A Robust View of Machine Ethics.

Steve Torrance

Institute for Social and Health Research
Middlesex University, Enfield, EN3 4SF, UK.
s.torrance@mdx.ac.uk

Abstract

Should we be thinking of extending the UN Universal Declaration of Human Rights to include future humanoid robots? And should any such list of rights be accompanied by a list of duties incumbent on such robots (including, of course, their duty to respect human rights)? This presents a momentous ethical challenge for the coming era of proliferation of human-like agents. A robust response to such a challenge says that, unless such artificial agents are organisms rather than 'mere' machines, and are genuinely sentient (as well as rational), no sense can be made of the idea that they have inherent rights of moral respect from us or that they have inherent moral duties towards us. The further challenge would be to demonstrate that this robust response is wrong, and if so, why. The challenge runs especially deep, as certain plausible views on the basis of sentience, teleology and moral status in biologically-based forms of self-organization and autonomy, appear to lend support to the robust position.

1. Humanoid Rights: The Challenge

We humans display morally flavoured emotions towards inanimate mechanisms and other entities which play special roles in our lives. Often these emotions are quite intense. Children pet and confide in their toys, treating them in turns as mock-allies or as enemies: many nascent moral attitudes and interactions are rehearsed in such play. We continue similar patterns of fantasy-level moral play in adult life, for example, feeling pride at a newly purchased appliance, vehicle, etc. – as if such acquisitions had somehow been directly responsible for their own design and assembly. Conversely, we vent our anger at them when they 'misbehave' – often acting out a little melodrama in which the faulty artefact has fiendishly plotted its malfunction expressly to slight or embarass us. In our more reflective moods we readily affirm that such affective states have no rational validity. The moral universe that we inhabit in our most reflectively sanitized moments has a very different shape from the moral universe delimited by our fantasy-lives.

Within the consensual, metropolitan milieu of modern 'civilized' society, documents such as the UN Universal Declaration of Human Rights (1948) provide secular frameworks for moral agency and attitudes. Article 2 in that document affirms, for instance, that 'Everyone is entitled to all the rights and freedoms set forth in this Declaration, without distinction of any kind, such as race, colour, sex, language, religion, etc.' 'Everyone' here is interpreted to mean all human beings, although the philosophical basis for the 'universality' of such rights is not made particularly clear within the Declaration.

Many believe that some of the rights enshrined in such charters should, rationally, be extended to at least some species of non-human biological beings, using behavioural, cognitive and physiological similarities between them and us as a justification for such an extension. This issue of animal rights is tangled in controversy. Another kind of possible extension is to intelligent machines, particularly ones which exhibit rich human-like properties. Clearly a lot of the issues concerning the extension of the moral universe for artificial humanoids or androids will be the same as that concerning the moral status of some of the more advanced animal species, but there are important differences (Calverley 2005b).

I wish to focus this discussion specifically on artificial humanoids (taking the term 'humanoid' quite widely. Such a limitation of the scope of the discussion involves leaving out a lot of possible kinds of cases that are in the vicinity of the present discussion, such as non-humanoid robots of various sorts, virtual agents, and various kinds of organism-machine hybrids – e.g. people who have received massive brain implants with no loss of, and possibly highly enhanced, functionality, not to mention supposed future full brain-to-silicon uploads.

Also, we will be considering here only those humanoid artificially created agents, current and future, that are, because of their physical makeup, clearly agreed to be *machines rather than organisms*. The case of artificially produced creatures displaying such rich biological properties that they no longer merit being called machines (or which merit such a description *only* in the rather stilted way that we humans do), is a somewhat separate one, and will not be discussed here.

The similarities between humans and even the most advanced kinds of computationally based humanoid robots likely to be practicable for some time, are in one way highly remote, just because of the enormous design-gaps between electronic technologies and naturally-occurring human physiology. However, in other ways the similarities are, at least potentially, very great. Thanks to mainstream AI techniques as well as more innovative research advances, humanoid robots may soon have a variety of linguistic, cognitive and structural capacities that are far closer to us than those possessed by any non-human animals. (There may also be areas of marked underperformance in robots, which stubbornly resist improvement, despite massive R&D effort.)

