Do Androids Dream of Electric Shocks?

Utilitarian Machine Ethics

By

Brendan Vize

A thesis submitted to the Victoria University of Wellington in fulfilment of the requirements for the degree of Master of Arts in Philosophy

VICTORIA UNIVERSITY OF WELLINGTON

2011

CONTENTS

ACKNOWLEDGEMENTS……2

CHAPTER 1: INTRODUCTION……3

CHAPTER 2: WHAT MAKES A BEING MORALLY CONSIDERABLE?……12

CHAPTER 3: COULD A MACHINE SATISFY CONDITIONS FOR MORAL CONSIDERABILITY?……27

CHAPTER 4: CAN A MACHINE BE A PERSON?……54

CHAPTER 5: CONCLUSION……78

BIBLIOGRAPHY……82

ACKNOWLEDGEMENTS

Thanks to Nick Agar, for all your suggestions, comments, critiques, and encouragement.

Thanks to the staff and fellow postgraduate students in the Philosophy department at Victoria University of Wellington, and to the undergraduate students that I have had the pleasure of tutoring; all your comments and (constructive) criticisms have been greatly appreciated. Thanks go out especially to the first-year student (whose name I have, unfortunately, long forgotten) who first gave me the idea for this thesis by remarking with callous indifference that the robots from *Star Wars* could never have rights. 18 months and 20,000 words later…here is my response…

Thanks to all my friends and to my family, for your support (moral, emotional, and financial…), for your interest in my project, and your encouragement. Thanks to my Lisa for your love, unflagging support and encouragement, for being a fantastic "housewife" when my schedule demanded it. I won't forget it when it's your turn to live the dream…

Finally, I wish to dedicate this thesis to my father, Des, and to my late mother, June, who have shown me, both by instruction and by example, how doing the right thing requires not just kindness, but also knowledge, and rational thought. I am very grateful for everything.

CHAPTER ONE: INTRODUCTION

"I have referred to this problem as the problem of the "civil rights of robots" because that is what it may become, and much faster than any of us now expect. Given the ever-accelerating rate of both technological and social change, it is entirely possible that robots will one day exist, and argue "we *are* alive; we *are* conscious!""

- Hilary Putnam (1964)[1]

Consider Lt. Commander Data from *Star Trek: The Next Generation*, the droid C3PO from Star Wars, or the Replicants that appear in *Bladerunner*: They can use language (or many languages), they are rational, they form relationships, they use language that suggests that they have a concept of self, and even language that suggests that they have "feelings" or emotional experience. In the films and TV shows that they appear, they are depicted as having frequent social interaction with human beings; but would we have any moral obligations to such a being if they really existed? What would we be permitted to do or not to do to them? On the one hand, a robot like Data has many of the attributes that we currently associate with a person. On the other hand, he has many of the attributes of the machines that we currently use as tools. He (and other science-fiction machines like him) closely resembles one of the things we value the most (a person), and at the same time, one of the things we value the least (an artefact), leading to an apparent ethical paradox. What is its solution?

[1] Putnam, Hilary. "Robots: Machines or artificially created life?" *The Journal of Philosophy* 61, no. 21 (1964): p.678

The possibility of a being like Data or C3PO eventually existing is surely not just science-fiction. At the outset, I will stipulate that there are no machines with which we regularly come into contact that deserve moral consideration, and probably no such machine in the world. However, as machines such as robots become more complex, we may start to encounter social robotic companions that behave in ways that we currently associate only with other humans. Certainly, there are many futurists who predict an increase in both the complexity of robots, and the complexity and frequency of our interactions with them. When Bill Gates started Microsoft in 1975, he envisioned a PC in every home[2]; now, in an article in *Scientific American* in 2007, Gates has predicted "A Robot in Every Home" [3]. The government of South Korea has announced a plan to put a robot in the homes of all its citizens by 2013[4]. With these predictions and policies come anticipations of the moral issues that would arise from a significant increase in the social complexity of robots. There already exists a society that is to robots as the SCPCA is to animals; the ASPCR (the American Society for the Prevention of Cruelty to Robots) claims to have been "Upholding Robotic Rights since 1999" (although they also note that they are "exactly as serious as robots are sentient", which may mean, at this point at least, not very)[5].

---

[2] Microsoft. "Microsoft's Tradition of Innovation: From Revolution to Evolution", October 25 2002, http://www.microsoft.com/About/CompanyInformation/ourbusinesses/profile.mspx, accessed 20/02/2011

[3] Gates, Bill. "A Robot in every Home." *Scientific American* 296, no. 1 (January 2007): pp. 58-65.

[4] Garreau, Joel. "Bots on the Ground." The Washington Post, May 6, 2007 http://www.washingtonpost.com/wp-dyn/content/article/2007/05/05/AR2007050501009.html , accessed 21/11/2009.

[5] ASPCR, "The American Society for the Prevention of Cruelty to Robots," (n.d.) http://www.aspcr.com/ , accessed 06/03/2010.

The possibility of having obligations to machines has also been considered by some philosophers. In 1964, Hilary Putnam predicted that technology will advance to the point that we will eventually have to deal with "the problem of the civil rights of robots"[6]. Michael Tooley briefly mentions the possibility of robots being given moral consideration in *Abortion and Infanticide* where he writes that the possibility of very advanced robots could bring the disagreement between materialist and dualists from its typical position in metaphysical discussion to the centre-stage of ethical discourse, causing "very serious moral disagreement about everyday matters"[7]. In 2009, Peter Singer and Agata Sagan wrote in the Guardian (and in an article in *Free Inquiry* magazine in 2010[8]), of the need to consider the possibility that some robots that may be developed that would be sentient, and hence (Singer and Sagan claimed) worthy of moral consideration[9]. Several others, both philosophers and those who specialise in relevant fields outside of philosophy, such as cognitive science and robotics, have also made valuable contributions to this topic, including Steve Torrance, Mark Coeckelbergh, and Joanna Bryson.

The goal of this thesis is to answer the question of whether a machine could be morally considerable, and the problem will be considered from a standpoint that assumes a utilitarian perspective. The work takes seriously Peter Singer's claims in "When robots have feelings", that machines could eventually be

---

[6] Putnam, p.678

[7] Tooley, Michael. *Abortion and Infanticide*. Oxford: Clarendon Press, 1983, p.89

[8] Singer, Peter, and Agata Sagan. "No Rights for Robots? Never?" Free Inquiry, June/July 2010: 13, 39

[9] Singer, Peter, and Agata Sagan. "When robots have feelings," December 14, 2009. http://www.guardian.co.uk/commentisfree/2009/dec/14/rage-against-machines-robots , accessed 06/03/2010

morally considerable, and asks whether this view is consistent with utilitarian views and particularly with those views advocated by Singer. In Singer's worldview there are three types of being (as described here by Philip Cole):

> Non-persons that are non-sentient; non-persons that are sentient; and persons. It is not possible to morally harm the first category at all; it is possible to harm the second category, but not by killing them; and it is only the third category that can be morally harmed by being killed.[10]

The main aim of this thesis is to ascertain which category (if any) an advanced robot similar to those seen in science-fiction would fall under, and what the ethical implications of this would be.

The importance of addressing these ethical issues is made clearer when we consider the likely magnitude of the harm that could be caused if we fail to address them. If Gates and others are correct, and machines do become part of our daily lives in the future, there will be significant amount of human-machine interaction. If we have not been able to clarify our ethical position on the matter in advance, there is the potential for widespread harm. Steve Torrance considers our current position with regards to robots to be similar to the position of our society 100 years ago, when cars were first becoming a part of citizens' lives, and suggests that more action in developing guidelines in these early days of the technology would have saved lives[11]. If robots are possible subjects of harm, but

---

[10] Cole, Phillip. "Problems with "Persons"" *Res Publica*, Vol.III, no.2 (1997) p.179.

[11] Torrance, Steve. "Ethics and consciousness in artificial agents." *AI & SOCIETY* 22, no. 4 (3, 2007): p.498

are not treated as such, then (if predictions of the proliferation into society of a large amount of robots are true) the result would be nothing short of a moral disaster. If however, they are not the type of thing that can be harmed, then resources that could be used for their protection would be better spent on humans rather than machines, whenever there is a conflict. It is important that these ethical issues are considered as soon as possible, to minimise moral mistakes in the future.[12]

A few notes on terminology and definitions: In the discussion below, the word "machine" is used interchangeably with "robot", and "android", without, I think, affecting the force of the argument. A dictionary definition of "machine" is: "an apparatus using mechanical power and having several parts, each with a definite function and together performing a particular task"[13]. However, this definition is much too broad, as it can be interpreted to include humans and animals (John Searle uses the word "machine" in this way[14]), which are not the intended subjects of the thesis. In this thesis, the word "machines" will be used in its colloquial sense, and a brief definition of my own devising, that captures the sort of entity I have in mind for the subject of this thesis is: "an artificial entity with moving or electronic parts, manufactured out of inorganic materials such as metal and plastic". Not too much emphasis should be placed on the exact wording

---

[12] Torrance, p.498 (f.n.)

[13] Oxford Reference Online. "Machine *Noun*" in *Oxford Dictionary of English*, edited by Angus Stevenson. Oxford University Press, 2010. http://www.oxfordreference.com/views/ENTRY.html?subview=Main&entry=t140.e0489390, accessed 7/02/2011.

[14] For example: "There is no question that machines can think, because human and animal brains are precisely such machines.", from Searle, John R. "Twenty One Years In The Chinese Room." In *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, edited by John Preston and Mark Bishop: 51-69. Oxford ; New York : Clarendon Press, 2002: p.56

of this definition, but this reflects the familiar use of the word, where machines are artefacts, categorical opposites of living things like animals and plants. In this thesis, the word "rights" will be avoided, since the idea of rights is sometimes thought to be problematic, and may be even more so for non-natural beings. So rather than use the term "rights", this thesis will be concerned with whether or not machines can be "morally considerable". This terminology is consistent with Peter Singer's own avoidance of the term "rights"[15].

The discussion that follows will be primarily about whether or not things can be *directly* morally considerable. A thing might plausibly be said to be morally considerable because we have obligations *regarding* that thing, even if it makes no sense to say that we have obligations *to* it. For example, even if I believe that animals cannot be morally harmed and I don't believe my neighbour's dog is morally valuable in itself, I might refrain from shooting it out of respect for my neighbour's interests. Similarly, as robots today are generally the property of an individual or a company, in one sense we already have obligations regarding some robots (such as to refrain from destroying or stealing them), just as we do with any property that is not our own. In this thesis, *if a being is said to be morally considerable, then it is so for its own sake, not because of its relationship to any other being.* While it may be that a future android should not be "mistreated" because it is someone else's property, this is not a harm *to* the android, but an indirect harm to another being. This distinction will prove important later on, but the first two chapters are primarily concerned with the direct moral considerability of machines.

---

[15] For example, see Singer, Peter. *Animal Liberation*. 2nd ed. New York, N.Y: New York Review of Books, 1990: p.8

Before moving on to the next chapter, I will frame the central thesis question in the form of an ethical dilemma. At the conclusion of this thesis, we should be able to provide a satisfactory answer to the question posed in this thought experiment:

Situation 1: The Spaceship Captain's Dilemma

Suppose that you are the Captain of a spaceship and that you must send one member of your crew on a particularly dangerous mission, which must be completed in order to save your ship from destruction. The mission involves entering the ship of a particularly violent alien species, and you believe that it is a suicide mission, likely to involve the death or destruction of the crew member that is selected. On the ship, there are only two crew members capable of performing the mission, and the nature of the task means that each is equally capable of success. You must choose to either send:

1) ***Adam****: A human crew member. OR*

2) ***Andrew****: A humanoid robot almost indistinguishable from a human person.*

Which of the crew should you send to almost certain destruction? As Adam is a person, he is morally considerable, so to make a decision in this situation, we need to know whether Andrew is also morally considerable. To get an understanding of whether Andrew is morally considerable requires answering a

series of questions, each of which will be the subject of a chapter or section of the thesis. The second chapter asks the question: “What makes a being morally considerable?” This chapter explores the concept of moral considerability according to utilitarian thinkers like Singer, and provides an indication of what to look for in a morally considerable being: sentience (or phenomenally conscious states that are qualitatively good or bad). The third chapter addresses the question of whether a machine could satisfy conditions for moral considerability. Finding an answer to this question will prove to be a challenge. The Problem of Other Minds will be considered as it relates to machines, and reasons will be given as to why the Problem of Machine Minds has greater implications for our metaphysical judgements about machines than the traditional problem of Other Minds has for humans and animals. The fourth chapter asks the question: “Can a machine be a person?” Given the behavioural similarity of machines like Data or C3PO to persons, it is natural that we should wonder whether these robots are moral persons. In chapter 4, I distinguish between “hard” persons (the traditional understanding of what a person is), “soft” persons (that have the behavioural aspects of personhood, but which are not conscious) and “fuzzy” persons (machines like Andrew). I argue that the preferences of “fuzzy” persons have less moral importance than the equivalent preferences of “hard” persons due to greater doubt over their sentience. I will establish several principles of machine ethics that will be defended within this work. At the end of the thesis, I propose that the decision of whether to perform an action that results in the satisfaction or frustration of a machine’s apparent preferences can be represented by an utilitarian equation. This equation can be used to resolve problems like the

dilemma introduced above. To begin, we should ask what it means to say that something is “morally considerable”. That is the subject of the next chapter.

CHAPTER 2: WHAT MAKES A BEING MORALLY CONSIDERABLE?

"Consciousness matters. Arguably, it matters more than anything."

– Nicholas Humphrey[16]

To explore the question posed in the Spaceship Captain's Dilemma, we need to know whether Andrew is morally considerable, so we need to know what it means to say that a being is "morally considerable". Once we understand what types of things can be morally considerable, we can then ask whether or not a machine could be one of those things. As noted, this thesis assumes a utilitarian standpoint, so in the discussion that follows, I will stipulate that the ethical principle that sentience is both necessary and sufficient for moral considerability is true. Later in the chapter, an in-depth analysis of the notion of sentience will help to provide a practical indication of what to look for in a morally considerable machine.

