own self-preservation or avoidance of pain, let alone the pain of others. In part, this can be explained by means

of arguing that they are not truly embodied, in the way that a living creature is. Parts of a robot could be removed from a robot's body without it suffering any pain or anxiety, let alone it being concerned about damage or pain to a family member or to a human. A living body is an integrated autopoeietic entity (Maturana and Varela, 1980) in a way that a man-made machine is not. Of course, it can be argued that the robot could be programmed to behave as if it cared about its own preservation or that of others, but this is only possible through human intervention. We return to a further discussion of the feasibility of programming morality below.

(ii) Societal responsibility: Johnson and Miller (2008) argue that robots, and other computational artefacts, are not full moral agents because they "are not ever completely independent from their human designers". They describe them as 'human-tethered' artefacts, and argue that responsibility cannot be offloaded onto the artefacts themselves since the behaviours and outputs of robots and computer systems necessarily depend on human designers and developers. A useful example that they consider is that of a door opener. A person who opens the door for someone carrying a package can be viewed as having performed a positive moral act. But if the door were opened by means of a sensor that detects the approach of a person, the mechanical door opener would not be considered to have performed a praiseworthy act. Related arguments about a lack of independence from human designers have been made in the past based on the way in which robots, unlike living machines, can never be considered to be fully embodied, since they have always required human intervention and involvement in their development (Sharkey and Ziemke, 2001). The point here is that robots, and their underlying control systems, depend on human intervention. The robots may be 'set loose' to make unpredictable decisions, but the decision to allow them to do so is a

human and societal one. Any decisions made by the robot will still depend on their initial design. Even if the robots are 'trained' or 'evolved' to make decisions, their training or fitness regime will still have involved human intervention at some point, and it is imperative that human responsibility is assumed and recognised. Johnson (2006) makes a useful distinction between moral agents and moral entities, and places robots and computer artefacts in the second category. Moral entities include the artefact designer, the artefact, and the artefact user, and moral responsibility cannot be offloaded onto the artefact itself.

(b) Is it sufficient to design robots to comply with existing laws and fundamental rights and freedoms, including privacy?

A major problem with the suggestion that robots should be designed to comply with existing laws and fundamental rights and freedoms, and the reason that it is not sufficient to do so, is that existing laws and human rights have not been formulated with technological developments such as robotics in mind. Although it is important to avoid unnecessary multiplication of ethical and regulatory instruments, there does seem to be a need to reconsider existing legislation in the light of such developments. For example, robots pose a particular risk to privacy, particularly when they are designed to appear as friends and companions and as a result are welcomed into our homes and intimate surroundings. There are many questions here to be answered about the extent to which the information they have access to will be accessible to others, and as yet little legislation to address this. Ethical concerns have been expressed about the risks of leaving vulnerable older people in the near-exclusive 'care' of robots, with little human contact, (e.g. Sharkey and Sharkey, 2012; Sparrow and Sparrow 2006), but the

Human Rights Act does not provide any explicit protection from such a situation. Similar concerns have been raised about leaving children in the 'care' of robots to the extent that their attachments to humans are compromised (Sharkey and Sharkey 2010) but again there is no legislation or rights that explicitly prevent such a possibility, other than that associated with child neglect. There is an urgent need for something like a digital bill of rights to ensure that there is some protection from the situations that could arise if humans place robots in positions of power over humans.

As well as concerns about whether existing legislation provides enough protection for humans from robot deployments, there is another set of reasons for believing that designing robots to comply with existing laws and fundamental rights and freedoms is not sufficient. These reasons are related to the earlier discussions about whether robots can be considered to be moral agents. Robots can be programmed with sets of rules that determine their behaviour, but this does not mean that they are capable of making moral decisions. When humans make decisions about how to act in social situations, they have to do more than follow a set of rules, or laws. They make decisions based on a moral understanding of what it is appropriate or inappropriate for them to do. They are sensitive to feedback about their decisions and their outcomes, and can reflect on it and adjust their future decision-making.