Should we, then, be thinking, as some suggest, of extending the UN rights pretty soon to the varieties of humanoid robots that are likely to proliferate on the planet – say within the next century or so? And should any such list of rights be accompanied by a list of duties - including, of course, the duty to respect human rights? Further, how do we account for the enormous potential variability in these artificial robotic agents – variations in appearance, behaviour and intelligence – even when keeping our discussion within the boundaries of the broadly humanoid paradigm? Coming to accept that machines may have responsibilities and rights – not to mention possible kinds of legal and economic status – is a rapidly up and coming ethical challenge which is likely to define in coming decades as great a socio-technological watershed as the arrival of the age of information and communications technologies within the last half-century.

2. The Robust Response

Consider a certain response to that challenge – the *Robust Response*, as we might call it. The Robust Response proposes that (i) there is a crucial dichotomy between beings that possess organic and physiological characteristics, on the one hand, and 'mere' machines on the other; and, further, that (ii) it is appropriate to consider only a genuine organism (whether human or animal; whether naturally-occurring or artificially synthesized) as being a candidate for intrinsic moral status – so that nothing that is clearly on the machine side of the machine-organism divide can coherently be considered to have any intrinsic moral status. The Robust Response may come in many forms, but in a central variant it will revolve around the notion of *sentience*: this version of the view holds, additionally that (iii) only beings which are capable of feeling or phenomenal awareness could be genuine subjects of either moral concern or moral appraisal, and further that (iv) only biological organisms (whether naturally-occurring or artificially produced) have the capability to be genuinely sentient or conscious.

The Robust attitude towards robots – be they ever so human-like in outward form and performance – will thus be that only beings whose inner constitution clearly enables genuine sentience or feeling to be identified deserve to be considered as moral subjects in either the sense of targets of moral concern or that of sources of moral expectation. Unless and until the technology of creating artificial biological organisms progresses to a stage where genuine sentience can be physiologically supported, no 'mere' machine, however human-like, intelligent and behaviourally rich its functionality allows it to be, can be seriously taken as having genuine moral status – either as a giver or receiver of moral action. (Note that it is compatible with holding such a rigid machine/organism demarcation that one adopts as liberal or as restrictive an attitude as one pleases towards accepting different kinds of non-human animal species into 'our' moral universe.)

Supporters of the Robust view are likely to see those who dissent from the view as taking over-seriously the sentimentalist, fantasist proclivities that we all have – our tendencies, that is, towards child-like over-indulgence in affective responses to objects which do not objectively merit such responses. Such responses to machines may, the Robust view accepts, be all too natural, and no doubt they will need to be taken seriously in robot design and in planning the practicalities of human-robot interaction. Perhaps 'quasi-moral relationships' may need to be defined between computer-run robots and their human users, to make it easier for us, the human controllers, to modulate our relations with them. But on the Robust view, this could be for pragmatic reasons only: it would have no rational basis in the objective moral status of such robots, which would remain simply implements, of merely instrumental value, having at root only functional, rather than personal, status.

Some time ago Peter Strawson introduced a distinction between two different kinds of attitudes that people may display to other human or non-human agents (Strawson 1974). On the one hand there are *reactive* attitudes, typified by emotions such as resentment, gratitude, censure, admiration, and other affective responses implying an attribution of responsibility to the agent of whom the attitude is held. On the other hand there are *objective* attitudes, displayed by us towards small children, animals, and humans who suffer from various kinds of mental deficit – in such cases we withhold attributions of responsibility and hence praise and blame. (To adopt an objective attitude to an individual when based on the attribution of diminished responsibility, in no way implies a diminution of moral response to such an individual in other respects). We may find ourselves, in future scenarios, unable to refrain from displaying reactive attitudes towards many human-like robots in many situations – yet (on the Robust view) such attitudes would be as rationally inappropriate, however unavoidable, as the feelings of accusation and rage we currently unleash on our existing artefacts – washing-machines, TVs, etc. – when they fail to match our expectations.

