The Ethical Problems Associated With The Creation of a Synthetic Consciousness

Gerard J. Brady

Acknowledgements

To Dr. Susan Stuart, whose support, advice, encouragement and infinite patience made this thesis possible. I offer my deepest and warmest thanks.

Dr. Stephen Bostock, who provided invaluable and entertaining insights into the consciousness and behaviour of animals.

Julie, who offered a unique perspective on various aspects of this thesis.

Kathleen Johnston and the people at the Faculty Office who helped me on more than one occasion with several things pertaining to postgraduate work.

Elaine Wilson, without whom nobody would have been acknowledged!

Finally, to my parents, two people to which I owe everything. If I spent the rest of my life trying, I could not hope to repay a fraction of what you have given me. Thank you.

Contents

1.Introduction

> *'...your scientists were so preoccupied with whether or not they could; they didn't stop to think if they should.'*
>
> (Jurassic Park, 1993)

In *Jurassic Park* a team of scientists are attempting to recreate the Jurassic period in order to open a theme park; a sort of Disneyland with dinosaurs. The quotation expresses the concerns of Dr. Ian Malcolm (played by Jeff Goldblum) about the apparent disregard the scientists have for the potentially dangerous consequences of meddling with nature. Of course, the scientists pay no heed to these concerns and, being Hollywood, many terrible consequences - including death, disaster and destruction - do occur. In spite of the glossy, overacted, special-effect laden nature of its origin, the quote does make a significant point with respect to science and, in our particular case, with respect to the creation of synthetic consciousness.

There has been a vast amount written on the nature of consciousness and how it might be created, but little has been said about the implications, both moral and practical, of success. Certainly, literature discussing whether our attempts to create a non-human consciousness can be morally justified is very thin on the ground. And even when morality is discussed with regard to machine consciousness, it often focuses on the steps that can be taken to protect us from the machines and very rarely do we see concern for the opposite scenario: the protection of our conscious creations from us.

In this thesis I intend to look at the moral questions that arise from synthetically creating a consciousness which possesses many or all of the capacities associated with human consciousness, and to argue that we are not morally justified in attempting to bring such a creation into existence. In the first section, I will argue that that there are no good reasons, moral or practical, for the creation of a synthetic consciousness and that the reasons given so far, other than merely to see if we can, are invariably bad ones.

I will then go on to explain the importance of consciousness in terms of where we draw moral boundaries with regard to the treatment of any artificially created consciousness, this explanation will involve an outright rejection of the epiphenomenalist position. Already we have everyday examples of the drawing of such boundaries, which appear to be based on the level of consciousness attributed to the organisms in question. An example of this may be found in any local supermarket that sells tins of tuna, on which can be found labels that proudly promote the fact that the tuna is 'dolphin friendly'; a fact most tuna would not find particularly reassuring. What does seem true however, is that there are people who opt to buy the 'dolphin friendly' variety and who feel that they have acted in a more 'morally correct' manner than those of us who do not make such a distinction and opt to buy whatever tuna is nearest at hand or the least expensive. This, I contend, is because dolphins are seen to possess characteristics and traits which are perceived to be more human than those possessed by the tuna. Such characteristics may include a highly developed intelligence, the expression of altruistic behaviour, and sociability. Further, we cannot overlook the fact that dolphins also appear to have a wide and open smile which may make it easier for humans to feel an affinity with them since this smile suggests a happy, playful nature.

Clearly then, we do make distinctions in our moral treatment of animals depending on, possibly among other things, the perceived similarity of their consciousness to that of humans and it is almost certain, I will argue, that we would do likewise with any synthetically created consciousness. I will argue that a major factor in this distinction comes from an unavoidable anthropocentric standpoint, a veil which we find very difficult to move beyond, and which plays a significant role in our moral judgements regarding non-human animals - and would do likewise with any man-made, conscious being. In addition to this, I intend to examine the moral relevance of pain and suffering as well as the role of desires, beliefs and emotions in the manifestation of these states, and how much can be said for a non-human organism placing any kind of significance on the continuation of its own life**.**

I will examine both the necessity, and the difficulties, of implementing any sort of moral code in machines; difficulties which arise both from a human standpoint and from the perspective of any potentially created consciousness. I will argue that ethics, by its very nature, does not lend itself easily to mathematics or formulaic expressions, and that until we uncover an infallible moral code of our own, any attempt to impart our moral 'wisdom' to machines would be morally naïve.