There have been discussions about the extent to which robots can be programmed or trained to make the right moral decisions in social situations. Arkin (2009), for example, has argued that in a battlefield situation, robot soldiers could be programmed to follow a set of rules that would result in more ethical behaviour than that sometimes shown by human soldiers in the heat of battle. His claim is that human soldiers can act badly as the result of their emotions – for instance being motivated by

revenge to carry out war crimes. A robot on the other hand would not respond emotionally and could be programmed, by means of an 'ethical governor', to evaluate actions before carrying them out, and to only perform those previously deemed (by the programmers) to be morally permissible.

Various authors have argued against the idea of being able to program robots to make moral decisions. In the context of autonomous weapons, Christof Heyns, the UN Special Rappoteur on Extrajudicial, Summary or Arbitrary Executions has argued against the use of autonomous robots to make lethal decisions on the battlefield on the basis that robots lack 'human judgement, common sense, appreciation of the larger picture, understanding of the intentions behind people's actions, and understanding of values and anticipation of the direction in which events are unfolding' (2013, A/HRC/23/ 47). The point is that the unpredictable variety of social situations that could arise on the battlefield means that it is unlikely that a set of pre-programmed rules about appropriate responses is likely to be applicable.

In an interesting paper about the requirements for creating robots with, what they term 'moral competence', Malle and Scheutz (2014) argue that, amongst other things, robots would require a network of moral norms, in order to know what is and is not morally acceptable. They suggest that it would not be practical to program this network, and that instead of programming robots with moral norms, they could learn and develop a network of moral norms on the basis of feedback given to them in response to their actions. They suggest that it might be necessary to raise the robots in human environments, since this may be 'the only way to expose them to the wealth of human moral situations and communicative interactions' (Malle and Scheutz, 2014). Others have suggested that robots' understanding of right from wrong could be

improved by training them on moral stories (Riedl and Harrison, 2016), and requiring them to reverse engineer the human values that they represent.

It is admittedly difficult to rule out the possibility that in the future a robot could be trained or raised to be moral, but there are reasons to be sceptical about the likelihood of success. Reasons for scepticism include the robot's lack of a biological basis for morality. As already discussed, an individual robot does not even care about its own body, let alone that of a human – it would suffer no pain if one of its wheels were to be removed for example. It could only be programmed to respond as if it cared about the effects of its actions on a human, or about any censure and moral disapproval of its actions. Another reason for scepticism is the complete lack of any convincing examples of robots developing a good, generalisable, understanding of the differences between right and wrong. All there is currently are examples of programmed behaviour, such as the robots programmed by Winfield et al (2014) to take actions to prevent other robots from falling into a hole, that have been described as exhibiting something that can be described as ethical behaviour. But the use of the term 'ethical' or 'moral' in this context does not mean that the robots in question could be legitimately praised or blamed for their actions.

(c) If robots are not responsible agents, should this limit the roles they are given and the situations in which they are deployed?

The original statement that robots are not responsible agents does not spell out what this implies for the deployment of robots. It is argued here that there are good reasons to

limit the social roles and decision-making powers of robots given their present capabilities, and those that are likely in the near future. As referenced above, Heyns (2013) argued that robots should not be allowed to make lethal decisions in battle, partly because of their lack of ability to understand social situations, but also because humans should have a right to have life and death decisions about them made by fellow humans. A related argument could also be made about robot policemen, who could be tasked with life and death (or serious injury) decisions away from the battlefield.

This argument can, and I argue should, be extended further to other kinds of decision where robots might restrict the freedoms of humans. A robot placed in the role of a teacher would have to make decisions about situations such as when to punish or restrain children, or when to praise them. A robot carer of older people might have to make decisions about when to share personal information about them with other people, or when to prevent them from doing something dangerous or risky. A robot nanny would have to make similar decision about its young charges. The point is that all these decisions are likely to involve moral judgements and evaluations of social situations, and for reasons already discussed the robot is unlikely to be able make good choices. Care should be taken to maintain human control, involvement, and responsibility in decisions that will affect the lives of humans. It is crucial that we find ways to ensure that robots are not placed in situations, or given social roles, that will result in allowing them to make moral decisions that will affect people's lives. There are already risks of automated decisions affecting our lives, but robots that can be given the appearance of competent social actors make these risks even more prevalent.