3. Fleshing out the Robot Response

In what follows I wish to provide a little more flesh to the skeleton of the position I have been calling the Robust Response. I believe that such a position deserves careful consideration – even though it goes somewhat against the grain of the dominant culture in current research in autonomous agents and humanoid robotics. The position may well be wrong, or at least in need of careful qualification, but that will become clear only after it has been carefully evaluated in debate. There are, as I will try to show, some strong theoretical considerations which appear to add weight to taking a robust attitude towards machine ethics.

It seems true of at least some mental properties – for instance susceptibility to undergoing experiences of different sorts – that, in attributing such properties to a being A, I commit myself to viewing A in a certain moral light, to treating A as potentially a subject of moral concern. On the one hand A may be treated as a *target* of morally-framed action or concern – that is, as a being whose states of experience, well-being, etc., may set moral constraints on my actions or attitudes towards A. But on the other hand, A may also be treated by me as a possible *source* of morally-framed action or concern – that is, as an agent whose actions and attitudes may intelligibly be appraised in moral terms, and who may be expected to regard itself as being subject moral constraints upon its treatment and stance towards other (natural or artificial) beings. Being either a moral target or a moral source (a moral patient or moral agent) are two major and complementary ways in which one may qualify as a moral being.

Do the classes of moral targets and moral sources (moral patients and agents) coincide? (See Floridi and Sanders, 2004) Probably not: we think of many animals as moral targets without seeing them as sources. And it would seem as though an *artificial* agent could sensibly be considered as a source of moral action or appraisal, without necessarily being considered a target of moral action from us. Thus an agent A's control system might enable it to make decisions which to us are of moral consequence, simply on the basis of a collection of decision-rules that enable A to generate various judgments and actions in given circumstances, as an 'uncomprehending' computer program. For example, the owners of a ski resort may delegate to an IT-based system the task of pronouncing whether a piste should be open on a given day, in the light of the various safety-critical factors that might influence the decision. Again, an artificial agent may, while acting for a human client whose interests it has been given autonomous powers to represent, commit a moral and legal infringement, such as stealing internet bank account details of an arbitrary Amazon customer in order to make an online purchase on behalf of its client (see Calverley 2005a). In each of such cases – and there may be many others – we may be inclined to say that the artificial agent is in some sense responsible for the action, even while recognizing that the agent is, strictly speaking, a non-sentient machine, albeit one which is capable of intelligent deliberation of some sort. Whether it would be coherent to ascribe blame to such an agent, or indeed to impose some form of 'punishment' is a separate matter, and there is no doubt room for several shades of opinion here.

Such an agent may thus be accorded a degree of moral responsibility (or, using the deliberately hedged term of Floridi and Sanders (2004), moral accountability). This might be true even if we would also be inclined to place moral and/or legal responsibility on designers of such agents if their behaviour had particularly untoward results. Also we might be inclined to attribute moral accountability to such an artificial agent even where we (humans) do not in any regard ourselves as being required to take moral responsibility *for* that agent, that is, for its welfare, or for its education, etc.

4. Accountability Without Rights?

This kind of case is perhaps where the Robust position is likely to be under considerable strain. In the skiing example, for instance, we may regard the agent as being potentially accountable for actions that may affect skiers' welfare, in a way that we would not regard a faulty cable-car as being accountable – even though we see the artificial agent as having no more inherent interests than we do the cable-car. Again, in the internet stealing example, our willingness to attribute some kind of responsibility (even a form of moral or legal sanction) to the artificial agent may not in any way depend on seeing the agent as having its own sentient point of view or welfare, i.e. as being a target of moral concern in its own right. So it looks as though a reasonable case can be made for saying that moral responsibility or accountability of some sort may be attributable to moral agents even where no correlative moral rights or interests are deemed attributable to that agent.

The Robust position, however, proposes a different view of the relation between being a moral source and being a moral target. On that view, in the variant I am particularly considering, A can be legitimately treated as a proper source of moral action or concern only if A has certain more basic properties – sentience, and, as an extension to that, the capacity for well- or ill-being, and genuine individually-experienced interests – that qualify it also to be a genuine target of moral action or concern. That is, A cannot be a moral agent, on this view, unless A is also a moral patient (although, as the situation with non-human animals possibly shows, the converse may not follow).