Finally, I will consider what the implications might be for our ethical treatment of any synthetically created consciousness and will argue that they, like human and non-human animals, should never be subject to unnecessary pain or suffering; this will include a rejection of Petersen's claim that Engineered Robot Servitude is morally permissible

2. **The Justification in Creating Machine Consciousness**

(i) Moral Agents and Moral Patients

Before going on to examine the moral implications of creating a synthetic consciousness, it is important to note that I argue from an externalist point of view. I do not wish to delve into the debate surrounding the necessary criteria for the creation of consciousness, but merely wish to assert that I take consciousness to require both embodiment and some manner of interaction with an external world. To be conscious, we must be conscious of something; the possibility of a consciousness existing independently of any affective stimulation – a consciousness not conscious of anything, not even sentient - would be analogous to the existence of a non-burning fire: that is to say, it would not exist at all. The idea of a distinct external world being necessary for the creation and/or existence of consciousness has fairly significant implications for the creation of moral agents: namely, that the creation of moral agency necessarily implies the creation of a moral patient, that is, a being with the capacity to experience pain and suffering and**,** thus, worthy of moral consideration. I will argue that, in order for an actor to be considered a moral agent, it is necessary for it to have the capacity for experiencing, understanding, and sympathising with the consequences (feelings, emotions, physical pain and so on) it *intentionally creates in others.* The fact that '*others*' are required necessitates the existence of an external world. Effectively, I can only be genuinely considered a moral agent if I am, in some way, able to experience for myself the effects my actions have on others, that I freely and intentionally create these experiences in others, and that I am aware of the fact I am doing so.

For moral agency, a being must possess the capacity for acting rationally upon a set of moral principles as well as the freedom to act upon these principles; it would be difficult, possibly even absurd, to hold morally culpable an actor who is entirely subject to deterministic laws which make an alternative course of action impossible. An actor of this kind may be a moral patient, able to experience pain or to suffer[1], but by lacking the capacities for reason and autonomy, it cannot be held morally accountable for its actions. Further, the level of moral agency, and so moral accountability, we attribute to a being is proportional to their ability to recognise others' pain and suffering, their perceived level of autonomy and the extent to which they *intend* to bring about the consequences that they do. The making of a hurtful comment, for example, made by a rational and self-aware adult intending to offend would be judged very differently from the same comment made by a very young child or a person with learning difficulties. One reason for this is that we believe the child or the mentally challenged adult to have less awareness of the potential effects of their actions on others, hence any harm or offence caused would not be taken as deliberate or intentional; they would lack the capacity to understand and internalise the negative feelings that those affected by their words or actions might feel, and so cannot anticipate the harm they may cause. Also, if a person has a disability, such as autism, which diminishes their autonomy, then we generally do not hold them accountable for the effects that their words, actions and behaviour produce. A case in point is that of Gary McKinnon, an Aspergers victim who, in 2008, hacked into the Pentagon looking for UFOs' and was later extradited to the US to stand trial for cyber-terrorism[2]. Although there were those who thought that the extradition was entirely just, there were many who believed that, as a result of his condition, McKinnon should not be held entirely for what's happened.

Intentionality then, is an important issue in all morally-appraisable circumstances; an intentional moral agent has an awareness and understanding of the potential consequences brought about by his or her actions and may *deliberately* set out to make these potential consequences a reality. An agent who merely behaves morally and who lacks any conscious intentionality is what Wallach and Allen (2009) describe as an Artificial Moral Agent (AMA), an agent who is programmed to behave in a morally correct manner but does not necessarily require any conscious awareness of the reasons for, or the results of, its actions. Admittedly, the former would be far harder to manufacture, for reasons I will examine later, and its presence would be far harder to prove[3], but only through a conscious understanding of the effects of its actions, as well as a deliberate, non-coerced attempt to bring these consequences about, can a being be defined as a moral agent. Such intentionality not only requires the existence of some external entity at which intentions can be directed, but also that the agents have the capacity for internalising, feeling, reflecting on, and understanding the negative effects brought about by their actions; this demands that *they themselves* are capable of imagining the possible outcomes with some associated affective experience. In short, the creating of a moral agent necessarily implies the creation of a moral patient.