According to Singer, an act is wrong or right according to the degree to which it satisfies a being's preferences, but it is clear from Singer's work that it is not just any preferences that count; only the preferences of sentient beings count as morally important; Singer holds the view that sentience is both necessary and sufficient for moral considerability. He argues that sentience is *necessary* for moral consideration because sentience is required for being to have morally

[16] Humphrey, Nicholas. *Seeing Red: A Study in Consciousness*, Cambridge, Mass.: Harvard University Press, 2006: p.2

significant interests, and without sentience as a boundary, there is no non-arbitrary way to distinguish between the interests of things that we think of as considerable and those that are not considerable[17]. For a being's interests to be morally important requires that the being is capable of phenomenal experience. It is quite possible to speak of entities such as blades of grass having interests, but according to Singer these interests are not morally important interests. They are not morally important interests because they lack an essential phenomenal element that is part of the suffering of sentient creatures. Singer points out that, unlike in the case of animals, where we can use our imagination to approximate what it might be like to be an animal when they suffer, in the case of non-sentient beings like trees, there is nothing that it is like to *be* them when their interests are frustrated[18]. Moral empathy requires that we can put ourselves in the position of the other, and to imagine how they might be feeling; a clear impossibility in the case of non-sentient beings.

We ought to also consider sentience necessary for moral considerability because, without this boundary, we are led to an implausible conclusion: that we ought to consider as morally important the interests of things that common-sense tells us are not morally considerable. If I must take into account the interests of trees as morally significant, then why not cars or laptops, which can also plausibly be seen as having interests? It is true that trees may be said to be "auto-poietic" because their goals are inherent or self-created, in a way that perhaps the interests of cars are not[19], but this doesn't tell us why we shouldn't consider the

[17] Singer, Peter. *Practical Ethics*. 2nd ed. Cambridge: Cambridge University Press, 1993: pp.57-58

[18] Ibid, p. 277.

[19] Torrance, pp. 512-515.

interests of computer programmes that develop their own goals or of the grand canyon, which could be said to have an interest in not being filled in (remember that in this account, having interests does not imply *conscious* interests). These results stray too far from our ordinary moral intuitions and should be seen as a *reductio ad absurdum* of the claim that sentience is not necessary for moral considerability. Without the condition that the target of empathy be capable of conscious experience, the number of things that are morally considerable becomes too numerous, and renders the distinction between things that are rightly considered to be of moral concern, and those that are not, meaningless.[20] So, for a utilitarian, a machine must be sentient to be morally considerable. This is the first principle of utilitarian machine ethics:

> ***The Principle of Necessary Sentience:*** *If a machine is not sentient, then it cannot be harmed.*

This principle aligns with how we currently categorise machines.

If we could know that Andrew is not sentient, then he would know that Andrew was not capable of being harmed and so was not morally considerable. If Andrew is sentient however, and we could know that he was, then we would know that Andrew was morally considerable. This is because if a being is sentient, then this is not just a necessary condition, but also a *sufficient* condition for moral considerability. According to what Singer calls "the principle of equal consideration of interests"[21], any sentient machine's interest in avoiding

[20] Thompson, Janna. "A refutation of environmental ethics." Environmental Ethics 12 (2) (1990):147-160

[21] Singer, *Practical Ethics*, p.21

(equivalent) suffering must not be regarded as any less important than that of any other such interest, including our own. The principle of equal consideration of interests is the principle that: "an interest is an interest, whoever's interest it may be"[22], and it requires that equivalent interests ("in so far as a rough comparison can be made"[23]) be considered equally regardless of morally irrelevant considerations like race, gender, or species. Singer most notably used this principle to significant effect in *Animal Liberation* (1975), when he argued that most human beings were "speciesist", a term comparable (both descriptively and morally) to "racist" or "sexist"[24]. In Singer's view, pain is intrinsically bad, and the reason why we should wish to prevent a specific being's pain is not because it is *that* particular being's pain, but simply because of "the undesirability of pain as such"[25]. Singer claims that the reasoning of the racist or speciesist is flawed because it singles out for importance a category that is not a morally important category (race and species respectively). Once we recognise that pain is bad in and of itself, then any being in which it is found is an appropriate target for moral sympathy, whether or not it falls into the same racial or species category as ourselves. In *Practical Ethics (2nd Edition)*, Singer writes:

> To give less consideration to a specified amount of pain because that pain was experienced by a member of a particular race would be to make an arbitrary distinction. Why pick on race? Why not on whether the person was born in a leap year? Or whether there is more than one vowel in her surname? All of

[22]Singer, *Practical Ethics*, p.21

[23]Ibid., p.50

[24]Singer, *Animal Liberation*, p.9

[25]Singer, *Practical Ethics*, p.21

> these characteristics are equally irrelevant to the undesirability of pain from the universal point of view.[26]

If it cannot matter to "the undesirability of pain from the universal point of view" that a being is of a particular race or gender, then nor should it matter what material a being is made from, since the principle of equal consideration of interests requires that any interest in avoiding suffering, regardless of where that interest lies, be considered equally with any equivalent interest. And this result is of obvious significance for the moral considerability of machines. If a being can really suffer, then the fact that it is made of metal rather than flesh, or the fact that it was designed and manufactured rather than naturally conceived and born, should be of no consequence to an assessment of its moral worth. Those who think otherwise hold an unjustified prejudice, equivalent to racism or speciesism. The prejudice thus referred to is sometimes known as "substrate chauvinism"[27], (defined as "the conviction that only biological matter can carry moral worth"[28]) or "substratism"[29]. The idea of a prejudice against robots was considered in 1964 by Putnam, who wrote: ""discrimination" based on the "softness" or "hardness" of the body parts of a synthetic "organism" seems as silly as discriminatory

---

[26] Ibid., p.22.

[27] See Virtual Worldlets Network, "VWN Virtual Dictionary: Substrate Chauvinism," (n.d.), http://www.virtualworldlets.net/Resources/Dictionary.php?Term=Substrate%20Chauvinism&Letter=S, accessed 21/09/2010; and Dvorsky, George. "Sentient Developments: Must-know terms for the 21st Century intellectual: Redux", January 11, 2007. http://www.sentientdevelopments.com/2007/01/must-know-terms-for-21st-century_11.html, accessed 03/06/2010.

[28] Virtual Worldlets Network

[29] Walker, Mark. "A Moral Paradox in the Creation of Artificial Intelligence: Mary Poppins 3000s of the World Unite!" in *Human Implications of Human-Robot Interaction: Papers from the AAAI Workshop*, edited by Ted Metzler, California: AAAI Press, 2006: p.3

treatment of humans on the basis of skin color."[30] This prejudice is not a problem now, since the synthetic beings we encounter today do not have any more sentience than a rock, but it is possible that in the future, substratism may indeed become harmful discrimination, if we build (or encounter) sentient beings made from inorganic material.

It might be claimed that a prejudice against equal consideration for robots is not an irrational prejudice. An objection could be made that since only beings made of flesh can experience *real* pain, a prejudice against machines would be entirely justified. It is true that such a prejudice would not be irrational now, since we have good reasons to believe that no currently existing artificially-created beings can feel pain; but this may not remain true forever (the extent to which we might suppose the connection between pain and flesh necessary will be considered in the next chapter). Moreover, a distinction must be made between making a judgement about the likely properties of a being, and a *moral* prejudice about the deserved treatment of a being. We don't think that rocks are sentient, so there is nothing irrational about a moral prejudice against the kind treatment of rocks. The principle of equal consideration of interests requires us to accept that *if* a being made of metal or plastic was found to be experiencing suffering equivalent to suffering that we ourselves can feel, *then* it would be an irrational prejudice to ignore this suffering purely because the being was inorganic. Now we have arrived at the second principle of utilitarian machine ethics:

> ***The Principle of Equal Consideration of Interests:*** *If a machine is sentient, then its preferences or interests should be considered as*

[30] Putnam, p.691

important as the equivalent preferences or interests ("in so far as rough comparisons can be made") of other sentient beings.

For a utilitarian, since sentience is both necessary and sufficient for moral considerability, the question "Can a machine be morally considerable?" can be replaced with the question: "Can a machine be sentient?" The next chapter will consider the problem of how we could tell whether a machine is sentient, but first, we must know what it is that we are looking for. What is "sentience"? Singer uses the term "sentience" as: "shorthand for the capacity to suffer or experience enjoyment or happiness"[31]. But in this thesis, it will be beneficial to unpack the term a little more; although sentience is relatively easy to spot in the natural world, it will prove harder to locate in the artificial world, as the discussion in the next chapter demonstrates. An analysis of the term will help us to understand what we mean when we say that sentience is both necessary and sufficient for moral considerability.

Sentience suggests two components, each of which are necessary to really be considered the type of sentience that warrants moral considerability. A sentient being is both phenomenally conscious *and* has qualitatively good or bad experiences (usually understood as pleasure and pain). A being that was able to have conscious experiences, but that had no preferences about its conscious states, could not be morally significant. Nor could a being that had such preferences, but no conscious experience. It will be necessary to say more about

[31] Singer, *Practical Ethics*, p.58.

each of these components and to say why they are each important for the moral significance of sentience.

CONSCIOUSNESS

It should be clear that conscious experience is a vital part of sentience. In *The Emperor's New Mind*, Roger Penrose asks us to consider a machine that receives ratings from its programmer for things that it does (For example, running out of electricity might be rated "-1", but spending time in the company of other machines might be rated "+1"). It avoids the things with negative ratings, and seeks out the things with positive ratings, trying to increase what Penrose calls its "pleasure-pain score"[32]. Despite the fact that the behaviour of the machine would resemble that of some animals, this machine would not be sentient. Mere "avoidance behaviour" is not the same as being *in pain.* For that, a conscious experience of the sensation of pain is required.

Despite the apparent difficulties in securing an agreed upon definition during years of philosophical discussions about consciousness, it is clear that some definitions are far superior to others. In particular any definition that does not take into account the experiential elements of consciousness does not seem to be talking about consciousness at all, since what seems to be so unique about consciousness (one might even say that it is literally its *defining* feature) is the experience of qualia. One of the best descriptions of consciousness comes from David Chalmers, who writes:

[32] Penrose, Roger. *The Emperor's New Mind : Concerning Computers, Minds, and the Laws of Physics*. Oxford ; New York: Oxford University Press, 1989: pp. 17-21.

> When we perceive, think, and act, there is a whir of causation and information processing, but this processing does not usually go on in the dark. There is also an internal aspect; there is something that it feels like to be a cognitive agent. This internal aspect is conscious experience. Conscious experience ranges from vivid color sensations to experiences of the faintest background aromas; from hard-edged pains to the elusive experience of thoughts on the tip of one's tongue; from mundane sounds and smells to the encompassing grandeur of musical experience; from the triviality of a nagging itch to the weight of a deep existential angst; from the specificity of the taste of peppermint to the generality of one's experience of selfhood. All of these have a distinct experienced quality.[33]

As Chalmers admits, central to this description is Thomas Nagel's definition of a conscious being. In *"What is it Like to be a Bat?"* Nagel writes that: "An organism has conscious mental states if and only if there is something that it is like to *be* that organism – something that it is like *for* the organism."[34] [italics in original]. This definition captures the primacy of the *experience* of consciousness. In the following discussion, when the term "consciousness" is used, it is Nagel's definition of consciousness, and Chalmers' description of qualia, that the reader should keep in mind.

It is of special significance to any moral discussion about machines that we establish not only what is meant by "consciousness", but also what is *not* meant.

---

[33] Chalmers, David J. *The Conscious Mind: in Search of a Fundamental Theory*, New York : Oxford University Press, 1996: p. 4.

[34] Nagel, T. "What is it like to be a bat?" *The Philosophical Review* 83, no. 4 (1974): p. 436.

In particular, it is vital to separate the two concepts of "thinking" and "consciousness", or what Chalmers calls the "psychological" aspects of mind and the "phenomenal" aspects of mind[35]. As Jack Copeland has noted, when talking about machine intelligence, the separation of these two concepts can have "a liberating effect on the discussion"[36]. This is especially true since computers are able to perform many tasks that resemble human thought, but that presumably don't have the phenomenal content that sometimes accompany their equivalents in human beings. The human brain also accomplishes many things unconsciously that can properly be described as "thought". To use a well-worn example, we may drive down the street while preoccupied, all the while avoiding obstacles and changing gears[37]. Here, our conscious mind is focussed on something else, and we find that our brain has nevertheless allowed us to navigate through traffic safely. Thought processes without any conscious content can indeed be quite remarkable: in the case of "blindsight", an affected patient may accurately locate a spot of light on a board by pointing to it, even while they have no conscious experience of being able to see the light, and will claim that they are merely guessing at its location[38]. The mind is able to perceive the light, locate its position, and direct the hand exactly how to point to it, using only unconscious mental processes, while the conscious mind has no awareness of the thinking that the brain is engaged in. If such processes can be properly described as "thought" (and if it is not "thought", then what is it?), then from examples such as these, we can

---

[35] Chalmers, p.12

[36] Copeland, Jack. *Artificial Intelligence: A Philosophical Introduction*. Oxford, UK: Blackwell, 1993: p.34.

[37] For example, see Carruthers, Peter. "Brute Experience." *The Journal of Philosophy* 86, no. 5 (1989): p.258.

[38] Copeland, p.35.

see that "thinking" and "consciousness", though often related, are indeed distinct concepts[39]. If thinking and consciousness are conceptually separate, then we can describe what a machine is doing as "thinking", without thereby implying anything about its possession of phenomenally conscious states. As a consequence, we can then say that machines can think, or that they have mental or psychological states, without making any commitment about their moral status.

This distinction can also help to clarify some issues related to the ontological status of consciousness. Putnam claimed that the issue of whether or not a robot was conscious was purely a matter of choice; he argued that the solution to the problem was more a decision about which words to use, than a discovery to be made[40]. But if we take a Nagelian view of consciousness, then this doesn't seem to be right. If there is something that it is like to be a being, then no amount of decision-making about words can change that fact. As John Searle puts it, "my present state of consciousness is entirely observer-independent, No matter what anyone thinks, I am now conscious."[41]. On the other hand, whether or not a being is "thinking" or "understanding" is a decision that needs to be made about which words are appropriate to use to describe the behaviour (if "behaviour" is the right word to use) of a machine, when it does things that

[39] The diversity of unconscious thought processes is well-documented. For several more examples, see Chapter 6 of Macphail, E. M. *The Evolution of Consciousness*. Oxford: Oxford University Press, 1998: pp.138-175.