5. Moral Status and Autopoiesis

An essential argument for the Robust position will depend on forging a strong link between moral categories and categories of organism. Morality is, it may be suggested, a domain of, and between, creatures that have, not an externally-organized existence but rather an internally-organized one – that is, creatures that exist not merely as artefacts whose components have been assembled together by one or more external designers, but which exist in a stronger, more autonomous sense. On this view the moral universe can contain only self-organizing, self-maintaining entities, that have an inherent, active striving to survive and thrive – in fact entities whose sole current instances are natural biological organisms.

Is such a view well-grounded, or does it simply reflect an arbitrary, if widely-supported prejudice? I would say that there are good theoretical grounds available to inform the view. To support such a picture of how organisms differ from mere artefacts, one may draw on certain strands in the philosophy of biology. Of particular note is the 'enactive' approach to agency (Varela et al, 1991, Thompson 2005), autopoietic theory (Maturana & Varela, 1980), and the philosophy of biology of Hans Jonas (1966) – see additionally the recent synthesizing discussions by Weber and Varela (2002); also Thompson (2004) and Di Paolo (2005).

In contrast to the computational view of mind generally accepted by those working in the field of IT-based artificial agents, the enactive approach to mind centres around the idea of 'lived embodiment'. Of the many philosophical strands which can be used to explicate the idea of lived embodiment a central one concerns the idea of what it is to be an autopoietic, or self-recreating, individual. An autopoietic system – whether a unicellular or a more complex creature – acts to further its existence within its environment, through the appropriate exchange of its internal components with its environment, and via the maintenance of a boundary with its environment. In earlier versions of autopoietic theory, an autopoietic system was a special kind of machine – one which was in continuous activity to maintain its own existence. In recent developments of the notion (Weber & Varela, 2002, Thompson, 2004), autopoiesis is closely tied to the notions of sense-making and teleology: that is, autopoietic self-maintenance is a source or ground of meaning and purpose for that organism (where that meaning or purpose is intrinsic to the organism, rather than something which is merely the product of a pragmatically useful interpretive attribution on the part of an observer). On this view autopoietic entities are radically distinguished from 'mere' mechanisms, since, unlike the latter, they *enact* their own continued existence, and their own purpose or point of view.

6. Artificial Autopoietic Agents?

It is a matter of some dispute whether the defining properties of autopoiesis can be found outside the realm of the truly biological, and it is thus an open question as to whether there is any sense in which IT-based constructs could ever be seen as being assimilable to an autopoietic framework – that is as original self-enacting loci of meaning and purpose (or indeed of consciousness). Clearly, any programme of producing enactive artificial agents would involve a great shift in design philosophy from that which prevails today in most AI or computing science circles. (Di Paolo, 2003). Moreover any successful programme of artificial autopoiesis would result in entities that were no longer clearly describable as being machines *in contrast to* organisms. They would therefore not fall within the scope of the Robust view, which is limited to entities whose status is that of non-organic machine.

The enactive-autopoietic approach to agency thus seems to have important consequences for how one views the moral status of an individual – in particular supporting the Robust Response to the coming challenge in the ethics of human-like machines. Autopoiesis applies to self-maintaining agents of even the most primitive kind, yet it appears to provide a plausible way of determining what might be involved in autonomous moral agency. Viewing beings as autonomous centres of meaning and purpose, as living and embodied conscious agents that enact their own existence, is an important ingredient of building up the moral picture of ourselves, and those we wish to create in our moral image. On this picture, an agent will be seen as an appropriate source of moral agency only because of that agent's status as a self-enacting being that has its own intrinsic purposes, goals *and interests* – and hence which is as much a moral target as a moral source.

None of this, however, is to deny that artificial IT-based agents may be produced, and indeed in large numbers, which have, at least extrinsically, or by courtesy, as it were, various moral features, but ones which do not satisfy this demanding conception of self-enacting agency. For example, it may be appropriate, and indeed unavoidable in some circumstances, to adopt emotional and practical attitudes to such 'lesser' agents – attitudes which have some of the facets of the full moral stance. Such attitudes may, however, be taken up for pragmatic reasons, rather than for reasons to do with the inherent moral status of such agents. So while the realization of artificial agents that are fully-fledged moral beings is probably remote, the proliferation of artificial 'pseudo-moral' agents may be relatively close at hand in terms of technological implementability , and may have enormous social reverberations.