This however, does not suggest that the reverse is true, i.e. that all moral patients are necessarily moral agents. The creation of a moral patient only requires the ability to experience pain or suffering, as well as a desire to avoid these experiences, whereas for moral agency, there must be the additional conscious capacities for rationality, intentionality and

[1] As a consequence of a moral patient's capacity for experiencing pain, moral agents have an obligation to ensure that no unnecessary pain or suffering is inflicted upon them.

[2] See http://www.nytimes.com/2009/08/01/world/europe/01britain.html

[3] Here we have something similar to the 'other minds problem'; how do we ever ascertain whether or not something is intentional, even when dealing with other humans, except through their own admisssion?

autonomy as well as a degree of empathy (even if such empathy feeds the desire to act immorally). We do not have to look too far for instances of moral patients who are not simultaneously conceived as moral agents. Family pets, for example, are seen as moral patients insofar as they can feel pain and suffer, but are not generally taken to be moral agents in virtue of the fact that they do not have the mental capacity to refrain from acting upon their natural impulses. Nor do they have the depth of understanding required to intentionally make their owners feel sad or upset, or to deliberately injure their feelings.

It is important to note that deservedness is not a relevant feature in the ascertaining of whether or not a being is a moral patient, and there is no moral contradiction in defining even the most wicked and sinister being as a moral patient. Such an agent may well act in a malicious way *because* it will cause harm or injury to others, but to be considered a moral agent still demands that the agent possesses awareness and understanding of the negative experiences their action causes in their victims. The fact that such actors are able to understand, experience and empathise with the negative or unpleasant experiences that they inflict on others is enough for them to be considered as moral agents, there is no demand for them to be *good* moral agents.

(ii) Synthetic Consciousness as Unjustifiable

Before looking at the moral aspect of manufacturing conscious machines, it is important to examine our reasons for creating such a being. Human history is teeming with undeniably incredible achievements, ranging from the invention of the wheel through to James Watt's steam engine, space travel, and to the present day innovations which have given us the internet, satellite navigation systems and mobile phones. However, not all of our wondrous technological advancements have arrived with beneficial intentions; indeed, some seem to have little purpose at all. Some would argue that the billions of dollars spent propelling Neil Armstrong and Buzz Aldrin to the moon forty years ago seems to have had little or no impact on our quest to improve and advance as a species, and that its main objective was an all too human tendency towards ostentation or, more succinctly, simply to show that we could. There have also been creations which have had more sinister purposes in mind. Such malevolent creations include gunpowder, the Atom Bomb, nuclear weapons and the mind-boggling equipment employed in espionage and surveillance; each one dangerously powerful in the wrong - or even the right - hands. Similarly, creating a conscious machine which remained stationary and inert in a laboratory would be one thing (which might constitute harm in its own way, but more on this later); but to manufacture active, powerful and conscious beings with the potential for malevolent objectives would be quite another.

Already I have suggested that the creation of consciousness for the sake of creating it does not provide us with particularly convincing justification, however, it might be true that consciousness could improve the performance of any machine into which it was placed, including robots designed to reduce the burden of our daily chores or reduce the risks involved in the undertaking of tasks we find perilous. If making life easier for humans is seen as a benefit, then robots designed to do housework, car mechanics or grocery shopping would provide a significant advantage, and the benefit of consciousness would lie in the fact that it would mean that the robot could learn and adapt to its surroundings more quickly and efficiently, thus settling into a pattern which fits its owner's lifestyle within a shorter timeframe. The moral implications of robot servitude will be examined later, but for now it is sufficient to say that making every aspect of human existence tranquil and undemanding does not necessarily provide sound justification for bringing something into existence.