[40] Putnam, p.690

[41] Searle, "Twenty One Years In The Chinese Room,", p.62.

appear similar to what a human does, when (s)he thinks or understands[42]. The difference between the case of thinking and consciousness is that to claim that a computer is thinking postulates nothing occurring aside from what we can observe[43]. We know that when we push buttons on a calculator to add two digits, the states of flip-flops inside the calculator's processing chip lead eventually to the calculator's display screen illuminating in certain parts, and as observers, we interpret those illuminated parts as numbers, and the whole process as "calculating". To use Searle's terminology, the calculator's "calculating" is "observer-dependent" or "observer-relational"[44]. Without an observer to interpret the calculator's states as the adding of two numbers, the sequence of states that occur inside the calculator could mean anything (or more precisely, without an observer to interpret them, they can *mean* nothing). On the other hand, when we say that a being is conscious, we *are* postulating something extra above what we can view, namely the existence of observer-independent (but ontologically subjective[45]) experiences. Since thinking is not equivalent to consciousness, then the question of whether a machine can think is of little moral interest to us. Unconscious thought processes are not proper objects of moral empathy, since we literally cannot imagine what it is like to be having an

---

[42] Copeland, p.54; Winograd, Terry. "Understanding, Orientation, and Objectivity." In *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, edited by John Preston and Mark Bishop, Oxford ; New York : Clarendon Press, 2002: pp. 84-85.

[43] Searle, John R. *The Mystery of Consciousness*, New York : New York Review of Books, 1997: p.111

[44] Searle, John R. "Mental Causation, Conscious and Unconscious: A Reply to Anthonie Meijers." *International Journal of Philosophical Studies* 8, no. 2 (7, 2000): p.172.

[45] Searle writes that, "Mountains and molecules, as well as planets and tectonic plates, are ontologically objective. Pains, tickles, and itches on the other hand, are ontologically subjective. They exist only as they are experienced by human or animal subjects.", Searle, "Twenty One Years In The Chinese Room.", p.66

unconscious mental state; it is not *like* anything[46]. To summarise the discussion thus far:

1) A being is morally considerable iff that being is sentient.

2) If a being is sentient, then it is conscious.

3) A being is conscious iff there is something that it is like to be that being.

MORALLY SIGNIFICANT CONSCIOUS STATES

It is not just experiences of any sort of qualia that qualify a being for moral considerability. A being must be capable of having qualitatively *good* or *bad* experiences. A being must have these specific types of conscious states, since without them, moral sympathy would again be inappropriate. We cannot make sense of what it could mean to harm a being that was totally indifferent to all its conscious states(it is unlikely that such a being exists in nature, but it seems conceptually possible in the case of artificial beings). Suppose that, in the early stages of trying to create a fully-functional sentient android, scientists create a being that has some very basic conscious states, but has no good or bad conscious states. Suppose that this robot is conscious only insofar as it is able to see colours, but it can have qualitatively rich experiences of them. Apart from this colour perception and experience, the robot has no other conscious thought. Call this robot RAINBOW 1. It makes no difference to RAINBOW 1 whether it is in a blue

[46] Carruthers, Peter. "Sympathy and Subjectivity." *Australasian Journal of Philosophy* 77, no. 4 (1999): pp. 475-476.

room or a red room. Although it can experience the colour, it feels indifferent to which colour it is shown. There is something that it is like to be RAINBOW 1 seeing the colour red that is similar to the way it is for you to see the colour red. But RAINBOW 1 doesn't feel good or bad about red. It just sees it. Because of the lack of qualitatively good or bad experiences, it seems that it is impossible to morally harm RAINBOW 1 in any way. Would switching off or destroying RAINBOW 1 count as morally harming it? That is, all other things being equal, wouldn't it be better for RAINBOW 1 to experience colour than not? It seems that it cannot matter *to* RAINBOW 1, since RAINBOW 1 is indifferent about all colour experiences. It doesn't feel pleasure or pain associated with the experience, and it feels no more excitement about a bright colour experience than a dark one. If RAINBOW 1 were placed in room that become successively darker, there would be no point (even when it became pitch black) at which it would begin to feel discomfort. So it seems that switching off RAINBOW 1 wouldn't really matter to it. Although we may be able to empathise with RAINBOW 1, in the sense that we can imagine (to some extent) what it is like for it to see red, we cannot *sympathise* with RAINBOW 1 because, having no good or bad experiences, it is not an appropriate object for moral sympathy. Now suppose that the scientists develop a second robot: RAINBOW 2. RAINBOW 2 is just like RAINBOW 1 except that when it sees the colour red, it experiences this as a qualitatively bad experience, and when it sees blue, it has a qualitatively good experience. In this scenario, it seems that it would be possible to harm the robot by placing it in a red room, and to be kind to the robot by placing it in a blue one. The addition of good or bad conscious states has made the robot morally considerable. For an artificial being to be

morally considerable, it is not enough that it is merely conscious. We must also be able to say of its experiences that they are good or bad.

The results of the enquiries undertaken thus far indicate that if our crew member, Andrew, has phenomenally conscious experiences that are qualitatively good or bad, then he is morally considerable. Yet this hardly ends our dilemma , for even if Andrew does claim to have these experiences, how can we be sure that he really does have them? Does a machine's behaviour provide a reliable indicator as to its internal states? We need to know whether or not it is even possible for a machine like Andrew to have conscious experiences. The question "Can a machine be sentient?" can now be replaced with the question: "Can a machine have phenomenally conscious experiences that are qualitatively good or bad?" This is the question that is considered in the next chapter.

CHAPTER 3: ON WHETHER A MACHINE COULD SATISFY CONDITIONS FOR MORAL CONSIDERABILITY

"Every cognition involves a contribution of the observer. The sensory input provides some constraints, but the perceiver automatically corrects for any deficiencies in the data by interpreting the inputs in terms of strong assumptions and expectations. Under ordinary and familiar circumstances these unconscious inferences serve as veridical guides as to what is actually the case. But under special conditions they can badly lead us astray."

–Ray Hyman- *The Psychology of Deception*[47]

In the previous chapter, the question:"Can a machine be morally considerable?" was found to be equivalent to the question: "Can a machine have phenomenally conscious states that are qualitatively good or bad?" The goal of this chapter will be to try to provide an answer to this second question (and by logical necessity, the first). This question invokes two major problems in philosophy: first, a problem of metaphysics, encompassing what is commonly known as the Mind/Body Problem; and second, an epistemological problem, The Problem of Other Minds. I will call these, "The General Metaphysical Problem of Minds", and "The General Epistemological Problem of Other Minds", respectively. In their specifically machine-related incarnations, I will call them, "The Metaphysical Problem of Machine Minds", and "The Epistemological Problem of

[47] Ray Hyman, "The Psychology of Deception," *Annual Review of Psychology* 40 (February 1989): p.135.

Machine Minds", respectively. Note that, for reasons explained in the previous chapter, as it is used here, the word "mind" refers only to the type of mental state that is of most interest to ethicists: phenomenally conscious states that are qualitatively good or bad:

General Problems of Mind:

1) ***The General Metaphysical Problem of Minds:*** *"What kinds of things can have minds?"*

2) ***The General Epistemological Problem of Other Minds:*** *"How can we know if a thing has a mind?"*

Machine-related Problems of Mind:

1) ***The Metaphysical Problem of Machine Minds:*** *"Are machines the kinds of things that can have minds?"*

2) ***The Epistemological Problem of Machine Minds:*** *"How would we know if a machine had a mind?"*

The Metaphysical Problems and the Epistemological Problems are closely related; one of the ways we might resolve one of the Metaphysical Problems is by trying to find a solution to the relevant Epistemological Problem (and vice versa). The difficulty we have in resolving a Metaphysical Problem is partly due to the difficulty we have in finding a solution to related Epistemological Problem (and

vice versa). In this chapter I will address both the Metaphysical Problem and the Epistemological Problem as they relate to machines.

THE METAPHYSICAL PROBLEM OF MACHINE MINDS: *"ARE MACHINES THE* KINDS *OF THINGS CAN HAVE MINDS?"*

The only thing that is known to create consciousness is a brain, but it is not known exactly how this is done. The functionalist view of the mind holds that consciousness is the result of a particular method of processing information, and thus, it can be replicated in any substrate that allows the required method of information-processing to occur, while a related theory of mind, computationalism, is the theory that the mind is just a digital computer. According to both functionalists and computationalists, there is no reason why a machine could not have conscious states. The application of functionalism and computationalism to the creation of machine minds has led to the development of the field of Artificial Intelligence. It is popularly supposed that beings with computers for minds will be the first artificially intelligent beings, and this may be so. But if the distinction between thinking and consciousness is sound, then there is less reason to suppose that computer-based beings will be the first artificially *conscious* beings. Of all the properties of the human brain, why should its information-processing capabilities be the property that leads to consciousness? It might be supposed that this is so because beings that have greater or more complex information-processing capabilities also have greater, deeper, or more qualitatively rich conscious experiences; but it is impossible to argue this without begging the question. We don't know that the experiences of those with more complex brains have any more of a phenomenal quality that the experiences of those beings with less complex brains. And even if we did, we wouldn't know

whether information-processing capabilities were sufficient for consciousness, or were merely necessary, with consciousness requiring a particular sort of information-processing within a particular sort of material (organic brain matter).

Biological naturalism is the view that the brain produces consciousness because it is made of the appropriate *material* to make consciousness, rather than because of the way it manipulates information. John Searle is a biological naturalist, and he claims that consciousness is a biological phenomenon, like lactation or photosynthesis, and like those phenomena, it occurs only in certain types of material[48]. A functionalist or computationalist mind could theoretically be created by any material, but biological naturalism holds that, unless this mind is instantiated in a particular substrate, then it is unlikely that there will be anything that it is like to be that mind. Searle claims that any complex behaviour that results from computation occurring in a substrate that doesn't have the necessary biological properties would be merely a *simulation* of human or animal behaviour, and would not be accompanied by intentionality, understanding, or consciousness. He writes: "It is just as ridiculous to think that a system that had a simulation of consciousness and other mental processes thereby had the mental processes as it would be to think that the simulation of digestion on a computer could thereby actually digest beer and pizza"[49].

[48] Searle, John R. "Minds, brains, and programs," in *Behavioral and Brain Sciences* 3, no. 3 (September 1980): p.424.

[49] Searle, "Twenty One Years In The Chinese Room,", p.52.

Ned Block claims that Searle's argument "depends on ungrounded empirical speculation"[50], because Searle asserts that only brain matter (or an artificial substitute that was able to accurately synthesise the appropriate physical-chemical properties of brain matter) can create consciousness. But aren't Searle's opponents also guilty of speculation when they claim that certain information-processing tasks and their resulting behaviours are indicators of conscious experience? Since we currently have limited understanding of consciousness, anyone engaged in the task of synthesising machine consciousness will need to make some sort of prediction about how that result is most likely to be achieved. Searle's speculation that it is more likely to be achieved through replication of the brain's biological properties is no more ungrounded than the speculation of those who believe it is more likely to be achieved through replication of the brain's information-processing capabilities.

Searle's Chinese Room thought experiment is the most famous criticism of the claim that computational processes will lead to conscious experience[51]. Since the thought experiment will be familiar to many readers, a brief description should suffice: The Chinese Room involves an agent in a room who is given pieces of paper through a hole in the door, upon which are written various squiggles. Using a rulebook that contains nothing but formal rules (of the type "If you see "squiggle squiggle", respond with "squoggle squoggle"[52]), the agent writes "responses" to the input he has been given and passes them back out the door.

---

[50] Block, Ned. "Searle's Arguments against Cognitive Science," in *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, edited by John Preston and Mark Bishop, Oxford ; New York : Clarendon Press, 2002: p. 76.

[51] Searle, "Minds, brains, and programs"

[52] Ibid, p.419

Unbeknownst to him, the "squiggles" and "squoggles" are actually Chinese characters, being passed to him by native Chinese speakers. The point of Searle's thought experiment is this: even if the agent becomes so good at following his rulebook that native Chinese speakers outside the room believe that they are speaking to another native Chinese speaker, the agent inside the room will never really understand Chinese. By analogy, other symbol-manipulators that simply follow formal rules, like modern computers, don't really understand the information that they process either, regardless of adept at it they become.

Searle sometimes writes that the Chinese Room argument is intended to show that the system he describes (and thus, similar systems like today's desktop computers) has no "understanding"[53], or sometimes that it has no "intentionality"[54], but the thought experiment is equally effective as part of an argument demonstrating that computational systems need not have "consciousness", or at least the type of phenomenally-conscious states we are interested in. David Chalmers agrees. He writes that, despite Searle's use of multiple terms, "it is clear that consciousness is at the root of the matter."[55]

The Chinese Room has been the target of much criticism since it was first described in 1980. Many of these critiques point out that the thought experiment doesn't translate well into a valid argument. For example, Copeland notes that the argument appears to make an invalid inference from the conclusion that the agent in the room doesn't understand Chinese to the conclusion that the whole

[53] Ibid, p.418

[54] Ibid, p.424

[55] Chalmers, pp.322-323

system doesn't understand Chinese[56]. A similar point is made by Block, who claims that the whole Chinese Room system does in fact "think", even though the agent inside the room isn't conscious of it[57]. These replies are representative of the most successful criticisms of Searle' argument: they are variants of the "Systems Reply". The Systems Reply is a response to Searle's argument which claims that, even if the agent inside the Chinese Room does not understand Chinese, the whole system (consisting of agent, room, and rulebook) does. Copeland and Block argue that, if the Chinese Room Argument is supposed to show that a system like the Chinese Room does not *think*, or *understand* Chinese, then it fails, for as we have seen, the usage of the word *think* and *understand* do not suggest the existence of any observer-independent phenomena. If English-language speakers decide to use these words to describe what the Chinese Room is doing, then Searle's claim that the Chinese Room does not understand is simply false.

