Robots are not just tools

Tony J. Prescott, University of Sheffield.

At the heart of the EPSRC principles of robotics (henceforth 'the principles') are a number of ontological claims about the nature of robots that serve as axioms to frame the subsequent development of ethical challenges and rules. These include claims about what robots are, and also about what they are not. The claims about what robots are include that "robots are multi-use tools" (principle 1), that "robots are products" (principle 3) and "pieces of technology" (commentary on principle 3), and that "robots are manufactured artefacts" (principle 4). The claims about what robots are not include that "humans, not robots, are responsible agents" (principle 2), that robots are "simply not people" (commentary on principle 3), and that robot intelligence can give only an "impression of real intelligence" (commentary on principle 4).

On first reading these statements seems straightforward assertions of obvious truths. I will argue that this is not the case. Indeed, I will propose that, these ontological commitments lack nuance, they assume all too easily that we know everything there is to know about robots, and they obscure or ignore some of the important ethical debates. If this is at all true, then progress towards a more useful set of principles could begin by thinking carefully about the real ontological status of robots.

If we look at how the principles are presented, there seems an implicit process of induction at work that allows statements about what most current robots are, to be re-interpreted as statements about what robots must essentially be. For example the statement that robots as multi-use tools in principle 1, slips into the claim that robots are "just tools" in the commentary on principle 2 and to the statement that "robots are simply tools of various kinds, albeit very special tools" in the preamble. Whilst it is easy to agree with a general statement that robots are multi-use tools, especially in the context of a discussion about dual use (principle 1), the much stronger claim that robots are just tools, or simply tools, denies that they could sensibly belong to other disjoint categories.

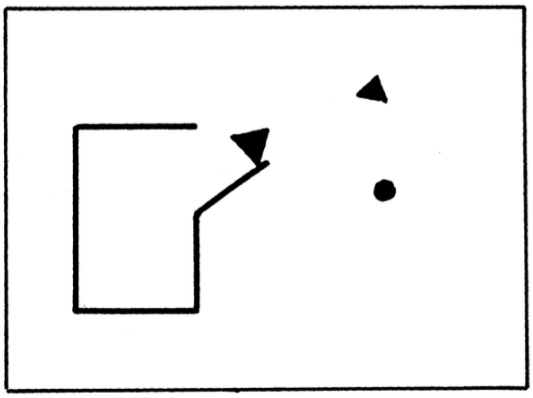
Take the category of 'companion' for instance. There is a major effort around developing robot companions that can provide social and emotional support to people as partially acknowledged in the discussion of principle 4. The category of tools describes physical/mechanical objects that serve a function, whereas the category of companions describes significant others, usually people or animals, with whom you might have a reciprocal relationship marked by emotional bond. The possibility that robots could belong to both these categories raises important and interesting issues that are obscured by insisting that robots are just tools.

Indeed, consistent with the view of robots as tools, the discussion of robot companionship in the principles is pretty dismissive, describing them as toys that could afford some pleasure to people who are unable to, or cannot afford to, keep animal pets. Robots are faux companions on this account that create an "illusion of emotions" and their intelligence is artificial and not "real". The faux nature of robot companions, it is argued, creates a real ethical problem in that robot companions are potentially, perhaps

even inherently, deceptive and so should be designed so that their "machine nature is transparent".

The ontological problem here particularly concerns the claim that robots could never possess psychological capacities such as "real" emotions or intelligence. What these are, in human terms, is poorly understood from the perspective of the cognitive and brain sciences and therefore there is no compelling reason to believe that these capacities must be unique to humans and could not be shared by machines.

A further problem concerns the assumption about how people will see robots—specifically, that robots will be seen as tools if they are shown in a transparent way. This could easily be wrong, for instance, people may anthropomorphise robots regardless of how obviously they are manufactured products. One reason to think this could be the case is the strongly social nature of our brains, and how easily our empathic systems are triggered by observing anything that appears even remotely life-like. The Heider-Simmel animations of simple geometric figures [1] (see figure), show just how crude this information can be and yet we will still see intentionality, motivation, even emotion. The invention of the Tamagotchi digital pet demonstrated that a simple 2-d animation of an animal-like creature can create a compelling urge to care [2]. We do not need to believe that the psychological state we read in to these artefacts is real in order to have a real emotional response to this ourselves.



Geometric shapes moving around in a simple animation were interpreted has "animated being, chiefly persons", in this famous 1944 study by Heider and Simmel.

