Available online at www.sciencedirect.com
ScienceDirect
Interacting with Computers 20 (2008) 338–341
Interacting with Computers
www.elsevier.com/locate/intcom

Robot ethics? Not yet
A reflection on Whitby's "Sometimes it's hard to be a robot"

Harold Thimbleby

Department of Computer Science, Future Interaction Lab, Swansea University, Singleton Park, Wales SA2 0SF, UK

Available online 13 February 2008

Abstract

Science fiction stories seductively portray robots as human. In present reality (early 21st century) robots are machines, even though they can do many things far better than humans (fly, swim, play chess to name a few). Any ethics for or of robots is therefore a seductive mix of fiction and reality. The key issue for rational discourse is to provide a rigorous framework for reasoning about the issues, including identifying flaws in the framework. We find such meta-reasoning in discussion about robot ethics to be ready for improvement.

This paper takes its inspiration from B. Whitby, "Sometimes it's hard to be a robot: A call for action on the ethics of abusing artificial agents," *Interacting with Computers*, this issue, 2008.

Keywords: Robot ethics; Meta-ethics

1. Introduction

Ethics is about doing what is right, but as there is little consensus in ethics, meta-ethics is the "higher-level" field concerned with choosing (and about rational debate on how to choose) consensual perspectives from which notions of right and wrong are self-evident. For example, if we select the perspective of *hedonism*, perhaps justified on the basis that life is short and apparently pointless, then personal enjoyment is paramount: it is right, for a hedonist, to enjoy themselves while they can. From a Christian perspective, human life is seen in the context of an eternal relationship with God, and this defines a particular *divine ethics*, where, for example, love of one's neighbour is crucial, and hence definitions of "neighbour" (e.g., Are slaves neighbours? Are robots neighbours?) stimulate lively debates within that perspective. Meta-ethics helps reason about the relations between and the merits of hedonism, divine ethics, and so forth.

Indeed, ethics come in many different "flavours," covering many different dimensions of interest. Thus *consequentialism* (what happens matters) and *virtue ethics* (my intentions matter; if I am virtuous, I do good) are at opposite ends of one dimension. *Utilitarianism* (a particular consequentialist ethics), *feminist ethics*, *environmentalism*, and many more, may be considered majority perspectives on ethics; while *sadism*, *Marxism*, *communitarianism*, and many more, might be considered relatively minority perspectives. Furthermore, there are many ethics that are of considerable historical interest, such as *Kantianism*, though one should not confuse the strength of following amongst academic specialists to reflect their claim to be satisfactory perspectives for contemporary society.

Important to our later discussion, below, there are circumscribed ethics, such as *professional ethics*, that make no claim to general application, but are focussed on particular domains. Professional ethics is concerned with what is considered right or good "professional" behaviour, usually as seen from the perspective of a particular professional body, such as the British Computer Society. Typically, professional ethics are *deontological*, that is duty-based ethics. In contrast to duty-based ethics, as a special (but important) case, *situational ethics* takes the view that there are no valid general positions, but ethical issues are to be debated and resolved with respect to particular situations – perhaps even as specialised as those that may concern the British Com-

E-mail address: H.Thimbleby@swansea.ac.uk

0953-5438/$ - see front matter © 2008 Published by Elsevier B.V.
doi:10.1016/j.intcom.2008.02.006

puter Society. Professional ethics often glosses the distinction between right and what is deemed right; for example, the BCS Code of Ethics glosses the difference between what action (or inaction) is acceptable to remain a member of the BCS and what is "actually" right action (or inaction). Of course it would not serve the purposes of the BCS as an effective professional organisation (e.g., where it behaves predictably in response to its members' behaviour) if its code of ethics was a meta-ethical document that opened up analyses of these issues – which have not been resolved for centuries, let alone with the novel features of computing technologies! Clearly these are very complex issues; we will revisit situational ethics again, in the Conclusions.

Of course, the very notion of a "majority perspective" on ethics, while a convenient turn of phrase in the introduction to a brief paper such as this, might be taken to suggest that meta-ethics is in some sense democratic! In fact, how one chooses a "right ethics" is itself an ethical issue and part of the rich fabric of moral philosophy to examine in depth. For example, while a divine ethics may claim to be right whatever anybody else thinks (it only matters what God thinks), sadism may hardly care what other people think – at least provided they do not like it!

Into this rich and diverse collection of ethical perspectives, we can follow Whitby and add (what we will term) *robotism*, and more specifically *Whitby robotism* as the particular form of robotic ethics presented in Whitby's recent paper "Sometimes it's hard to be a robot" (Whitby Whitby, in this issue).