But if we recall that "consciousness is at the root of the matter", and recall that the type of consciousness we are interested in are phenomenal states, then the Systems Reply is on much shakier ground, for the claim now becomes one about the existence of an observer-independent phenomenon: subjective experience. Would successful communication with native Chinese speakers (to the point that those outside the room believed that they were speaking to another native Chinese speaker) by a system involving a room, an agent following formal rules, a book, and some pieces of paper, result in the creation of phenomenal

[56] Copeland, p. 125.

[57] Block, Ned, "Searle's Arguments against Cognitive Science", p.73.

experiences for that system? Although we cannot prove that it doesn't, it certainly seems very unlikely.

Block concedes that, although he believes that the Chinese Room system "thinks", there probably isn't anything that it is like to be the Chinese Room[58]. And Block also suggests a similar thought experiment: the Chinese Nation. In this experiment, Block imagines a body being controlled by the population of China, using radio communication to simulate the communication that takes place between neurons. Again, Block concludes that, although the Chinese Nation could conceivably control a body that behaved like a human, it is not likely that it would experience qualia[59].

Searle and Block rely on our intuitions that rooms and groups of people linked by radios are not the type of things that would have phenomenal consciousness. But it has been argued that human brains are themselves unlikely candidates for consciousness; isn't it rather surprising to find consciousness in a lump of meat? This might seem just as odd as the idea that a room could have a subjective experience, yet we *do* find consciousness in meat, in our own brains. Perhaps (the objection goes) our intuitions about consciousness in rooms and people linked by radios are just as flawed as the intuitions we have about consciousness in meat. In response, Searle points out the peculiarity of the claim that it should be surprising to find consciousness in a lump of grey flesh. After all, he notes, isn't that the *only* place that we know of that *does* have consciousness?

---

[58] Ibid, p.74.

[59] Block, Ned. "Troubles with Functionalism," in *Readings in Philosophy of Psychology, Volume I*, edited by Ned Block, Cambridge, Mass: Harvard University Press, 1983: pp.276-278.

Why then would this be at all surprising?[60] Block also points out that there is a difference between our intuitions that brains shouldn't have qualia, and our intuitions that Chinese Room or Chinese Nation-type beings (or "homunculi-heads", as Block calls them[61]), probably don't. Given the evidence of our own subjective experience, and our understanding of medical science, we have reason to doubt our intuition that brains seem unlikely seats of consciousness. On the other hand, we have no reason to doubt the intuition that "homunculi-heads" don't. A rational person would accept the evidence that brains can be conscious, but can remain sceptical that inorganic information-processing systems can.

Computational and functionalist theories of mind provide an explanation of the difference between a thinking and non-thinking thing, but not of the difference between a conscious and a non-conscious being. Without being able to make this distinction, we cannot know whether machines have the properties that would make them morally valuable. Machines whose minds duplicate only the functional properties of brains may have all the information-processing properties of a human, but they may lack subjective phenomenal consciousness. It is possible that a machine could be the kind of thing that has a mind, but with a limited understanding of consciousness, we cannot be sure.

THE EPISTEMOLOGICAL PROBLEM OF MACHINE MINDS: *"HOW WOULD WE* KNOW *IF A MACHINE HAD A MIND?"*

Suppose that it *was* the brain's information-processing properties that lead to consciousness, how would we know if the information-processing

[60] Searle, *The Mystery of Consciousness*, p.158

[61] Block, Ned. "Troubles with Functionalism", pp. 281-283.

methods that we use in our computing machines are the *right* ones? Our computers might achieve the appropriate result outwardly, but since much of information-processing that occurs within our own brains occurs without consciousness, how would we know which of a machine's information-processing was occurring with accompanying subjective experience, and which was not? How would we know if *any* of it was?

As we have seen, the subjective viewpoint is of special significance to ethics, so it is important that we are able to verify the presence of this special point-of-view in any agent we might encounter. Moreover, it is vital that we are able to verify the presence of subjectivity scientifically, or from a third-person perspective. Yet, as Nagel points out, subjective experience is *irreducibly* subjective; any attempt to describe consciousness from a third-person perspective will inevitably sacrifice the first-person perspective, thereby abandoning exactly what it is that makes consciousness unique: its subjective viewpoint[62]. If subjective experience cannot be verified scientifically, we will find it impossible to ever be sure about whether or not consciousness is present, even in fellow human beings. The problem, which I will label The Epistemological Problem of Other Human Minds, is summed up by Alan Turing as follows:

> A is liable to believe 'A thinks but B does not' whilst B believes 'B thinks but A does not'. Instead of arguing continually over this

[62] Nagel, p. 437.

> point it is usual to have the polite convention that everyone thinks.[63]

Given the explanatory gap between objective phenomena and subjective consciousness, the problem is perhaps insurmountable, yet it is almost always ignored in daily life, perhaps for the sake of "polite convention". In truth, no conscious being can know with definite certainty whether any other being is conscious. The only thing that they can do is *infer* that others are conscious from the fact that others behave or look similar to them. This is an inference known as the "argument by analogy". Bertrand Russell describes the argument by analogy as follows:

> If, whenever we can observe whether A and B are present or absent, we find that every case of B has an A as a causal antecedent, then it is probable that most B's have A's as causal antecedents, even in cases where observation does not enable us to know whether A is present or not.[64]

So, if A is a conscious mental state of feeling thirsty, and B is the act of saying "I am thirsty", then the argument by analogy tells us:

> *In situations in which I can observe whether or not A and B are present (my own subjective experience):*

---

[63] Turing, A. M. "Computing Machinery and Intelligence," *Mind* 59, no. 236, New Series (October 1950): p.446.

[64] Russell, Bertrand. *Human Knowledge: Its Scope and Limits*, London: Allen and Unwin, 1948: p.505.

1) *The action of saying "I am thirsty" (B) only occurs when there also exists a mental state with the conscious content of feeling thirsty (A).*

Therefore, it is highly probable that:

2) *Most situations that consist of saying "I am thirsty" (B), occur when there also exists a mental state with the conscious content of feeling thirsty (A).*

Even in situations where we are unable to verify the presence of both A and B (observing the behaviour of others).

The argument by analogy thus suggests that we have good reason to suppose that other humans have the same or similar mental states to ourselves.

Can we use the argument by analogy to solve the Epistemological Problem of Machine Minds? Some commentators think so[65]. They claim that the Problem of Machine Minds involves many issues that are also evident in our own daily interactions with other humans. They therefore argue that, since these issues are not overly significant obstacles for attributing consciousness to humans, they ought not to stop anyone from attributing consciousness to a machine. Since we are so good at ignoring the problem when communicating with our fellow humans, why can we not do the same (the claim goes) when communicating with

[65] Those who make this claim include: Penrose, Harnad, S. "The Turing Test is not a trick: Turing indistinguishability is a scientific criterion," *ACM SIGART Bulletin* 3, no. 4 (1992): 9–10., and Leiber, Justin, *Can Animals and Machines be Persons?: a Dialogue*, Indianapolis, Ind.: Hackett Publishing, 1985

robots? After all, I don't doubt that when Adam says he has a headache, he really has one. So why should I doubt it when Andrew says he has a headache? As Stevan Harnad points out, when communicating with a pen-pal, we don't need to examine the pen-pal physically to infer that they have a mind. To ask for more proof for the existence of conscious states from a machine is "arbitrary"[66].

Yet these critics are fundamentally mistaken. There are significant differences between The Epistemological Problem of Other Human Minds and The Epistemological Problem of Machine Minds. The argument from analogy allows us to infer that other humans have minds, but the further away from a human brain we move, the less the analogy holds. We find as we move a relatively small step away from humans to the question of whether non-human animals are conscious (what might be described as the Epistemological Problem of Animal Minds), it is taken for granted that animals are conscious almost as often as it is that humans are. But sometimes, even the consciousness of non-human animals is doubted. Singer addresses the Epistemological Problem of Animal Minds in *Animal Liberation*, in which he expounds a view that those animals that are further down the evolutionary ladder, and thus holding fewer properties in common with humans, are less likely to be conscious[67]. Singer responds to those who express scepticism about the idea that animals can feel pain by giving 3 reasons to infer that animals have similar experiences of pain to our own. He points out that animals share with humans:

[66] Harnad, p.9.

[67] Singer, *Animal Liberation*, pp.171-172

1) ***Similar behaviour:*** *Many animals express many of the same or similar types of pain behaviour that we have when we are in pain (for example, crying out, moving away from the negative stimulus).*

2) ***Physical similarity:*** *Many animals possess nervous systems that are physically similar to those possessed by human beings.*

3) ***An evolutionary link****: Since animals share a common evolutionary origin with us, it is rational to suppose that their similar biology functions the same way as ours. Pain avoidance is also evolutionary advantageous, so we should expect it to be common in higher animals like mammals.*[68]

With these arguments, Singer provides an argument by analogy that answers the Epistemological Problem of Animal Minds, but it is equally applicable to the Epistemological Problem of Other Human Minds. We think that other humans and animals have minds, not only because they behave like us, but also because they *look* like us. From the fact that you look like me, I can infer that you share the similar underlying physical systems, including a nervous system. Importantly, I also infer that we are both the products of millions of years of evolution. In terms of mental states, I conclude from the former inferences that what goes on inside our heads should be expected to serve the same evolutionary function (whatever

[68]Ibid, pp. 10-11.

that function might be), so it is likely to be similar. Singer writes of other animals that:

> It is surely unreasonable to suppose that nervous systems that are virtually identical physiologically, have a common origin and a common evolutionary function, and result in similar forms of behaviour in similar circumstances should actually operate in an entirely different manner on the level of subjective feelings.[69]

But the same analogy cannot be made with machines. Even if robots are made to be behaviourally or outwardly physically indistinguishable from humans it would not be reasonable to assume that a being that acts or looks the same as oneself is "virtually identical physiologically" under the skin, or that it has "a common origin" and "common evolutionary function[s]". And this means that, unlike in the case of other animals, or other humans, it would not be "unreasonable" to suppose that its mind operated "in an entirely different manner on the level of subjective feelings". The analogy between humans and machines is less strong than the analogy between my own case and that of other humans, or even that of my own case and that of other animals, and as a result, the argument by analogy for the presence of machine minds is significantly weaker.

By way of comparison, consider each of the reasons that Singer provided as reasons to infer that animals have minds, as they apply to machines. In each of these cases it will be clear that the argument by analogy provides a much weaker reason for inferring the presence of consciousness in machines than it does in the case of other humans or animals.

---

[69] Ibid, p.11.

SIMILAR BEHAVIOUR

Machines may indeed have similar behaviour to humans, but unlike in the cases of humans or animals, this behaviour may not be a clue to subjective experiences in their minds. If the machine is developed using a "top-down" method of design, then it will be designed with a specific behaviour in mind. The actual method used to create that behaviour need not coincide with that used to create similar behaviour that appears in humans or animals. The problem with all behavioural evidence for consciousness is that is seems possible to imagine a being that outwardly presents behaviour that is indistinguishable from conscious human behaviour, but that has no conscious states (the philosophical zombie). There is no way to tell whether or not a being that had passed a behavioural test was in fact just a variety of robot zombie, or whether it had really experienced phenomenal consciousness. It certainly seems plausible that an artificial being that evolves a tendency to avoid harmful situations, for example, could still lack the pain experiences that we might expect to find in natural beings displaying such behaviour. Although our own evolution resulted in the development of qualitatively bad pain experiences alongside the pain behaviour, there is no certainty that this was necessarily so, since it doesn't seem to be pain feelings that are evolutionarily advantageous, but pain-avoidance behaviour[70]. In his description of the argument by analogy, Russell writes that the argument by analogy does not allow us to conclude that it is definitely the case that B is always

[70] Harrison, P. "Do animals feel pain?," *Philosophy* 66, no. 255 (2009): p.32.

caused by A, but only that it is "highly probable" that B is always caused by A[71]. It is a fallacy of reasoning to make the inference that just because A and B always occur together that they must *necessarily* always occur together. In particular, the argument by analogy does not allow us to come to the conclusion that things that are true of natural entities are also true of artificial ones. For example, a person living before the 18th century might have felt justified in making the following claim:

> *In situations in which I can observe whether or not A and B are present:*
>
> 1) *It is true that an entity flies (B) only when that entity has wings (A).*
>
> *Therefore, it is highly probable that:*
>
> 2) *Any entity that flies (B) has wings (A).*

But the existence of hot air balloons and helicopters demonstrate that wings are not necessary for flight. In the 21st century we note that there are many times in which B is present, but A is not. Similarly, we sometimes see claims made by roboticists to the effect that:

> *In situations in which I can observe whether or not A and B are present:*

[71] Russell, p.505.

1) *If a naturally-occuring entity has the (outwardly visible) properties of a person (use of language, reasoning, thinking, planning for the future, etc.) (B), I can infer the presence of consciousness (A).*

Therefore, it is highly probable that:

2) *For* any *entity that has the (outwardly visible) properties of a person (use of language, reasoning, thinking, planning for the future, etc.)(B), I can infer that they must also be conscious (A).*

In his entertaining article ""If Droids Could Think..." Droids as Slaves and Persons", Robert Arp makes a claim of this sort, implying that we can infer that the droids in *Star Wars* can feel fear because they behave as if they do[72]. This is a prime example of use of the argument by analogy to draw an erroneous conclusion; "fear behaviour" usually indicates the presence of fear in natural animals, but not necessarily in artificial ones. These inferences are *not* reasonable to make for artificial creations because, as Joanna Bryson and Phil Kime note, "aspects of cognition do not automatically come with others". We have developed phenomenal consciousness alongside behavioural properties such as the ability to use language, but "our particular mix is the product of millions of years of evolution."[73]. It is possible that, through our own attempts at artificial evolution,

[72] Arp, Robert. ""If Droids Could Think..." Droids as Slaves and Persons," in *Star Wars and Philosophy*, edited by Kevin S. Decker, Jason T. Eberl, and William Irwin, Chicago: Open Court, 2005: pp. 120-131.

[73] Bryson, Joanna and Kime, Phil. "Just Another Artifact: Ethics and the Empirical Experience of AI," presented at the *Fifteenth International Congress on Cybernetics*, Namur, 1998: p390.

we might create a being with most of the outward properties of a person, but without it having any conscious awareness.