Take the internet and the advances made in telecommunications. There certainly have been a number of positive things that have arisen since the inception of the worldwide web: the housebound have an outlet for communicating with the outside world, it provides us with the ability to communicate with people of different cultures from around the globe, it allows us access to unprecedented volumes of information on virtually any subject, and makes it easy for us to remain in touch with friends and loved ones who now live at a distance from us. However, this revolutionary invention is not without its drawbacks. Obesity levels are higher than ever[4] due to people spending endless hours transfixed by a computer screen, social and interpersonal skills are diminishing as we become ever more reliant on virtual 'chat' rooms and it might be argued that our language is being desecrated as it gets further and further abbreviated to fit the contemporary instant messaging culture. Similarly, the creation of conscious machines which are designed to assist us in our daily routines would also come with their own particular sets of pros and cons; on the one hand they could be of great assistance to the elderly and infirm but, on the other, they could make us, as a society, even more indolent and slothful, thus exacerbating the health problems we currently have. I am sure, with enough time and thought, a long list of the advantages and disadvantages associated with burden-reducing machines could be compiled, but providing a comprehensive list is unnecessary to make the point that it isn't all beneficial.

A further possible reason for the creation of conscious, active machines concerns their fungibility and the advantages this brings with respect to their deployment in dangerous tasks or missions. In 2005, 118 in every 100,000 fishermen died in America alone, making it the most dangerous job in that particular country, finishing just ahead of the logging industry where there were 90 fatalities for every 100,000 workers[5]. Arguably, an even more perilous occupation is that of a soldier in times of conflict, with the war in Iraq alone claiming the lives of over 4000 American soldiers between 19th March 2003 and 14th February 2009[6] and other conflicts such as the two world wars and the war in Vietnam would surely provide an even lower ratio of combatants to fatalities. Here we seem to have ideal circumstances in which to replace humans with conscious machines which, in the event of damage or destruction, can either be repaired or replaced with others of their kind. Without question, the idea of reducing the risk to human life in these situations has a strong appeal, as does the idea of manufacturing beings capable of calculating these risks far more efficiently than we can, thus becoming less likely to overreach, stand in the wrong place, shoot an innocent bystander, or make any decision that goes catastrophically wrong. Indeed, there would likely have been far less moral outcry and protestation against the wars in Iraq and Vietnam if they had been fought between two sets of robots, with no harm to any civilians; harm which is euphemistically defined as 'collateral damage'. However, there is a considerable difficulty in reconciling consciousness with fungibility, which involves the likelihood of conscious machines being in possession of memories of a subjectively-lived life.

Presently, we have several types of machines capable of performing certain tasks far more efficiently than human beings. These range from basic pocket calculators, which can perform arithmetical tasks quickly and accurately, to more advanced robots found in places such as

[4] For more on this see the World Health Organisation's website: www.who.int/dietphysicalactivity/publications/facts/obesity/en/
[5] Figures taken from http://money.cnn.com, accessed November 2008
[6] Figures taken from http://www.antiwar.com accessed November 2008

car manufacturing plants or those sent to explore the surface of Mars. One thing currently missing from even the most advanced of these machines however, is consciousness. Although there is considerable debate surrounding the qualities and capacities necessary for a being to be considered conscious, it would come as little surprise for us to learn that in order for any being to function most effectively in human-centred tasks (e.g. war, housework, fishing, logging, piloting etc.), it would require a consciousness of the type and level which was on a par (at least) with that of humans, otherwise evolution would have equipped us with a consciousness other than the one that we have. However, as soon as such a consciousness is created, the potential exists for it to think, learn, feel emotions, form beliefs and desires and, crucially, form memories. These in turn could allow it to form opinions, sympathise and empathise with others and to act in a manner appropriate to the circumstances in which it finds itself, which may also include the giving of meticulous consideration to moral dilemmas. A conscious machine will have its own experiences of the world, even if they are only minutely different to those of its 'clones', and will almost certainly have personal memories of these experiences. It might even have formed a non-genetic disposition for survival, which may provide the motivation for flight instead of fight. Essentially, if a being has a unique set of memories of a past history, of which it is consciously aware, then it is difficult to see how it can be in any way replaceable; its fungibility is compromised. Consciousness involves having our own particular, subjective experiences of the world and there appears no way to reconcile the subjectivity with fungibility. Robots would either be conscious or they would not. If they are, then along with this comes the potential for forming memories within a unique personal history, in which case they cannot be considered fungible and could never be treated as such morally. If they are not conscious, then we have merely improved on what we already have: powerful, efficient, emotionless collections of wires, nuts and bolts which can be programmed to behave in a particular way, but are far less reliable in unusual or novel situations, where 'breaking the rules' might be the ethically correct course of action. There is no way then, to justify the creation of conscious, thinking and feeling machines through appeal to their fungibility for it would be the loss of a unique existence.