Summary: It is easy to agree with the EPSRC principle about robots not being responsible agents, but this brief consideration finds it to be insufficient to guide future

action. It does not refer to any reasons for claiming that robots are not responsible agents, nor consider the implications for the deployment of robots and for human choices about the social roles they should be given. At present, and in the near future, even when efforts are made to program robots to follow the law and to respect individuals' rights and freedom, those robots are not going to be able to understand social situations and consequently will not be able to consistently make the right moral decisions about human social situations. It is therefore important to avoid placing robots in social roles and situations in which moral decisions are required. Care should be taken to avoid or minimise automatic and algorithmic decision making in situations in which human judgement is required. Even greater care is needed in the case of robots that create the illusion that they understand. Humans do sometimes make flawed decisions, but they can reflect and learn from them and develop a better moral understanding in a way that a robot cannot.

References

Arkin, R. (2009). Governing lethal behavior in autonomous robots. Chapman-Hall review. Computers and Education, 58(3), 978–988.

Asaro, P. M. (2006). What should we want from a robot ethic? International Review of Information Ethics, 6, 9–16.

Bekoff, M., and Pierce, J. (2009) Wild Justice: The Moral Lives of Animals. The University of Chicago Press, London.

Brosnan, S.F. and de Waal, F.B. (2003). Monkeys reject unequal pay. Nature, 425, 297-99

Churchland, P. (2011) Braintrust: What Neuroscience tells us about Morality. Princeton University Press, Oxford.

Heyns, C. (2013). Report of the Special Rapporteur on extrajudicial, summary or arbitrary executions, A/HRC/23/47

Johnson, D.G. (2006). Computer Systems: Moral Entities but not Moral Agents. Ethics and Information Technology, 8(4): 195–204

Johnson, D.G., and Miller, K.W. (2008) Un-making artificial moral agents. Ethics and Information Technology (2008) 10:123–133

Malle, B. F., & Scheutz, M. (2014). Moral competence in social robots. IEEE International Symposium on Ethics in Engineering, Science, and Technology (pp. 30–35). Presented at the IEEE International Symposium on Ethics in Engineering, Science, and Technology, June, Chicago, IL: IEEE.

Maturana, H. R. & Varela, F. J. (1980). Autopoiesis and Cognition - The Realization of the Living. Dordrecht, The Netherlands: D. Reidel Publishing

Riedl, M.O., and Harrison, B. (2016) Using stories to teach human values to artificial agents. In Proceedings of 2nd International Workshop on AI, Ethics and Society, Phoenix, Arizona

Sharkey, A. J. C., & Sharkey, N. E. (2012). Granny and the robots: Ethical issues in robot care for the elderly. Ethics and Information Technology, 14(1), 27–40.

Sharkey, N. E., & Sharkey, A. J. C. (2010). The crying shame of robot nannies: An ethical appraisal. Interaction Studies, 11(2), 161–190.

Sharkey, N. E. & Ziemke, T. (2001). Mechanistic vs. Phenomenal Embodiment - Can Robot Embodiment Lead to Strong AI Cognitive Systems Research, 2, 4, 251-262

Sparrow, R., & Sparrow, L. (2006). In the hands of machines? The future of aged care. Mind and Machine, 16, 141–161.

Sullins, J. P. (2006). When is a robot a moral agent? International Review of Information Ethics, 6(12), 23–30.

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

Winfield, A.F., Blum, C., and Liu, W. (2014) Towards an ethical robot: Internal models, consequences and ethical action selection. In M. Mistry, A. Leonardis, M. Witkowski, & C. Melhuish (Eds) Advances in autonomous robotics systems: Proceedings of the 15th annual conference, TAROS 2014 (pp 85–96). Birmingham, UK, 1–3 September