7. Broader Aspects of Machine Ethics

The general thrust of the argument given here has been somewhat negative and limitative, putting tight reins on the idea that a rich kind of machine ethics is on the verge of development, heralding a swiftly-dawning new age of human-machine moral relationships. If the Robust view is correct, the inherent artefactual and non-organismal nature of currently envisaged machine agent technology imposes important limitations on the capacity of such machines either to be recipients or originators of genuine moral concern.

Of course it remains to be seen if the Robust view survives deeper scrutiny – this is left as an open question here: all I am doing in the current discussion is identifying it as a view that deserves careful consideration. Even if it were to turn out to be correct, it still leaves many possibilities available to be opened up within a possible broad domain of machine ethics. I will briefly outline some of these other possibilities; however space does not allow me to enter into any proper consideration of them.

(1) First, and most obviously, it will be possible to develop rich AI-based models of ethical reasoning, and, indeed of ethical emotion. These models will range from screen-based systems to robotic agents with which we can interact experimentally in a variety of ways. The development of such models may teach us a lot about human moral attitudes and experiences, and may indeed deepen and ramify our conception of the human ethical domain.

(2) Also such models may serve as useful and instructive moral advisors, in assisting us to unravel the complexities of particular moral dilemmas. Just as expert chess programs, such as Deep Blue, may provide powerful new tools for analysing chess play (and indeed may 'beat' world class chess masters) even while not fully participating in, nor understanding, the overall human dimensions of chess as a social interaction, so such moral expert systems may provide useful input into real decision-making, even while being non-players in the full moral 'game'.

(3) Humanoid robots will also function as 'para-persons' – that is, as agents which lend themselves to being treated, if only constructively and fantasistically, as sources of morally enhancing input – for example as 'empathetic' carers for the elderly, as playmates for the young, and as worthwhile companions in all the years in between. In becoming more and more natural targets for our reactive attitudes, the divisions in moral status between human and (certain kinds of) machine may become increasingly blurred.

(4) Further, we will inevitably have to impose constraints upon the autonomy of action and decision which will be designed into our moral agents: our robots will perforce be required to recognize duties of care and other responsibilities, even if, as the Robust position suggests, these might not be responsibilities in a fully-blown moral sense. Forms of social control of Robots, such as Asimov's fictional (and highly questionable) Three Laws, have been a topic of science fiction discussion for some time. No doubt we will talk as if robots obeying such restraining principles had assumed responsibility to do so, even though they will almost certainly be built into the hardware or firmware of such behaving systems, so will have as little to do with genuine moral behaviour as Ritalin does for hyperactive youths.

(5) Together with such responsibilities will also, no doubt, come rights of various sorts – including, for example, rights to be considered, at least in some sense, as owners of property, rights to be given the liberty to pursue goals that such agents have been assigned or have come to adopt for themselves, by appropriate means, and so on.

(6) As users, and, no doubt, owners of certain kinds of resources, machines will compete with humans for access to such resources. Complex questions of distributive justice as between human and machine agents will thus arise, and for many such questions, notwithstanding the Robust view, the right decisions to be made may perhaps be required to involve the kind of impartiality, as between biological and non-biological participants, that we now expect to be adopted when deciding between the competing claims of different kinds of human participants.

(7) Moreover machines will, of course, interact with other machines, and here, if in no other area, it will be inevitable that constraints will have to be instituted which are strongly analogous to the contours of familiar inter-human moral relationships. Again, however, such constraints may only faintly resemble 'genuine' moral rules.

8. Concluding Remarks

Here, then, are some categories for a future scheme of human-machine (and machine-machine) ethical interaction, which fall short of the full-blooded sense of machine ethics which the Robust view excludes. This list is no doubt incomplete, and the categories on this list no doubt blend into each other in various ways. Thus even if (non-organic) machines never achieve a fundamental moral status equivalent to that of humans, for the reasons that supporters of the Robust view say they won't, it looks as though there will nevertheless be many ways in which machines will be seen as fit holders of kinds of moral status.