A further attempt to justify creating conscious machines, which does not rely on the 'just to see if we can' principle, is that it provides us with the potential for learning more about our own consciousness and about our own morality. Since computers are so well suited to testing the results of consistently following patterns or set of rules, they can be used to quickly and accurately evaluate the effectiveness of any particular moral code (Anderson & Anderson, 2007). Also, the processes involved in creating conscious machines could help us understand more about the nature of our own consciousness. However, although it may be true to say that machines provide us with the ideal testing arena for ascertaining the most appropriate moral code to which to adhere, there is surely a question mark over any attempt to impart moral and ethical wisdom (through programming or any other means) to other beings in the absence of an infallible moral code of our own. Sending a 'test' consciousness out into the world, abiding by an erroneous moral code, could have disastrous consequences, particularly if the consciousness were placed in a physically powerful vehicle. Anderson and Anderson attempt to navigate around this problem by suggesting that we should see the moral code as updateable, and that until we uncover an infallible version of the code, machines should be kept away from situations which demand answers to difficult moral dilemmas. But there seems little point in creating a machine that only has the capacity to tell us things that we already know and gets as perplexed as we do when faced with more testing ethical conundrums. We might object that we already pass on our values and moral convictions to our children, even though we do not as yet have an infallible set of moral rules; but children, as they get older, have the potential to question these values and to form ethical codes of their own; it is not obvious that a morally pre-programmed machine could do likewise, particularly if it were created as an 'adult', a point which I will examine in more detail later.

As far as gaining a better understanding of our own consciousness via the manufacturing of another, I have two points to make: (1) it would surely be easier to manufacture a synthetic consciousness if we already possessed an understanding of our own, in the same way that a greater understanding of the working of the limbs has allowed the development of prosthetic arms and legs and led us away, thankfully, from the days of wooden legs. (2) even if we did manage to successfully create a synthetic consciousness, this does not guarantee that we would be any the wiser as to the inner workings of consciousness in humans; manufacturing a contraption which allows a car to get from A to B in no way implies that I would gain a better understanding of the inner workings of a conventional car engine. [See section 3(ii) for more on this].

Essentially then, before we put copious amounts of time, money and effort into the creation of a non-human consciousness, we have a moral duty to examine our motives for doing so. Creation for creation's sake does not provide us with enough justification for ploughing what could be billions of pounds and a vast amount of resources into a project which has no apparent benefit to our development. Endeavouring to construct things which strip away the effort and exertion from our everyday living does not necessarily constitute such justification, since this can result in taking away the things that make us distinctly human - such as complex social interactions, relishing the challenge of solving difficult problems, our apparently boundless creativity and the ability to develop and utilise complex and intricate language - and reduce us to obese, anti-social sloths, unable to communicate without the comfort blanket of a keyboard and a dizzying array of abbreviations: the linguistic equivalent of grunting. We cannot rely on conscious machines taking the place of humans in perilous situations, at least not with any genuine moral conviction, since if a being is conscious then it has the potential to form its own subjective and unique personal history, meaning that we have a very hard time in defining it as fungible. Consciousness would also endow these beings with the capacity for experiencing pains, fears and concerns as well as a desire for survival equal in intensity to that of the humans they replace, and there is no justification for subjecting one consciousness to these disagreeable experiences in order to protect another. In any case, how would the fishermen, loggers, pilots, soldiers and those with equally hazardous occupations manage to earn a living if they were replaced with machines? We must also tread carefully with the development of conscious machines as a tool for the attainment of a better understanding of our own consciousness and of our own moral world. In the first instance, there is no guarantee that creating conscious machines would provide us with such an understanding and, in the second, perhaps we should aim to perfect our own understanding of our moral universe before we arrogantly attempt to impart our fragile moral convictions on other, powerful, beings who could end up beyond our control.