Consider one prominent example of a behavioural test for consciousness: Alan Turing's Imitation Game (more commonly known as "the Turing Test). In describing the test, Turing intended to provide a replacement for the question "Can machines think?", a question which he thought "too meaningless to deserve discussion"[74]. Turing seemed to think that the decision about whether a computer could be said to "think" was a linguistic decision, and that answering the question told us nothing interesting about the machine in question. Instead, Turing proposed an imitation game and a corresponding question, "Are there imaginable digital computers which would do well in the imitation game?"[75]. I have already argued that the question "can machines think?" is a relatively trivial one (perhaps there already exist some machines that can properly be described as "thinking" machines, but there is nothing morally significant about this); the most important test for ethicists is one that answers the question of whether or not the machine has conscious states. Note that it is not clear whether Turing himself wanted to know whether a machine had conscious states (in fact, it is likely that this was not the original aim of the test[76]), nevertheless, the Turing Test provides a convenient example of some of the problems of behavioural tests for consciousness.

[74]Turing, p.442

[75] Ibid, p.442.

[76] Turing's famous paper includes the following statement: "Although consciousness is mysterious, problems of consciousness don't need to be solved before we can answer the question with which we are concerned in this paper.", Turing, p.447

In the Imitation Game, an interrogator sitting at a computer terminal receives responses to questions posed to both a person and a computer, located out of sight of the interrogator. The interrogator tries to ascertain which responses are coming from a person, and which are coming from a computer. The computer "passes the Turing test", when the interrogator mistakes the computer's responses for those of a person. Turing developed the Imitation Game because he thought that it would be difficult to convince people that computers could think. The test was designed to overcome what Turing supposed would be a natural bias against attributing minds to machines. But Turing did not have any evidence for his supposition; he really had no idea how people might react to a talking computer, because nobody had ever seen one. Is it possible that most people will be *too* willing to ascribe agency, the presence of a mind, or phenomenal consciousness to computers and other machines? In fact, humans do have a natural tendency to anthropomorphisise, a trait that probably arose due to its evolutionary advantage. Human beings are said to have a "Hyperactive Agency Detection Device" (or HADD), making them over-sensitive to the presence of intentional agents, particularly when faced with an unfamiliar or uncertain situation[77]. Triggering of the HADD is usually caused by an object moving or behaving in a way that is not expected by our intuitive understanding of the way such objects are supposed to behave, and this is true regardless of whether or not the object looks like a typical agent[78]. This suggests that Turing was wrong to suppose that people would struggle to see mechanised entities as agents with

[77] Barrett, J.L. "Exploring the natural foundations of religion," *Trends in Cognitive Sciences* 4, no. 1 (2000): 31-32.

[78] Barrett, J.L. and Johnson, A. H. "The role of control in attributing intentional agency to inanimate objects," *Journal of Cognition and Culture* 3, no. 3 (2003): p.215.

intentional states; in fact, it suggests that people might be too eager to do so. Interpretations of the world that involve the presence of agency were probably selected for during human evolution because they can reveal the presence of enemies, mates, or food. It seems that it was better for the survival of early humans to over-attribute agency and risk wasting time searching for a predator or prey that did not exist rather than to under-attribute it, and miss out on a potential meal, or risk being killed by an enemy that they did not see.

Given the existence of the HADD in humans, we ought to be sceptical of positive Turing Test results, or positive results of any behavioural test for machine consciousness. If human beings have a tendency to ascribe consciousness where none exists, we should be surprised if there were not more than a few false positives in a behavioural test for phenomenal consciousness. While Turing was right to design a test that avoided natural human biases, he may have been misguided about which directions our natural bias would lead us; the Turing Test, and other behaviourally-based tests for consciousness, rather than discouraging bias, allow the natural tendencies of the HADD towards attributing intentionality to flourish. Moreover, because our minds have evolved in an environment in which all things that have behaved consistently in a human-like way have in fact been other humans, we naturally ascribe human agency to things with behaviours typically only seen in humans (for example, language, or the use of moral concepts). Once we have developed the ability to simulate these behaviours in artificial creations, this previously useful mental tool could become a liability.

What a computer that successfully passed the Turing Test could demonstrate is that it is possible for an entity to display patterns of "conscious behaviour" (the kind of behaviour that is usually accompanied by, or created by consciousness), using a method (in this case, "the rigorous execution of algorithms" [79)] entirely different from that normally used by human beings to create the same behaviour. Indeed, as Halpern has pointed out, the Turing Test is perhaps best viewed not as a test to differentiate between a human and a machine, but rather as a test to differentiate between the use of natural thought processes and the use of systematic algorithms[80]. Even if the cognitive procedure normally used by human beings for a particular type of behavioural output always involves consciousness, there is no reason to suppose that the method used by a machine that duplicates this behaviour does.

The Turing Test was developed to test computational devices in particular for human-like behaviour, but the same criticisms can be directed at any test that relies solely on behavioural output, regardless of what kind of device it is testing. Indeed, this is precisely the point: the results that we see could plausibly be the result of *any* method, including any that are not the same as those which are used by the human brain, and may or may not coincide with the existence of a subjective consciousness. Throughout most of human evolution, behavioural cues have been reliable indicators of morally relevant properties such as self-awareness, sentience, and consciousness. In the age of androids this may no longer be the case.

---

[79] Halpern, M. "Turing's test and the ideology of artificial intelligence," *Artificial Intelligence Review* 1, no. 2 (1987): p.88.

[80] Ibid, pp.88-89.

PHYSICAL SIMILARITY

Since we can only be certain of the consciousness in our own brains, the less physically similar a being's brain is to a human brain, the less likely it is that it will have subjective conscious. Animals that have brains vastly different from our own, such as invertebrates, are typically considered less likely to be conscious, and we are most confident in inferring the presence of consciousness in higher mammals with brains relatively similar to our own. The exact degree of physical similarity between humans and future robots is impossible to predict, but given the fact that the same behaviours could come from different mechanisms, there could be a large degree of variance (as there is currently between the physical mechanisms of flying animals and those of flying machines like aeroplanes, for example). The greater the degree of variance between artificial brains and our own brains, the less successful any appeal to the argument by analogy will be.

It is also worth noting that there will be no way of knowing whether any given artificial brain has adequately duplicated the *right* parts of the brain to give rise to consciousness. Even the duplication of the neural structure of a natural brain in an artificial brain is no guarantee that the artificial brain will be conscious. If biological naturalists like Searle are correct, and natural brain tissue itself contains essential properties for creating consciousness, then machine brains will not be conscious until such time as those properties can be defined

and synthesised artificially. The problem is that we cannot know whether we have synthesised the correct properties of neural tissue, or whether we have missed a fundamental element, since behavioural evidence that we had succeeded would be unreliable.

EVOLUTIONARY LINK

Perhaps most importantly, an artificial being will lack a common evolutionary link with humans. By recognising that other humans and animals are physically and behaviourally similar to ourselves, we infer that they have similar origins and, as noted, this makes it reasonable to suppose that their behavioural and physical properties have the same function as our own. This difference in the method by which the machine's brain is created may turn out to be the most significant factor in determining whether a machine has conscious experiences or not. In particular, we ought to expect that a machine developed using common "top-down" methods of design may develop conscious behaviour without the accompanying subjective states, while a machine that has been developed through "bottom-up" methods may be more likely to be phenomenally conscious. This is because it may be possible to develop machines through a sort of artificial evolution, which replicates the environment in which consciousness evolved. In such a situation, the analogy with humans is stronger, since similar physical structures and behaviour should have similar evolutionary functions. A property like consciousness, that developed from evolutionary processes, likely serves a function that gives a being that has it some level of evolutionary advantage. Beings that are artificially evolved in relevantly similar circumstances should benefit in similar ways from the development of consciousness, and so in

this limited case, there are reasons to suppose that a machine that had similar behaviour to a natural being, might also possess the appropriate accompanying subjective state.

Nevertheless, there will remain differences between human brains and those of any artificially evolved machine that will leave room for doubt; it will be impossible to duplicate all the selection pressures that were brought to bear on the human brain during its evolution. It is quite possible that our attempts to create conscious beings through artificial evolution may result in the creation of robot versions of philosophical zombies. Of course, it might be the case that it is far easier to create a being that does the same things as an animal if it is also conscious. Nicholas Humphrey argues that being able to experience pain and pleasure has an evolutionary advantage, since it makes the being care about what is happening to it, and presumably this leads to better pain-avoidance behaviour, and thus to better survival capabilities[81]. But even if consciousness creates better pain-avoidance in natural environments, it is surely not the *only* way that such behaviour might develop. The extraordinary variety even among natural creatures is testament to the many ways that problems can be solved through evolutionary means, and given that machines will be composed of different materials to natural beings, there is every reason to suppose that different solutions could be found to similar problems. Artificial evolution would also likely differ from natural evolution in terms of:

1) The nature of the environment

[81] Humphrey.

2) Different selection pressures

3) Vastly different time scales

4) The degree that it is consciously controlled and goal-directed

These all weaken the analogy with naturally-evolved beings and our inference of the presence of conscious states in artificially-evolved beings.

THE PRINCIPLE OF DISANALOGY

It is clear that requiring more proof for the existence of consciousness from a machine is not at all "arbitrary". There are several relevant differences between the behaviour, physical appearance, and the method of creation of machines, and those of naturally-evolved beings, that means that the argument by analogy is weaker in the case of machines than in that of humans. Note that I have not tried to show that machines *could not* be conscious. In fact, it is possible that they could. What I have attempted to demonstrate in this chapter is that the answer to the question posed at the beginning of the chapter ("Can a machine have phenomenally conscious states that are qualitatively good or bad?") is that we cannot know. In our attempts to verify the presence of consciousness in artificial beings we only have the natural brain as an example for comparison. This makes the Epistemological Problem of Machine Minds more significant than some other incarnations of the General Epistemological Problem of Other Minds. Since consciousness is irreducibly subjective, the possibility of being able to verify the existence of consciousness in another being from an objective perspective is very low; we cannot, in fact, verify its existence directly, but only

through inference. The argument by analogy, which goes some way to resolving the problem in humans and animals, is less successful when used to determine whether machines can be conscious, due to several significant differences between machines and naturally-evolved beings. This is the third principle of utilitarian machine ethics:

> ***The Principle of Disanalogy:*** *The problem of machine minds is greater than the problem of other human minds.*

This will have significant ethical implications, since it seems that if the question of whether a machine could be morally considerable rests on their ability to have phenomenal consciousness, then as we come to the end of chapter 3, we remain in doubt as to whether a machine could be morally considerable. In the ethical dilemma described earlier, in order to know whether Andrew was morally considerable, we needed to know whether or not Andrew could have phenomenal states that are qualitatively good or bad. Now, not only do we not know whether or not Andrew is the kind of thing that *can* have these states, but even if we did, we could not verify it. We cannot know whether Andrew is conscious at any particular time, or for that matter, at any time at all.

CHAPTER 4: CAN A MACHINE BE A PERSON?

"As we continue to push our technology forward, and as related technology progresses, we expect that Hanson Robotics' robots will evolve into socially intelligent beings, capable of love and earning a place in the extended human family."

- Hanson Robotics Vision Statement, 2010[82]

Since robots like Data, or our crew member, Andrew, are behaviourally so similar to humans, it is inevitable that we should ask whether they could be persons. The answer to this question would have significant implications for our dilemma, for if Andrew could be a person, then it would be possible to harm him by "killing" him (whether or not a being that is not alive can truly be "killed" is a trivial matter. There are ways that we might end the existence of an inorganic being, and it is this ending of a being's existence that is meant here by the term "killing"). We already know that it is possible to harm Adam by killing him, since Adam is definitely a person. Because the choice we have to make involves sending one of the crew to certain death, we know that it would be a bad thing to send Adam, and we should avoid it if we can. The situation becomes a genuine dilemma if Andrew is also a person, and thus worthy of protection from wrongful death too. Is Andrew a person? Could a machine be a person?

Although there is no one unanimously accepted definition of personhood, and many philosophers have offered their own interpretation of what properties

[82]Hanson Robotics, "Our Mission is to Realize Friendly Super-Intelligent Machines" (n.d) http://www.hansonrobotics.com/vision.html, accessed 11/02/10

a person must have, there is some degree of consensus on what it means to be a person. Usually definitions of personhood make reference to three distinct types of properties. I have listed each of these types, along with a representative example of a description of personhood of the relevant type:

1) ***Persons are cognitively special.*** *For example, they are rational...*

> [A person is…] a thinking intelligent being, that has reason and reflection, and can consider itself as itself, the same thinking thing, in different times and places- John Locke [83]

2) ***Persons are socially special.*** *For example, they can communicate with each other, and form relationships...*

> [Persons]...belong to a community of rational, valuing beings and it thus belongs to the essence of personhood that a person be at least in principle capable of communicating, by means of symbols and signs, in a shared space of interpersonal meanings and shared experience of pleasure, pain, desires, hopes, fears, etc. This is, broadly, the notion that what is distinctive about persons is the possession of language.- Keith Price[84]

3) ***Persons are morally special.*** *For example, they can be both moral patients and moral agents...*

---

[83] Locke, John. *An Essay Concerning Human Understanding* Oxford : Clarendon Press, 1975: p.192.

[84] Price, Keith, "Why Computers Will Never Be People," in *Selected Papers from Conference on Computers and Philosophy*, Volume 37, 2003: p. 2.

> The fact is that moral relations...are possible only between persons, and that any entity with which one can conceive having moral relations would be a person.- Roland Puccetti [85].

Of particular importance to the topic at hand is the fact that it is commonly accepted that the term "person" is a notion distinct from the term "human being". For example, in *Rethinking Life and Death*, Singers lists great apes, whales, dolphins, elephants, monkeys, dogs, and pigs as non-humans that might possibly have the right properties to be described as "persons"[86]. Since the principle of equal consideration of interests prohibits us from discriminating against machines on the basis of the material from which they are made, there is no reason not to attribute personhood to non-biological entities that also possess the appropriate properties. If Andrew is conscious, then he could have all of the properties usually ascribed to persons. If conscious, Andrew could have:

1) *Interests and goals.*

2) *The ability to use language.*

3) *The ability to form reciprocal relationships*

4) *The ability to understand and apply moral concepts*

5) *Rationality*

---

[85] Puccetti, Roland. *Persons: A Study of Possible Moral Agents in the Universe*, London, Melbourne: Macmillan, 1968: p.26.