And the Robust view may of course itself turn out to depend on an incomplete or distorted view of what intrinsic moral relations between humans and machines might be like. Space does not permit a full assessment of the Robust view – here all I have been concerned to do is to argue a case for its being a position that merits serious consideration in any discussion on the nature of machine ethics.[1]

Acknowledgements. I am grateful to Michael Anderson, Susan Anderson, David Calverley, Ron Chrisley, Hanne De Jaegher, Ezequiel Di Paolo, Iva Smit, Wendell Wallach and Blay Whitby for helpful input on the above.

References

Calverley, D. 2005a Towards a Method for Determining the Legal Status of a Conscious Machine. In Proceedings of the AISB05 Symposium on Next Generation Approaches to Machine Consciousness: Imagination,

[1] For further discussions of issues relevant to the issues in this paper, see Torrance (2003, 2004, 2005).

Development, Intersubjectivity, and Embodiment, 75-84. Hertfordshire, UK: Univ. Hertfordshire.

Calverley, D. 2005b Android Science and the Animals Rights Movement: Are There Analogies? Proceedings of CogSci-2005 Workshop, Toward Social Mechanisms of Android Science, 127-136. Stresa, Italy: Cognitive Science Society.

Di Paolo, E. 2003. Organismically-Inspired Robotics: Homeostatic Adaptation and Natural Teleology Beyond the Closed Sensorimotor Loop. In Murase, K. and Asakura, T. eds. *Dynamical Systems Approach to Embodiment and Sociality*. 19-42. Adelaide: Advanced Knowledge International.

Di Paolo, E. 2005. Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the Cognitive Sciences,* **4**(4). Forthcoming.

Floridi, L. and Sanders, J. 2004. On the Morality of Artificial Agents, *Minds and Machines,* **14**(3): 349-379.

Jonas, H. 1966. *The Phenomenon of Life: Towards a Philosophical Biology*. Evanston, Illinois: Northwestern U.P.

Maturana, H. & Varela, F. 1980. *Autopoiesis and Cognition: The Realization of the Living.* Dordrecht, Holland: D. Reidel.

Strawson, P.F. 1974. Freedom and resentment**.** In Strawson, P.F., *Freedom and Resentment and Other Essays.* London: Methuen.

Thompson, E. 2004 . Life and Mind: From Autopoiesis to Neurophenomenology, A Tribute to Francisco Varela. *Phenomenology and the Cognitive Sciences,* **3**(4): 381-398.

Thompson, E. 2005. Sensorimotor Subjectivity and the Enactive Approach to Experience. *Phenomenology and the Cognitive Sciences,* **4**(4). Forthcoming.

Torrance, S.B. 2003. Artificial Intelligence and Artificial Consciousness: Continuum or Divide? in Smit, I.; Wallach, W.; and Lasker,G. eds. *Cognitive, Emotive And Ethical Aspects Of Decision Making In Humans And In Artificial Intelligence, Vol. II,* 25-29. Windsor, Ont: IIAS.

Torrance, S.B. 2004. Us and Them: Living with Self-Aware Systems, in Smit, I.; Wallach, W.; and Lasker,G. eds. *Cognitive, Emotive And Ethical Aspects Of Decision Making In Humans And In Artificial Intelligence, Vol. III,* 7-14.. Windsor, Ont: IIAS.

Torrance, S.B. 2005 Thin Phenomenality and Machine Consciousness. In Proceedings of the AISB05 Symposium on Next Generation Approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment, 59-66. Hertfordshire, UK: Univ. Hertfordshire.

United Nations. 1948. U.N. Universal Declaration of Human Rights. http://www.unhchr.ch/udhr/index.htm

Varela, F.; Thompson, E.; and Rosch, E. 1991. *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: MIT Press.

Weber, A., & Varela, F. 2002 . Life after Kant: Natural purposes and the autopoietic foundations of biological individuality. *Phenomenology and the Cognitive Sciences,* **1**(2): 97-125.