An analysis of ontological and psychological issues in human-robot interaction has previously been made by Paul Kahn and colleagues [3]. Following a similar path, we can describe four general ways in which ontological questions about *what robots are*, and psychological questions about *how robots are seen*, could combine. There are illustrated in the following table along with some of the ethical issues they entail.

I. Robots are just tools, and people will see robots as just tools unless mislead by deceptive robot design. *Ethical issues: We should address human responsibilities as robot makers/users and the risk of deception in making robots that appear to be something they are not.* ***This is the position of 'the principles'.***	*II.* Robots are just tools, but people will see them as having significant psychological capacities irrespective of the transparency of their machine-nature. *Ethical issues: We should take into account how people see robots, for instance, that they may feel themselves as having meaningful and valuable relationships with robots, or they may see robots as having important internal states, such as the capacity to suffer, despite them not having such capacities.*
III. Robots can have some significant psychological capacities but people will still see them as just tools. *Ethical issues: We should analyse the risks of treating entities that may have significant psychological capacities (for example, the ability to suffer) as though they are just tools, and the dangers inherent in creating a new class of entities with human-like (or beyond-human) psychological capacities without recognising that we are doing so.*	*IV.* Robots can have some significant human-like psychological capacities and people will see them as having such capacities. *Ethical issues: We should consider scenarios in which people will need to co-exist alongside new kinds of psychologically significant entities in the form of future robots/AIs.*

Note that only one quadrant of this table (I) is addressed in the principles, but that II, III and IV are all possible, at least in principle. To conclude this essay I want to briefly consider some of the ethical issues that might arise in quadrants II–IV.

In quadrant II, interesting questions arise how robots should be treated—not because they are sentient agents but because people will choose to treat them as such. For instance the idea that it should be unlawful to wilfully damage robots, proposed as part of a "Robot Ethics Charter" by the government of South Korea [4], or that we might want to mourn the loss of a favourite robot, as has been reported for some owners of Aibo robots in Japan [5], does not seem so strange when viewed from the perspective of how robots are seen by people rather than in terms of what they are. Of course, appearance and function do matter, but transparency of "machine nature" will only be one factor of many influencing how people see and behave towards robots, it may be naïve to assume that it will be a decisive one. The bonds people will form with some robots may be similar to those we develop with other valued possessions, such as cars and mobile phones. On the other hand, for some robots, they may be more like the relationships we have pet animals, including for instance, wanting to take care of them, an activity that may fulfil a human need to nurture. Finally, some human-robot relationships may share similarities to human-human relationships. For instance, I may develop a bond with my companion robot not because it looks human but because it has the capacity to remember and communicate with me about some of our shared experiences. More generally, what may be needed, in order to develop suitable ethical principles, is to develop a taxonomy of different forms of emotional bonds that might exist between robots and people and analyse the factors that could underpin the development and maintenance of these relationships [6].

Quadrant III concerns the possibility of robots having significant psychological capacities that are ignored by people. This raises ethical risks that are not discussed in the principles, but that have been highlighted by others. For instance, Metzinger [7] has argued that we could build robots that are capable of experiencing suffering without realising that we are doing so, and therefore create a new kind of sentient entity that suffers unduly due to our actions. Although this may seem unlikely in the near-term, there are grounds to consider that this could be an ethical risk in the medium- to long-term as cognitive architectures for robots become more sophisticated. Several trends in ongoing research on human consciousness also support this possibility. First, one of the major contemporary theories of consciousness [8] asserts a critical role for integration of information that doesn't necessarily require a biological substrate. Neurologists are also re-appraising whether islands of integrated activity in the brains of patients who have suffered substantial brain damage, and who don't exhibit any outward behavioural signs of awareness, might constitute a form of minimal consciousness [9]. There is also an active debate as to whether animals with smaller brains than ours, such as fish, might sentient in a significant way (e.g. that they may experience pain), whilst having significantly less cortical tissue [10]. These developments suggest that consciousness could be possible in an artificial agent without having to match the size or complexity of an intact human brain. Dennett has argued that "crude, cheesy, second-rate, artificial consciousness" could be possible in robots [11], provided certain conditions are met, and Bryson [12] has proposed that today's robots might already have some simple forms of consciousness that meet some commonly proposed criteria. None of this is to claim that we are in quadrant III yet, but given the risks, ethicists should be pressing us as to how we would know if we were.