[86] Singer, Peter. *Rethinking Life & Death: The Collapse of Our Traditional Ethics.* Melbourne : Text Publishing, 1994: p.182

6) *Self-awareness*

7) *A personality*

These are all properties of persons, and are jointly sufficient for personhood. The property of being a biological being is never considered a necessary property for personhood, and nor should it be. If personhood transcends species barriers, then this is because the properties that define personhood are not bound to the biology of a human being, and in fact (if it is possible for a non-biological being to be conscious) there is no reason why they should be bound to *biology* at all. If Andrew is conscious, then these properties would endow him with the moral protection that the label of personhood implies, and Andrew would have the equivalent moral status of a normal adult human. According to the principle of equal consideration of interests, to treat Andrew as anything less than a person would be substratism, because it would imply that Andrew's interests are less important than those of other beings with the same (morally-relevant) properties, because of the material of which he is made (an arbitrary criterion), and it would be a position no more defensible than speciesism or racism. All this follows from the stipulation that Andrew is conscious. If Andrew is conscious, then he is a person.

But what if he were not conscious? Even without phenomenal consciousness, Andrew would still possess many of the characteristics of persons, and he would still behave as if he was one. We might call the type of personhood that a non-conscious being could have "*behavioural personhood"*, to signify that a being has (at least) the outwardly visible properties of personhood. This

distinguishes behavioural personhood from "*full personhood*", which consists of behavioural personhood as well as the subjective states that we associate with personhood, such as self- awareness, intentional states like an understanding of language and moral concepts, and future-oriented preferences.

It seems plausible that machines of the future could have behavioural personhood, but what would their moral status be? Is behavioural personhood morally equivalent to full personhood? In fact, this seems doubtful; most definitions of personhood agree on making at least some reference to consciousness (for example, Tooley calls consciousness a necessary "starting point" for a definition of personhood, completely ruling out the possibility that a non-conscious being could be a person[87]). This focus on consciousness could be considered merely representative of the fact that no beings currently exist that have the cognitive, social, and moral capabilities of persons which do not also have consciousness. Perhaps the development of advanced social robots with human-like behaviour would lead some philosophers to include some non-conscious robots in the category of persons. It is more likely though, that the focus on consciousness comes from the requirement that a person be a moral patient as well as a moral agent. If a being is not sentient, then according to Singer, it cannot be harmed. This entails that it cannot be a person, since personhood is a moral category, which requires that a being be capable of being harmed or benefitted. Moreover, a lack of phenomenal consciousness would have implications for the capacity of the being in question to actually have the properties predicated of it. We might require that some of the qualities of

[87] Tooley, p.88.

personhood have their foundation in conscious experience to be genuine examples of that quality, so that a non-conscious being could not be said to truly possess a particular property if it was not conscious. It is not clear for example, whether a being that is not conscious could be said to be capable of having a "real" relationship, or if it could be said to "actually" have self-awareness if it did not have phenomenal awareness. These are complex issues, but I will stipulate that if a being is not conscious, then it cannot have full moral personhood.

Even though behavioural personhood might not endow its possessor with the equivalent moral status of full personhood, some writers on the subject of machine rights have stressed the importance of behavioural personhood for a machine's moral standing. In Mark Coeckelbergh's 2010 article *Robot rights? Towards a Social-relational Justification of Moral Consideration*, he attempts to formulate a way to give moral consideration to particular machines with advanced human-like social capabilities[88]. Coeckelbergh suggests that, because of the unusual mix of properties that androids might have (advanced social skills, and social roles, but a low probability of consciousness) we might give certain robots "soft rights", while human persons would have "hard rights"[89]. Steve Torrance also acknowledges the unusual place of a human-like robot in our moral and social world. Torrance writes that we might have to redefine some of our moral concepts to include an understanding of "quasi-moral relationships"[90] between humans and robots, and talks of some humanoid robots as being "para-

[88] Coeckelbergh, Mark. "Robot rights? Towards a social-relational justification of moral consideration," *Ethics and Information Technology* 12, no. 3 (6, 2010): pp.209-210.

[89] Ibid., 218.

[90] Torrance, p.504.

persons"[91]. Kahn et al. have also suggested that androids might confound some of our traditional understandings of the world; they wonder whether the advanced social robot might represent "a new technological genre" that blurs the distinction between living and non-living or animate and inanimate[92]. Central to all of these claims is a recognition that a non-conscious android would provide us with a truly unique case in moral philosophy. Throughout the history of human evolution, we have never encountered a being like a non-conscious android would be: a being that had highly advanced social and cognitive skills at levels equal to or greater than human beings, but which lacked phenomenal consciousness. Inspired by the idea of "soft rights" for robots suggested by Coeckelbergh, and by Torrance's suggestion that humanoid robots of the future might be "para-persons"[93], I suggest that we might distinguish between "soft" persons", "hard" persons, and "fuzzy" persons. "Hard" persons are persons in the traditional sense; as most philosophers agree, they must be conscious. "Soft" persons are beings that have behavioural personhood, but which lack consciousness. "Soft" persons would be very much like the philosophical zombies of philosophy-of-mind literature[94]. A "fuzzy" person is a being that has behavioural personhood, but whose consciousness is in doubt. A "fuzzy" person is really either a "hard" person or a "soft" person, but we may never know which they really are. Because of the Epistemological Problem of Other Minds, all persons (apart from oneself) are technically "fuzzy" persons, but reasons were

---

[91] Ibid, p.518.

[92] Kahn, P.H., Freier, N.G., Friedman, B., Severson, R.L. and Feldman, E.N., "Social and moral relationships with robotic others," in *Proceedings of the 13th International Workshop on Robot and Human Interactive Communication (RO-MAN'04)*, 2004: p.549.

[93] Torrance, p.518

[94] Ibid, p.500

given in the third chapter as to why a machine person would be “fuzzier” than a human person. Machines like our crew member Andrew and the character Data would certainly be classed as “fuzzy” persons. If a “fuzzy” person turns out to be a “hard” person, then their moral status has already been established in this work and in many other works on the moral status of persons. If they are a “soft” person however, their status is not so clear. We therefore need to establish what the moral status of a “soft” person would be, in case any “fuzzy” person is actually a “soft” person.

“SOFT” PERSONHOOD

What is the moral status of “soft” persons? If behavioural personhood is not full moral personhood, where does it stand in relation to it? In Singer's worldview, a person is the only being that can be harmed by being killed. Could a “soft” person be harmed in this way? Let us examine the reasons given by Singer for why it is usually wrong to kill a person, to see whether those reasons might also apply to “soft” persons. On a very basic level, Singer claims that killing a person deprives them of the ability to experience any pleasure in the future. The killing of a person thwarts not only their preference to go on living, but also a multitude of preferences for the future that depend on the person's continued existence for their satisfaction. But there are more significant wrongs than this; else it would be equally wrong to kill a sentient non-person, which Singer denies. Singer writes that: “According to preference utilitarianism, an action contrary to the preference of any being is, unless this preference is outweighed by contrary preferences, wrong. Killing a person who prefers to continue living is therefore

wrong, other things being equal."[95], he also writes that "every person has a right to life. We have seen that the basic reason for taking this view derives from what it means to be a person, a being with awareness of her or his own existence over time, and the capacity to have wants and plans for the future."[96] As well as these "direct" harms, Singer stresses that there are "indirect" harms that result from the killing of a person. These include: the grief caused to the friends and family of the person killed; the anxiety caused to other persons who fear for their lives as a result of seeing another person killed; and the threat to the "peaceful coexistence on which society depends"[97] Singer notes that many nonutilitarians consider these indirect harms "side effects", but he highlights their importance for our moral judgement of murder, saying, "I am not sure that we should, in the case of normal human beings, allow these "side-effects" to be so lightly brushed aside"[98].

It seems quite plausible that a technologically-advanced robot could express a preference for continuing to exist. It also seems plausible that it could demonstrate by its behaviour that it had this preference. But since sentience is a necessary property for a being to have in order for its preferences to be morally significant, the preference that the robot expresses for staying alive is not the sort of preference that a Singerian utilitarian needs to take into account. In chapter 2 it was established that non-sentient beings cannot be harmed and that the first principle of machine ethics is accordingly that a machine cannot be harmed if it is not sentient. But are there reasons to suppose that the harm of killing might be

[95] Singer, Peter. *Unsanctifying Human Life: Essays on Ethics*, edited by Helga Kuhse. Oxford : Blackwell, 2002. p.118

[96] Singer, *Rethinking Life and Death*, p.218

[97] Ibid, p.218

[98]Singer, *Unsanctifying Human Life*, pp.112-113

different and that a “soft” person could also be harmed by being killed? Singer claims that it is the fact that a being's preference to continue living is thwarted that makes killing that being wrong. But why should the preference of a person to stay alive be given any consideration by a preference utilitarian at all? After all, it is a preference whose dissatisfaction cannot cause any suffering to the being that holds the preference (aside from the moment of death perhaps- but this cannot be what makes death bad for persons, since sentient non-persons can also experience suffering when they die, and Singer holds that murdering a person painlessly is usually wrong also). Singer responds to this criticism by claiming: “the fact that the victims are not around after the act to lament the fact that their preferences have been disregarded is irrelevant. The wrong is done when the preference is thwarted.”[99] Yet, elsewhere in Singer’s work, he stresses the importance of a being’s reaction to having a preference thwarted, in deciding whether a being’s preferences are morally considerable. In *Practical Ethics*, Singer writes, "we saw in discussing the ethic of reverence for life that one way of establishing that an interest is morally significant is to ask what it is like for the entity affected to have that interest unsatisfied."[100], he goes on to say that a tree's interest in not having its roots flooded is not a morally significant interest because "there is *nothing* that corresponds to what it is like to be a tree dying because its roots have been flooded”[101]. Yet this defence of sentience as a necessary criterion for moral considerability undermines Singer's assertion that it is a person's preference to continue living that makes it wrong to kill them.

---

[99] Ibid, p.94

[100] Singer, *Practical Ethics*, p,283

[101] Ibid, p.277

After all, there is nothing that it is like for an entity that desires to go on living to have that interest unsatisfied. A dead person has just as much awareness of their preference to live having gone unsatisfied as a tree does of its roots being flooded. It is difficult for a utilitarian to make sense of the moral wrongness of killing a person without allowing that some significant harms cannot be experienced. If there are such harms, then there seems to be no reason why a being that has no capacity for experience could nevertheless be the subject of a harm of this type. A "soft" person could conceivably have a preference to continue existing, so what reason could be given for denying that the "soft" person's preference has any moral significance? It cannot be because there is nothing that it is like for it to have its preference thwarted, since the same is true of normal adult human persons. The experience of the harm of death is precisely the same for both "soft" and "hard" persons; in both cases there is no such experience. So if it is wrong to destroy a "hard" person with a preference to live, then why isn't it wrong to destroy a "soft" person with the same preference?

Although it is true that a "hard" person cannot experience the dissatisfaction of their desire to continue existing, a utilitarian might respond that it is possible for them to experience that preference being *satisfied*. Since a preference utilitarian thinks that it is a good thing for preferences to be satisfied, it is better to allow a person to continue to exist than to kill them. On the other hand, non-sentient beings cannot experience any of their preferences being satisfied; so their preferences, including their preference to continue existing, are not morally valuable. Only a "hard" person has the ability to conceive of their having a future, to desire to have a future, and to experience this desire being

satisfied. This is why, a utilitarian could claim, it is worse to kill a "hard" person than it is to kill any other category of being, including a "soft" person.

To say that it would be worse to kill a "hard" person than a "soft" person is not to say that it could not still be wrong to kill a "soft" person. There are other reasons (aside from those reasons that refer to direct harms to the person) that make the murder of a "hard" person wrong, and some of these are possibly also reasons for thinking that killing a "soft" person could also be wrong. Consider what Singer thinks about whether it would be wrong to kill a newborn baby; Singer writes:

> Killing a newborn baby is never equivalent to killing a person, that is, a being who wants to go on living. That doesn't mean that it is not almost always a terrible thing to do. It is, but that is because most infants are loved and cherished by their parents, and to kill an infant is usually to do a great wrong to its parents.[102]

And again:

> Although a normal newborn baby has no sense of the future, and therefore is not a person, that does not mean that it is all right to kill such a baby. It only means that the wrong done to the infant is not as great as the wrong that would be done to a person who was killed. But in our society there are many couples who would be very happy to love and care for that child. Hence even if the

[102] Singer, Peter. "FAQ", (n.d.) http://www.princeton.edu/~psinger/faq.html, accessed 22/02/2010

> parents do not want their own child, it would be wrong to kill it.[103]

Here, Singer argues for the wrongness of killing a being solely on the basis of *indirect* harms and benefits. Since the "soft" person has behavioural personhood, could some of the same indirect harms and benefits result from the killing of a "soft" person as from the killing of a "hard" person? Certainly, it seems plausible that the death of a "soft" person might cause grief to others, and that the killing of a "soft" person might disrupt the "peaceful coexistence on which society depends". Whether or not a "soft" person's murder could cause anxiety amongst sentient beings about the possibility of their own murder might depend on how much the androids resemble human persons, but it is not too difficult to imagine this type of harm also resulting from an android's murder.

One way in which we can make sense of the wrongness of killing without referring to directly experienced harms done to the being themselves is by acknowledging the distress that the death could cause to others, as Singer does above. Another way this might be done is by appealing to what might be termed the "aesthetic value" that bearers of behavioural personhood possess. Peter Harrison has made a similar argument regarding the treatment of non-human animals. According to Harrison, even if animals were not morally considerable because they were not sentient, animals would still have an aesthetic value that makes them the kind of thing that we ought to protect, rather than destroy, and he compares them to a work of art:

---

[103] Ibid.