In brief, the aim of Whitby appears to be to make a call to establish a professional ethics that accommodates robotism, and he considers doing so a matter of some urgency. His paper, then, can be considered to raise four pertinent issues: what is robotism; what is Whitby robotism; the urgency of accommodating robotism in professional ethics; and, finally, the validity or coherence of his arguments considered more generally.

Environmental ethics and animal ethics are both recent developments in ethics that emphasise there is a sense of right and wrong action with respect to inanimate objects as well as to living but non-human objects, and moreover that this sense resonates with all thoughtful and informed people. In other words, these ethical positions are stimulating and indeed valid in some way. (In contrast there are plenty of trivial ethics, which are of negligible interest to anybody else, such as *my* selfishness determining what I think is right for me – a view since Kant is that ethics that are not universalisable are inconsequential.) Whitby has implicitly raised the issue of robotism, or robot ethics, and he contends it is as plausible as environmental or animal ethics. All these ethics – environmental, animal, robot – are about right action by humans towards these things.

In my reading of Whitby, this is a good analogy. The environment deserves human respect. Robots deserve human respect. The environment itself has no ethics (e.g., environmental ethics is not divine ethics) and, for example, if somebody is abused by the environment, environmental ethics sees this, if at all, as a consequence of earlier human abuse. A good illustration here would be the unhealthy consequences of acid rain, but the normal ethical focus here would be the earlier human cause of the pollution – not the agency of the environment. Similarly, Whitby does not consider the ethics of robot action, which was classically discussed at length by Isaac Asimov (see Thimbleby et al, 1995), but of human action *to* robots. He takes it that mistreatment and abuse are self-evidently bad. He seems to see the relation of humans and robots as asymmetric: humans are the designers; humans are the abusers.

This may be no more than a word game. While "dismantling" a robot seems to have neutral ethical implications, call that dismantling "mistreatment" and it evokes a sentimental response. If we build robots to destroy each other then at first sight this seems negative, but if we run competitions to engineer more ingenious robots this view of exactly the same event, is very constructive. We may say we need more inspired and engaged engineers (we certainly do if robotism is to become a real issue!) and if such people are inspired by "robot wars" then good on them. Similarly, if we choose to say that robots "suffer" then "abuse" is an appropriate word for how we may make them suffer. If we choose to say that robots are no more significant in the world than stones, then "suffering" is as an inappropriate word as "abuse" is.

A more realistic view is that there is here an issue of degree. I may be happy killing millions of bacteria with an antibiotic; I may even celebrate the apparent destruction of entire species such as smallpox, bedbugs or wasps. Yet as we advance up some notional evolutionary scale, to culling seals, say, then things become emotive. Rationally debating the rights and wrongs of that issue thus involves the establishment of an animal ethics as a coherent point of view. Similarly, while I do not worry that I might crash my car or consider crashing in itself to be an ethical issue, if I could have conversations with my car about where I could drive today, I might come to regret the loss of its (albeit narrow) companionship following "abuse" of the navigation system. If so, then a robotism would provide a framework for discussion of such ethical issues in a reliable way; given the assumptions of robotism, one makes reliable inferences, rather than continually changing one's perspective to make (what are usually) sentimental points.

Whitby recognises that robotism is more interesting when robots resemble humans. Let us briefly take this the other way around. Thus: human ethics is very interesting when humans resemble robots. On what grounds, if any, are we justified in aborting a cognitively-defective human? When does unconsciousness or vegative state, if ever, move a human personality from "human" to "non-human"? When is, or may be, euthanasia a good? If we had clear notions of when and under what circumstances it is good to terminate a human life, we might have clearer ideas on when and under what circumstances it is good to terminate a robot. Tellingly, Whitby's discussion at this point is concerned with the more mundane ethics of the robot's human designer, not the robotism itself.

Whitby makes frequent comparison with cars; robots have, from an ethical viewpoint, many of the properties of cars. However, we do not have a specific ethics of cars (if we ignore the deontological perspective of national legislation), and therefore the analogy is weaker than he appears to wish. Cars by their sheer volume do raise interesting ethical trade-offs. The infrastructure that supports cars provides employment and economic growth; the energy consumed in car travel threatens a global shortage of energy (and further down the line, famine and war over shortage of natural resources); they cause injury; they exaggerate the ethical consequences of certain otherwise private acts (e.g., not wearing seat belts increases national health costs). Robots today do not have the volume and range of social consequences that raises any ethical issue unique to robotism that is worth addressing as such yet.