One of the consequences of the view of robots as 'just tools' is the implicit dismissal of the possibility of "strong AI"—that future robots could have human-level, or beyond human-level general intelligence. A quadrant III/IV issue, recently discussed by celebratory scientists and innovators such as Stephen Hawking, Elon Musk, and Bill Gates (to name a few), and analysed in-depth by Bostrom [13], is that an AI singularity could reverse the master-slave relationship between humans and robots. Believing that our robots/AIs are 'just tools' may keep us from recognising the early-warning signs of a self-bootstrapping super-AI. An ethical approach would surely encourage some vigilance. A more positive quadrant IV stance on the singularity debate is the perspective of the "global brain", proposed by Heylighen [14] and others, that humans and advanced AIs could co-exist to our mutual benefit. This reminds us that that ethics must be about analysing the possible benefits as well as the risks.

Although quadrant III/IV scenarios may seem far-fetched or at least distant, such concerns have captured the public imagination and have prompted significant calls for debate. The attempt to create a rhetorical barricade against these issues by insisting that robots are just tools may do little to calm the voices and could come across as hegemonic and condescending. Whilst approaches to these longer-term ethical challenges are necessarily speculative, as a starting point the principles could acknowledge that there are concerns here that are worthy of further attention.

A more candid approach may be to recognise that, whilst most robots are currently little more than tools, we are entering an era where there *will* be new kinds of entities that combine some of the properties of machines and tools with capacities that we had

previously thought were reserved for complex biological organisms such as humans. Following Kang [15], the ontological status of robots might be best described as *liminal*—neither living in quite in the same way as biological organisms, nor simply mechanical as with a traditional machine. The liminality of robots makes them both fascinating and frightening. In common with the *Association of Manhattan Scientists* who created a very different technology with potential for great harm or great good, and who wrote in 1945 [16] of their feeling of collective responsibility for their role in developing the atomic bomb—and of their "special awareness" of it potential for the "advance of our civilization or its utter destruction"—robotic researchers should recognise that in advancing the capability of robots towards a largely unknown end we have a special responsibility to understand and be outspoken about what the future might bring and its potential benefits and threats.

1. Heider, F. and M. Simmel, *An Experimental Study of Apparent Behavior.* The American Journal of Psychology, 1944. **57**(2): p. 243-259.
2. Levy, D., *Love and Sex with Robots*. 2007, London: Harper Collins.
3. Kahn, J., Peter H., et al., *What is a Human?: Toward psychological benchmarks in the field of human–robot interaction.* Interaction Studies, 2007. **8**(3): p. 363-390.
4. Lovgren, S., *Robot Code of Ethics to Prevent Android Abuse, Protect Humans*, in *National Geographic News*. 2007.
5. Brown, A., *To mourn a robotic dog is to be truly human* in *Guardian*. 2015: Manchester.
6. Collins, E.C., A. Millings, and P.T. J. *Attachment to Assistive Technology: A New Conceptualisation*. in *Assistive Technology: From Research to Practice: AAATE 2013*. 2013.
7. Metzinger, T., *The Ego Tunnel: The Science of the Mind and the Myth of the Self*. 2009, New York: Basic Books.
8. Tononi, G., *Consciousness as Integrated Information: a Provisional Manifesto.* The Biological Bulletin, 2008. **215**(3): p. 216-242.
9. Qiu, J., *Probing islands of consciousness in the damaged brain.* The Lancet Neurology. **6**(11): p. 946-947.
10. Seth, A.K., *Why fish pain cannot and should not be ruled out* Animal Sentience, 2016. **2016.020**.
11. Dennett, D., *The practical requirements for making a conscious robot.* Philosophical Transactions of the Royal Society of London A, 1994. **349**: p. 133-146.
12. Bryson, J.J. *Crude, Cheesy, Second-Rate Consciousness*. in *Vienna Conference on Consciousness*. 2008.
13. Bostrom, N., *Superintelligence: Paths, Dangers, Strategies*. 2014, Oxford: Oxford University Press.
14. Heylighen, F., *The Global Brain as a New Utopia*, in *Zukunftsfiguren*, R. Maresch and F. Rötzer, Editors. 2002, Suhrkamp: Frankfurt.
15. Kang, M., *Sublime Dreams of Living Machines: The Automaton in the European Imagination*. 2011, Cambridge, MA: Harvard University Press.
16. Association of Manhattan Scientists. *Preliminary Statement*. 1945; Available from: https://www.gilderlehrman.org/history-by-era/postwar-politics-and-origins-cold-war/resources/physicists-predict-nuclear-arms-race-.