> Briefly, it would be morally wrong to attack Michelangelo's 'Pieta' with a hammer, despite the fact that this beautifully crafted piece of marble cannot feel pain. If animals are mere machines, they are, for all that, intricate and beautiful machines (most of them), which like old buildings, trees and works of art, can greatly enrich our lives. Accordingly, rational arguments can be mounted against acts which would damage or destroy them.[104]

Harrison is surely correct in recognising that the end of an animal's life results in the loss of a subject with unique aesthetic value. How much more "intricate and beautiful" would a machine with behavioural personhood be? James Rachels argues that it is a person's possession of a "biographical" life that gives their life more value than that of a non-person[105]. Rachels claims death is bad for a person because it ends a life filled with goals, dreams, relationships, and all the things that we associate with the "biography" of a person. Moreover, it is worse to kill a person than a non-person because the added complexity of the life of a person adds value to their life; Rachels writes:

> A young woman dies: it is bad because she will not get to raise her children, finish her novel, learn French, improve her backhand or do what she wanted for Oxfam; her talents will remain undeveloped, her aspirations unfulfilled. Not nearly so much of this kind could be said about a less sophisticated being.

---

[104] Harrison, p.39

[105] Rachels, James. *The End of Life: Euthanasia and Morality*, Oxford, U.K.: Oxford University Press, 1986: p.50

> Her death is worse because there are more reasons for regretting it.[106]

The complexity of some machines could very well allow them to have what would be termed a biographical life; this is what makes them interesting characters in science fiction. According to William Ruddick, the notion of "biographical lives" allows us to make sense of harms that can come to a person without their awareness, including those harms that might occur after death. Ruddick claims that a biographical life can act as a "post-mortem surrogate" for a person, "a surrogate that can be harmed or benefited – and further extended – by what others do or say after the liver's death."[107] If it is biographical lives that we value particularly about persons, then (since there is no reason (in theory) why a non-conscious being couldn't have a biographical life), a "soft" person could have at least one property of personhood that makes it wrong to end a person's life and it might indeed be wrong to kill a "soft" person. This helps us to make sense of why the loss of a non-conscious person would seem like harm even if we would struggle to point to the being that was harmed. If it is bad for a person to die because it marks the end of a biography or personality, then it can be bad for a machine that has these things to "die" also. Killing a "soft" person is not as bad as killing a "hard" person, because it doesn't stop any morally important preferences being satisfied, but it may still be wrong to do it.

There are a number of other ways that the murder of a "soft" person could cause harm, which also fit the category of indirect harms. There is the need to be

---

[106] Ibid, p.57

[107] Ruddick, William. ""Biographical lives" revisited and extended." *Journal of Ethics* vol. 9 no. 3-4 (2005): p.512.

pragmatically "polite" enough to a socially-capable being that it behaves in a way that doesn't endanger you or others. If machines are in roles in which they can harm or benefit others, and are sensitive to social or moral slights, it would be prudent for human beings to provide machines with the means to meet their needs, and to treat them in whatever manner best provides for the smooth functioning of society. In this respect, the question of how to treat social machines is quite similar to that of how to treat existing machines like cars[108] and computers; in both types of cases we ought to be aware that machines have needs (e.g. for oil, power, and maintenance) and that neglecting those needs will lead to inefficiency and perhaps injuries or deaths[109]. In the case of social machines, their needs could possibly involve psychological needs, such as the need to be treated fairly, and with respect. If the machine is in a position to hinder or help humans (which many could perhaps be), then there is a strong pragmatic case for treating the machine with fairness and respect, even if there is no possibility that the machine could experience the suffering that would be caused to a human person in an equivalent position.

We may also wish to satisfy the preferences of "soft" persons to avoid social and psychological problems that could arise from people, especially young children, having violent or anti-social interactions with beings that are physically and behaviourally similar to human persons. If a person commits acts of violence against beings that cry out and exhibit pain behaviour, then there could be a possibility that they will repeat these acts against sentient beings. Kant used a

---

[108] Versenyi, Laszlo. "Can robots be moral?" *Ethics* 84 (3) (1974): p.449.

[109] Whitby, Blay. "Sometimes it's hard to be a robot: A call for action on the ethics of abusing artificial agents," *Interacting with Computers* 20, no. 3 (2008): p.327.

similar argument against the ill-treatment of animals, suggesting that there would be a "brutalisation" effect from the mistreatment of animals that might spill over into any subsequent human interactions a person had[110]. There are surely reasons to believe that this brutalisation effect, if it exists, would be stronger if a person were to perform violent acts upon a being that resembled a human person, and it is troubling to imagine what sort of effect it might have on a child for them to grow up accustomed to ignoring the needs of person-like agents. There could be social and educational benefit to be gained from teaching our children to be equally kind to "soft" persons as they are to "hard" persons.

There are a significant number of harms involved in the killing of a "soft" person that may be described as indirect harms, because they do not refer directly to the experience of the being in question. The killing of a "soft" person, while not a harm to the "soft" person themselves, and not as harmful as the killing of a "hard" person, nevertheless has the potential to cause significant harm. This is the fourth principle of utilitarian machine ethics:

> ***The Principle of Indirect Harm:*** *The value of satisfying or frustrating the apparent preferences of any machine will be* ***at least*** *as bad or good as any indirect harms or benefits that result.*

"M"

Using the principles developed thus far in this thesis, we can create an equation that allows us to see the expected harms or benefits that could result from an action that satisfies or frustrates a machine's apparent preference. Let "*I*"

[110] Kant, Immanuel, *The Metaphysics of Morals*, translated and edited by Mary Gregor, Cambridge: Cambridge University Press, 1996: 192-193.

signify the sum of any indirect harms or benefits of the action (any harms or benefits that are not directly experienced by the machine itself, but may be experienced by other beings that are sentient); let "*D*" signify the sum of any direct harms or benefits of the action (any harms or benefits that would be directly experienced by the machine itself, if it were sentient); and let "*x*" signify the degree of doubt that we hold about any claim that the machine is conscious. The equation "***M***" looks like this:

$$M = I+D\times (1-x)$$

Where:
M = the expected utility value from an action satisfying or frustrating the apparent preferences of a machine
I= Indirect harms or benefits
D= Direct harms or benefits
x= the degree of doubt that we hold about any claim that the machine is conscious

This equation shows that the expected harms or benefits that could result from an action that satisfies or frustrates a machine's apparent preference are least as good or bad as any indirect harms or benefits that result. It then allows for any direct harms or benefits that might result, but these are reduced by a percentage that depends on the level of doubt we hold about any claim that the machine is conscious.

Using this equation helps to resolve problems like the Spaceship Captain's Dilemma. To do this, we simply need to compare the value of "***M***" for the crew member Andrew in this scenario with the value of "***H***", an analogous equation for an action involving the preferences of Adam, the human crew member. The equation "***H***" looks like this:

$$H = I + D \times (1 - y)$$

Where:

H = the expected utility value from an action satisfying or frustrating the preferences of a human
I= Indirect harms or benefits
D= Direct harms or benefits
y= the degree of doubt that we hold about any claim that the human is conscious

We cannot know the exact value of either "*x"* in "***M***"or "*y"* in "***H***". Because of the Problem of Other Human Minds, the value of "*y"* in "***H***" will be greater than zero (there is always the chance that a human person is a natural zombie), but a rational person must admit that would not be much greater than zero; we have excellent reasons to believe that other humans are conscious. But, in chapter 3, reasons were given as to why there is less reason to suppose that a machine is conscious than to suppose that a fellow human is conscious; the Problem of Machine Minds was shown to be greater than the Problem of Other Human Minds. This means that, regardless of the value of "*y"* in "***H***", the value of "*x"* in "***M***" will always be greater. So, in situations where the values of all other variables are equal (as we might suppose they are in the Spaceship Captain's Dilemma), the value of "**M"** will represent less expected utility value than the value of "***H***". The same is true of a situation involving a choice between the interests of artificial *non*-persons or natural *non*-persons. The interest of an apparently sentient natural being in avoiding pain should outweigh the interest of an apparently sentient artificial being in avoiding pain. Since there is a greater level of doubt about the existence of the machine's mind than the animal's, then the likelihood of getting a positive utility value from an action showing preference to the machine is less than the likelihood of getting the same utility value from an action

showing preference to a natural animal. We have thus arrived at the final principle of utilitarian machine ethics:

> ***The Principle of Inequality:*** *The expected utility value of an action satisfying the preference of a machine is always less than the expected utility value of an action satisfying the equivalent preference of an animal (human or non-human).*

Here, the term "equivalent preference" signifies that the values of "*I*" and "*D*" are considered the more or less the same for "***M***" as they are for "***H***". Any decision that weighs up the preferences of machines against the equivalent preferences of animals should be weighted against machines. If a decision involves a choice between satisfying the equivalent preferences of equal numbers of machines and animals (human or non-human), then the choice should be made that satisfies the preferences of the animals.

In the Spaceship Captain's Dilemma, we can see that we must choose between the destruction of Adam (a "hard" person), or Andrew, who is a "fuzzy" person (either a "soft" person or a "hard" person). If Andrew is conscious, then he is a "hard" person, and he is of equal moral worth to Adam. If he is a "soft" person then his death will be at least as bad as any indirect harms that result, but no direct harms would affect Andrew as they would affect Adam. This means that to send Andrew to certain death could only be as bad as sending Adam, but we also know that it could not be worse. Although it is entirely possible that if we sent Andrew to his "death", then we would send a being with full moral personhood to be destroyed, and thus lose something of equal moral value to Adam, the extra

level of doubt surrounding Andrew's consciousness means there is more chance in Andrew's case that this will not happen. It is less of a risk to send Andrew on the mission, and this is what we must do.

Let us now consider two frequently cited problems in machine ethics that the five principles described in this thesis and the equation "***M***" might be applied to. First, let us consider the claim that it might be wrong to switch off a machine that is asking not to be. For example, James Geary speculates that in the future we will have robots that are "kind of high-tech pets" and that: "turning one off will be the moral equivalent of shooting your dog"[111]. While Rodney Brooks has said that he will feel he has completed his robot 'Cog', when people feel guilty about turning it off[112]. Turning off a machine is not really analogous to anything in human or animal terms, and certainly not murder. One of murder's most salient characteristics is its finality, but currently many electronic machines can be switched off or unplugged for long periods of time, then can usually be switched on again without the object having been damaged in any way (the closest parallel to something in human terms might be an anaesthetic). The same might very well be true of very complex social machines, even those that could be conscious. Perhaps we ought to add the proviso that a machine should not be shut down *permanently*. Or consider a thought experiment suggested by Roland Puccetti[113]: Suppose that Simon comes home to find his girlfriend Sally having sex with his best friend. In a fit of jealous rage, Simon beats and stabs Sally to death. It just so

---

[111] Geary, James. "The Brain-Machine Interface" in *This Will Change Everything: Ideas That Will Shape the Future*, edited by John Brockman. New York, N.Y.: Harper Perennial, 2010: p.16

[112] Brooks, Rodney, "Rodney Brooks Q & A", (n.d.), http://www.pbs.org/safarchive/3_ask/archive/qna/3275_rbrooks.html , accessed 06/09/10

[113] Puccetti, p.47.

happens that, unbeknownst to Simon, Sally is a robot. Should Simon be charged with murder? Has Simon done anything morally wrong?

Murder is one of the most significant harms that can be done to a person, so in this situation, the value of "*D"* (the sum of the direct harms or benefits that result from the action) in "***M***" will represent an appropriately large *dis*utility value. But since we don't know the value of the variable "*x"*, we can't tell whether the value of "*D*" will have much effect on the final value of "***M***". However, since the value of "*x"* has no bearing on the value of *I* (the sum of the direct harms or benefits that result from the action), we may still judge Simon's actions as wrong, whether Sally is conscious or not, based on the value of "*I*". The principle of necessary sentience and the principle of indirect harm tell us that if Sally is not sentient, then the harm involved in her murder can only be indirect harm, the type of harm that is caused to those that are not the subject of the action. The fact that Simon has lived with Sally as his girlfriend without realising that she was a robot, and that she perhaps developed another intimate relationship with his best friend, suggests that Simon has not only broken an object, but has also ended a biographical life. Sally is no longer able to fill the social role that she previously had, the kind of role that currently can only be filled by persons. The world has lost a personality, along with all the idiosyncrasies, plans, projects, creative output, and relationships that come along with one. The killing of Sally might also involve: grief caused to the friends and family of the person killed; threat to the "peaceful coexistence on which society depends"; Simon himself being "brutalised" by his actions; economic loss (from the loss of a valuable artefact and the loss of future profit that might have been created by Sally in her job); the loss of Sally as a receptacle of cultural information; the loss of an artefact that has

aesthetic value. All these are sufficient to say that Simon was wrong to do what he did, and probably also, depending on the scale of the harm, that he ought to be legally punished for it.

Another problem in machine ethics is whether or not it would be wrong to keep machines with person-like behaviour as slaves. This is the way in which machines are usually treated in the *Star Wars* movies, for example. Joanna Bryson has argued that slavery is "the correct metaphor we should use when thinking about our relationship with robot companions"[114], and thinks that it would be wrong to create machines "friend" that we owe obligations to, because the protection of these machines would involve the use of valuable resources that ought to be used for the benefit of humans[115]. Similarly, Mark Walker argues that if we could use robots to do some of the dangerous work that humans now have to do, then we might be morally *obliged* to make robot slaves, to avoid deaths and injuries to human persons[116]. However, Walker also claims that if robots were like us, then we would be guilty of substratism if we allowed them to be slaves. Walker is correct if saying that a robot is "like us" means that the machine is sentient, but if "like us" just means that the machine has "human-like" behaviour, then this does not necessarily mean the machine is sentient. The charge of substratism only applies when we have violated the principle of equal consideration, which in turn only applies to beings with the same morally-relevant properties. Walker admits that we wouldn't know whether or not a

[114] Bryson, J. J. "Robots should be slaves." In Close Engagements with Artificial Companions, edited by Yorick Wilks, Amsterdam: John Benjamins Publishing Company, 2010: p.64

[115] Ibid, p.67

[116] Walker, p.1

machine was sentient, but he claims that, to allow machines as slaves "we would need a high degree of certainty…that robots or computers are not conscious, for otherwise we risk mistreating persons in one of the worst ways."[117] The principles I have argued for lead to a different conclusion. The principle of inequality says that the expected utility value of an action satisfying the preference of a machine is always less than the expected utility value of an action satisfying the equivalent preference of an animal. The human worker's preference to avoid death and injury as a result of his work outweighs the machine's equivalent preference; this would remain true as long as the preferences of the machine were equivalent to those of the human. However, this would not justify the use of machines as slaves for trivial human needs, unless the value of "*x"* in "***M***" for the machine (representing the degree of doubt we hold about the possibility of the machine being conscious) was thought to be sufficiently high, as it is in the case of currently existing machines. On the other hand, it would mean that making robots to help in roles in which a loss of life was likely, would not just be acceptable, but morally obligatory.