As interactive robots become as ubiquitous as cars (as no doubt they will) because of the nature of capitalism, robotism will surely have a broad future significance, if only because of the aggregate effects. A single robot being slow (to take an arbitrary property of their behaviour), for instance, is tedious, but if millions of robots are slow, then the aggregate impact on society of robot sloth might add up to human lifetimes. When the impact has this large multiplier, then "sloth" and other vices and virtues currently may well have application to robots. Surely we are not in a position today to identify with any useful certainty what the relevant properties would be? There is no urgency to define robotism.

Indeed, Whitby's analogy with cars ironically makes the point if it is thought through. Had early workers in cars defined professional or other ethical stances with respect to cars with any urgency *before* cars became a widespread technology, it is very likely ethics of cars would have focussed on now obsolete issues, such as the professional behaviour of chauffeurs or on scaring horses when cars travel faster than 4mph. Indeed one contemporary source (French, 1908) wondered why steam engines were not more popular as they were then very much more reliable and easier to use than the internal combustion engine. A little-known fact about early car technology was the important role of motorised ambulances in World War I, which significantly reduced mortality; the ethical issues, had they been addressed explicitly with "urgency" then, might well have referred to the humanitarian issues, or even to the loss of "essential" equine skills in the military. In other words, "urgency" before the fact is likely to make us miss the bigger picture.

Perhaps when Whitby says "urgent" he means "important"?

2. Debate and discourse

As my introduction makes clear, characteristic of ethics is the debate between different points of view. Whitby's paper is disappointing in that it does not provide a review of opposing or contradictory views. For example, speaking from a perspective of divine ethics one might claim that robots are made by humans and treating them as in some sense equals to living creatures (even humans) is tantamount to being well down the path towards idolatory. Or from the *communitarian* ethics perspective, we might wonder how – and how good or bad – it would be that robots contribute to community.

Whitby makes repeated use of the analogy with cars. Yet cars and the contexts in which they are found are not homogeneous. Human destructiveness aimed at a scrap car surely has a different status to destructiveness addressed to a vintage car? Or to a racing car *during* a competitive race, when in some sense it is designed to be abused? Similarly, there seems little ethical impact in hitting a stone with a hammer – but (i) if everybody did it, the landscape would be changed, hence the legal protection of places like the King Arthur's Stone (a megalith near Swansea University), and (ii) if the stone has human values fixed to it – for instance, the stone in question is a sculpture – then the stone has (at least) sentimental value, and most people would adopt a right/wrong view of its being hit, and they would also unproblematically distinguish between the sculptor hitting it and a vandal hitting it. As a piece of art, it may have been designed *to be* vandalised.[1] Similarly, we would argue that robots and the contexts in which they are found have a very wide range of variation that Whitby's approach does not admit.

William Perry claims that as we develop in *ethical sophistication*, the second step (out of *ten* steps) is to acknowledge disagreement (Perry, 1999; discussed in Thimbleby, 2007). The lack of opposing ideas expressed in the Whitby paper suggest that the arguments are not well-calibrated or balanced against alternatives. Indeed, Jürgen Habermas has raised this to an ethical position: *discourse ethics* is the view that ethics cannot be distinguished from and in fact is the engaged dialogue (i.e., community discourse) of reflecting on the facts of the world and right and wrong within it. In a sense this is merely legitimising discussion on ethics; that is what ethics *is*.

Another view is that of Hans Moravec (1998). Moravec views robots as a natural evolutionary advance over humans, and just as humans may consider themselves ethically "superior" to lower animals, robots will be "superior" to humans. Indeed, once robots start building robots to their own designs, their evolutionary path will accelerate away from limited human conceptions: Moravec anticipates super-intelligent, emotionally complex "cyberbeings." Accepting Moravec's futuristic point of view but also retaining a human-centred perspective, then, it is the humans who need robots to be ethically constrained. If so, the urgent issue now is how to embed and ensure for perpetuity human conceptions of ethics into robots – which is the other way around from Whitby's position.

[1] Community artists have the interesting problem of creating pieces that survive as art in sometimes hostile community settings – vandalism, fires, and so forth.

Whitby is right to start the dialogue on robotism, but he then fails the academic tradition of recruiting a sufficiently-diverse range of views to compare and contrast his new perspective – and this is particularly damaging to his case given the rich and very diverse nature of ethics.