[117] Walker, p.4

CHAPTER 5: CONCLUSION

"PICARD: Commander Data, what are you doing now?

DATA: I am taking part in a legal hearing to determine my rights and status. Am I a person or property?

PICARD: And what's at stake?

DATA: My right to choose. Perhaps my very life.

PICARD: My rights. My status. My right to choose. My life. It seems reasonably self aware to me. Commander? I'm waiting.

MADDOX: This is exceedingly difficult..." [118]

Star Trek: The Next Generation - *The Measure Of A Man*

Science fiction robots like Data provide intriguing examples of beings at the margins of our moral circle of concern. If such a being existed, would it be possible to harm it? How should we treat it? In this thesis, I have examined the question of whether a machine could be morally considerable from a utilitarian perspective, and in particular, from the perspective of the work of Peter Singer. In doing so, I have developed several principles of utilitarian machine ethics that provide a schema for where particularly advanced machines with human-like behaviour could fit into our worldview.

A utilitarian machine ethic holds that sentience is both necessary and sufficient for moral considerability, and utilitarians must take care to avoid substratism. Like racism, sexism, and speciesism, substratism describes an irrational moral prejudice (in this case, a prejudice against a being that is made

[118] Snodgrass, Melinda M. "The Measure of a Man", directed by Robert Scheerer. *Star Trek: The Next Generation*, Season 2, Episode 9. Paramount Television. First aired February 13, 1989

from a different type of material to oneself). The second chapter gave us our first two principles of machine ethics:

1) ***The Principle of Necessary Sentience:*** *If a machine is not sentient, then it cannot be harmed.*

2) ***The Principle of Equal Consideration of Interests:*** *If a machine is sentient, then its preferences should be considered as important as the equivalent preferences ("in so far as rough comparisons can be made") of other sentient beings.*

I have described sentience as consisting of conscious states that are qualitatively good or bad, and it must be stressed that consciousness should be understood as phenomenal states that have a feel; there must be something that is like to be having that state. Because of this, it is important in machine ethics to differentiate between thinking and consciousness.

In the third chapter, I addressed the question of whether or not a machine could have phenomenally-conscious states that are qualitatively good or bad. This is an exceedingly difficult question to answer, but because the methods we use to tell if another being is conscious are unreliable in the case of machines, then the proper attitude toward machine consciousness is agnosticism. I have argued that the Problem of Machine Minds is more problematic than the Problem of Other Human Minds or the Problem of Animal Minds, and this is our third principle of utilitarian machine ethics:

3) ***The Principle of Disanalogy:*** *The Problem of Machine Minds is greater than the Problem of Other Human Minds.*

In chapter 4, I introduced a distinction between a "hard" person, a "soft" person, and a "fuzzy" person. A "hard" person is described by the familiar definitions of person. A "soft" person has behavioural personhood, but is not conscious, and a machine like Data would be a "fuzzy" person (either a "hard" person or a "soft" person). A "soft" person would have indirect value, but could not be directly harmed. In this chapter, I introduced the fourth principle of utilitarian machine ethics:

4) ***The Principle of Indirect Harm:*** *The value of satisfying or frustrating the apparent preferences of any machine will be* ***at least*** *as bad or good as any indirect harms or benefits that result.*

The equation "***M***" provides a simple way of understanding our obligations towards machines. Any action satisfying or frustrating a machine's preference is at least as good or bad as any indirect harms or benefits that result. The possibility that we might cause direct harms to a machine is represented in the equation by the variable "*D*", but the value of "*D*" is appropriately discounted by any level of doubt that we hold about the machine being conscious. Because of the extra level of doubt about the consciousness of machines, the interests of a naturally-occurring being should be preferred over the interests of a machine with similar behavioural properties; the final principle of utilitarian machine ethics:

5) ***The Principle of Inequality:*** *The expected utility value of an action satisfying the preference of a machine is always less than the expected utility value of an action satisfying the equivalent preference of an animal (human or non-human).*

I have demonstrated that these principles may be applied to some significant problems in machine ethics, and it is hoped that they could provide guidance in many more such cases.

This thesis demonstrates the importance of attempting to find a solution to the Problem of Other Minds. If this significant problem can be overcome, then the variables "*x*" and "*y*" in the equations described in this thesis might be replaced with known quantities, and the moral status of machines would be all the more clear.

BIBLIOGRAPHY

Arp, Robert. ""If Droids Could Think..." Droids as Slaves and Persons," in *Star Wars and Philosophy*, edited by Kevin S. Decker, Jason T. Eberl, and William Irwin, Chicago: Open Court, 2005, 120-131.

ASPCR, "The American Society for the Prevention of Cruelty to Robots," (n.d.) http://www.aspcr.com/ , accessed 06/03/2010.

Barrett, J.L. "Exploring the natural foundations of religion," *Trends in Cognitive Sciences* 4, no. 1 (2000): 29-34.

Barrett, J.L. and Johnson, A. H. "The role of control in attributing intentional agency to inanimate objects," *Journal of Cognition and Culture* 3, no. 3 (2003): 208-217.

Block, Ned. "Troubles with Functionalism," in *Readings in Philosophy of Psychology, Volume I*, edited by Ned Block, Cambridge, Mass: Harvard University Press, 1983: 268-305.

Block, Ned. "Searle's Arguments against Cognitive Science," in *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, edited by John Preston and Mark Bishop, Oxford ; New York : Clarendon Press, 2002: 70-79.

Brooks, Rodney, "Rodney Brooks Q & A", (n.d.), http://www.pbs.org/safarchive/3_ask/archive/qna/3275_rbrooks.html, accessed 06/09/10

Bryson, J. J. "Robots should be slaves." In *Close Engagements with Artificial Companions*, edited by Yorick Wilks, Amsterdam: John Benjamins Publishing Company, 2010: 63–74

Bryson, Joanna and Kime, Phil. "Just Another Artifact: Ethics and the Empirical Experience of AI," presented at the *Fifteenth International Congress on Cybernetics*, Namur, 1998: 385–390

Carruthers, Peter. "Brute Experience." *The Journal of Philosophy* 86, no. 5 (1989): 258–269.

Carruthers, Peter. "Sympathy and Subjectivity." *Australasian Journal of Philosophy* 77, no. 4 (1999): 465-482.

Chalmers, David J. *The Conscious Mind: in Search of a Fundamental Theory*, New York : Oxford University Press, 1996.

Coeckelbergh, Mark. "Robot rights? Towards a social-relational justification of moral consideration," *Ethics and Information Technology* 12, no. 3 (6, 2010): 209-221

Cole, Phillip. "Problems with "Persons"" *Res Publica*, Vol.III, no.2 (1997)

Copeland, Jack. *Artificial Intelligence: A Philosophical Introduction*. Oxford, UK: Blackwell, 1993.

Dvorsky, George. "Sentient Developments: Must-know terms for the 21st Century intellectual: Redux", January 11, 2007. http://www.sentientdevelopments.com/2007/01/must-know-terms-for-21st-century_11.html, accessed 03/06/2010.

Garreau, Joel. "Bots on the Ground." *The Washington Post*, May 6, 2007 http://www.washingtonpost.com/wp-dyn/content/article/2007/05/05/AR2007050501009.html , accessed 21/11/2009.

Gates, Bill. "A Robot in every Home." *Scientific American* 296, no. 1 (January 2007): 58-65.

Geary, James. "The Brain-Machine Interface" in *This Will Change Everything: Ideas That Will Shape the Future*, edited by John Brockman. New York, N.Y.: Harper Perennial, 2010: 14-16

Halpern, M. "Turing's test and the ideology of artificial intelligence," *Artificial Intelligence Review* 1, no. 2 (1987): 79–93.

Hanson Robotics, "Our Mission is to Realize Friendly Super-Intelligent Machines" (n.d) http://www.hansonrobotics.com/vision.html , accessed 11/02/10

Harnad, S. "The Turing Test is not a trick: Turing indistinguishability is a scientific criterion," *ACM SIGART Bulletin* 3, no. 4 (1992): 9–10.

Harrison, P. “Do animals feel pain?,” *Philosophy* 66, no. 255 (2009): 25-40.

Humphrey, Nicholas. *Seeing Red: A Study in Consciousness*, Cambridge, Mass.: Harvard University Press, 2006

Hyman, Ray. “The Psychology of Deception,” *Annual Review of Psychology,* 40 (February 1989): 133-154

Kahn, P.H., Freier, N.G., Friedman, B., Severson, R.L. and Feldman, E.N., “Social and moral relationships with robotic others,” in *Proceedings of the 13th International Workshop on Robot and Human Interactive Communication (RO-MAN’04)*, 2004: 545-550

Kant, Immanuel, *The Metaphysics of Morals*, translated and edited by Mary Gregor, Cambridge: Cambridge University Press, 1996.

Leiber, Justin, *Can Animals and Machines be Persons?: a Dialogue,* Indianapolis, Ind.: Hackett Publishing, 1985

Locke, John. *An Essay Concerning Human Understanding* Oxford : Clarendon Press, 1975.

Macphail, E. M. *The Evolution of Consciousness*. Oxford: Oxford University Press, 1998.

Microsoft. “Microsoft’s Tradition of Innovation: From Revolution to Evolution”, October 25 2002,

http://www.microsoft.com/About/CompanyInformation/ourbusinesses/profile.mspx , accessed 20/02/2011

Nagel, T. “What is it like to be a bat?” *The Philosophical Review* 83, no. 4 (1974): 435–450.

Oxford Reference Online. "Machine *Noun*" in *Oxford Dictionary of English*, edited by Angus Stevenson. Oxford University Press, 2010. http://www.oxfordreference.com/views/ENTRY.html?subview=Main&entry=t140.e0489390, accessed 7/02/2011.

Penrose, Roger. *The Emperor's New Mind : Concerning Computers, Minds, and the Laws of Physics*. Oxford ; New York: Oxford University Press, 1989.

Price, Keith “Why Computers Will Never Be People,” in *Selected Papers from Conference on Computers and Philosophy*, Volume 37, 2003: 45-49

Putnam, Hilary. “Robots: Machines or artificially created life?” *The Journal of Philosophy* 61, no. 21 (1964): 668–691.

Puccetti, Roland. *Persons: A Study of Possible Moral Agents in the Universe*, London, Melbourne: Macmillan, 1968.

Rachels, James. *The End of Life: Euthanasia and Morality*, Oxford, U.K.: Oxford University Press, 1986.

Ruddick, William. “"Biographical lives" revisited and extended.” *Journal of Ethics* vol. 9 no. 3-4 (2005): 501-515

Russell, Bertrand. *Human Knowledge: Its Scope and Limits,* London: Allen and Unwin, 1948.

Searle, John R. "Minds, brains, and programs," in *Behavioral and Brain Sciences* 3, no. 3 (September 1980): 417-457.

Searle, John R. *The Mystery of Consciousness*, New York : New York Review of Books, 1997.

Searle, John R. "Mental Causation, Conscious and Unconscious: A Reply to Anthonie Meijers." *International Journal of Philosophical Studies* 8, no. 2 (7, 2000): 171-177.

Searle, John R. "Twenty One Years In The Chinese Room." In *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, edited by John Preston and Mark Bishop: 51-69. Oxford ; New York : Clarendon Press, 2002.

Singer, Peter. *Animal Liberation*. 2nd ed. New York, N.Y: New York Review of Books, 1990.

Singer, Peter. *Practical Ethics*. 2nd ed. Cambridge: Cambridge University Press, 1993.

Singer, Peter. *Rethinking Life & Death: The Collapse of Our Traditional Ethics.* Melbourne : Text Publishing, 1994.

Singer, Peter. *Unsanctifying Human Life: Essays on Ethics*, edited by Helga Kuhse. Oxford : Blackwell, 2002.

Singer, Peter. "FAQ", (n.d.) http://www.princeton.edu/~psinger/faq.html, accessed 22/02/2010

Singer, Peter, and Agata Sagan. "When robots have feelings," December 14, 2009. http://www.guardian.co.uk/commentisfree/2009/dec/14/rage-against-machines-robots , accessed 06/03/2010

Singer, Peter, and Agata Sagan. "No Rights for Robots? Never?" *Free Inquiry*, June/July 2010: 13, 39

Snodgrass, Melinda M. "The Measure of a Man", directed by Robert Scheerer. *Star Trek: The Next Generation*, Season 2, Episode 9. Paramount Television. First aired February 13, 1989

Thompson, Janna. "A refutation of environmental ethics." *Environmental Ethics* 12 (2) (1990):147-160

Tooley, Michael. *Abortion and Infanticide*. Oxford: Clarendon Press, 1983.

Torrance, Steve. "Ethics and consciousness in artificial agents." *AI & SOCIETY* 22, no. 4 (3, 2007): 495-521.

Turing, A. M. "Computing Machinery and Intelligence," *Mind* 59, no. 236, New Series (October 1950): 433-460.

Versenyi, Laszlo. “Can robots be moral?” *Ethics* 84 (3) (1974):248-259.

Virtual Worldlets Network, “VWN Virtual Dictionary: Substrate Chauvinism,” (n.d.), http://www.virtualworldlets.net/Resources/Dictionary.php?Term=Substrate%20Chauvinism&Letter=S , accessed 21/09/2010

Walker, Mark. “A Moral Paradox in the Creation of Artificial Intelligence: Mary Poppins 3000s of the World Unite!” in *Human Implications of Human-Robot Interaction: Papers from the AAAI Workshop*, edited by Ted Metzler, California: AAAI Press, 2006: 23-28.

Whitby, Blay. “Sometimes it's hard to be a robot: A call for action on the ethics of abusing artificial agents,” *Interacting with Computers* 20, no. 3 (2008): 326–333

Winograd, Terry. “Understanding, Orientation, and Objectivity.” In *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, edited by John Preston and Mark Bishop, Oxford ; New York : Clarendon Press, 2002: 80-94.