Finally, Whitby provides no new grounds for reasoning, no clear principles or rules of inference, no conceptions of consistency of application, no meta-principles (such as appeals to rational agency, justice, reversibility, etc. – or even new ones or new variations of old ones, such as energy conservation, lubrication or power, issues that robots may care about). How are we then to think clearly about robotism and robotic ethics? (See, for example, Danielson, 1992, who proposes an "artificial morality.") It does not seem to this reader that Whitby's arguments or analogies empower the reader to build on his contribution; at least, not in the way he probably intends.

3. Conclusions

Whitby has not convinced this reader that robotism is distinct from any other technology-oriented ethics. He has not provided persuasive or provocative case studies, e.g., "it's not the gun that kills but the user" that arise frequently in other technoethical discussions. This reader does not share his sentimental interpretation of "abuse"; robots are not kittens. Robots are easy to build and mass produce; there are no grounds, at least as presented, to justify Whitby's sentimental title – "sometimes it's hard to be a robot..." (Aw, I feel sorry for them already.) I imagine, if robots had any notion of what "hard" meant that they'd say it was easy enough to be a robot. Even that view is contingent on a modest level of consciousness still inaccessible to them.

Robots clearly share with cars many of the emphatically ethical concerns of environment, care, and power. However, Whitby has made no clear case that robots raise new ethical issues. (We might consider cognitive prostheses, lifetime memories, injected nano-robots, exoskeletons used for social advantage, and so on as possible topics of concern.) He does repeatedly mention the potential robots have for abuse by humans, but without a grounding in (say) animal ethics, there is no perspective where abuse is self-evidently an ethical issue beyond the situational ethics any particular abuse raises.

By which I mean: I can chop down a tree, and I may do so aggressively with an axe. I may well be "abusing" the tree in the way I fell it. I might fell it slowly, taking pleasure say, in the way it weeps sap; I might anthropomorphise it into representing some human character I despise. Yet one can hardly see *in general* this raising any interesting ethical concerns. Yet if I choose to chop down my large beech tree, which happens to have a preservation order on it, and is indeed a majestic tree, then the situation raises all sorts of issues. Am I chopping it down because I am selfish and inconsiderate of the pleasure others have in it? Or am I a vandal chopping it down *because* doing so destroys the pleasure other have in it? Similarly, Whitby has no general case for robotism, but has made clear that there are situations where ethically interesting issues arise. However, that such issues may be expressed, as the situation demands, in the terminology of robotics does not in itself legitimise robotism as a stance of any broader interest.

Humans are complex and contradictory, and ethics attempts to provide various frameworks for a coherent intellectual discourse on what is right. Consider this: most people believe murder is bad, yet most of us enjoy watching movies that portray murder, sometimes on massive scales. One is criminal, one is entertainment. (The extreme category of snuff movies challenges us to be clear where we draw the line.) We may engage with consequentialist arguments concerning the impact violence or its portrayal has on viewers. Having thought thus about the portrayal of humans, we can now imagine robots – particularly robots as represented in movies – that have many human or even super-human skills. Such imaginary robots could participate fully in human society, and they would doubtless have rights.

In reality, and certainly for the time being, our science fiction-inspired imagination is far and away ahead of what is presently possible; real robots are mundane, and real robots are more useful than simulated humans. Calling a human "robotic" generally dehumanises and depersonalises them. In conclusion, then, it should be clear that we disagree with Whitby: there is no urgency to define robotism or any ethics of robots. But as nothing is certain, we should end on a careful note: there is no urgency until the robots themselves start to ask for it.[2]

Acknowledgements

The author gratefully acknowledges the inspiration of Whitby's fascinating paper. Hopefully the errors and other limitations of the present paper will serve as further inspiration, whether for a riposte or for new developments to the dialogue.

References

Danielson, P., 1992. Artifical Morality: Virtuous Robots for Virtual Games. Routledge.

French, J.W., 1908. Modern Power Generators. Gresham Publishing Company.

Moravec, H., 1998. Robot: Mere Machine to Transcendent Mind. Oxford University Press.

Perry, W.G., 1999. Forms of Ethical and Intellectual Development in the College Years. Jossey-Bass.

Thimbleby, H., 2007. Press on. MIT Press.

Thimbleby, H., Pullinger, D.J., Witten, I.H., 1995. Concepts of cooperation in artificial life. IEEE Transactions on Systems, Man & Cybernetics 25 (7), 1166–1171.

Whitby, B., in this issue. "Sometimes it's hard to be a robot: a call for action on the ethics of abusing artificial agents," Interacting with Computers.

[2] Robots asking for robot ethics? As Moravec (1998) says, even the wildest predictions of the future are not as unhinged as, in hindsight, they should have been.