3.Why Consciousness Matters

In this section I will examine the question of whether consciousness is important in both a moral and an evolutionary sense. From an evolutionary standpoint, Darwinian (1859/1988) thought would dictate that, in order for consciousness to have persisted through the ages – and for it to have existed in the first place - it must have endowed and continue to endow conscious beings with an advantage which maximises their chances of survival and increases the possibility of outliving their non-conscious competitors. From a Darwinian standpoint, if consciousness did not enhance the prospect of survival for its bearer, then natural selection dictates that it would have been discarded, though there is, of course, the possibility that consciousness could have survived as something which neither harmed nor benefitted its bearer, rather than being selected. Bostock (2008) cites Stephen Jay Gould's term 'spandrel'[7], to describe a features '*which hadn't actually been selected for, but were by-products of the construction of the sort of animal they were part of*''. However, consciousness seems to play such a major part in our lives that it seems unlikely that it survived merely as a by-product or spandrel [See section 3 (iii) for more on this]. One problem then, arises from the attempt to identify the functional role consciousness plays in the enhancement of these prospects; more succinctly, why is it better to have consciousness than not? Further, in order for conscious mental activity to boost an organism's chances of survival and so win out in the battlefield of natural selection, it must have been able to influence the physical aspects of its organic vehicle in some meaningful way. By identifying this functional and beneficial role of consciousness in humans, the task then becomes to identify any similar benefits to non-humans. Effectively, if conscious, mental activity can be shown to have been critical to human evolution, then perhaps some of its advantages and uses can be applied to the evolution of non-human organisms. There is, on the other hand, the possibility of abandoning all talk of the evolution of consciousness, denying the admission of the mental into the world of cause and effect, and going down the path of *Epiphenomenalism***:** we shall see if it comes to that**.**

From a moral perspective, it is important to look at how consciousness in other organisms or bodies would affect the ethical attitudes of people towards them and, crucially, at what point we would permit other non-human agents into our sphere of moral consideration. For some, it may simply be a matter of aesthetics and that the harming of a dog, for example, is a matter of moral concern because dogs are nice to look at and pleasant to have around. For others, the ability to feel physical pain may be enough to deem an organism worthy of our moral concern; yet even this may not be sufficient for those who demand that the organism be able to reflect on this pain, being consciously aware not only of the pain, but also of its duration and the anticipation of its cessation. And there are those who argue for the exclusion from our moral realm of any being which cannot be shown to possess the level and type of consciousness associated with humans. There will always be those who stipulate that the price of entry into our moral dominion is that an organism can walk, talk, think and act like a human, as well as being capable (and eligible) to apply for a passport.

I will argue that consciousness plays a fundamental role in the formation of our ethical and moral outlook, that having the capacity for consciousness makes it possible for the organism to form even basic action-related, non-linguistic beliefs and experience desires and emotions, and that an organism capable of this is one with a non-trivial experiential life which is open to injury and harm.

(i) Epiphenomenalism and Mental Causation

Huxley (1874) likened consciousness to a steam whistle from a locomotive in that, mental events, despite being caused by physical events, are equally inert with regard to causality. In fact, according to epiphenomenalism, mental events are powerless to cause even other mental events. Consequently, each and every one of our actions come about as a result of purely

[7] An architectural term used to describe the space between two arches, which artists used to paint pictures of saints in St. Mark's Cathedral in Venice.

physical causes, while our wants, needs, desires and our volition to act are themselves caused by physical causes and are themselves causally inert. This commits those who subscribe to epiphenomenalism to the anti-Darwinian idea that mindless bodies could, and probably would, have evolved to perform at least as well as those containing fully-conscious minds. Despite this strongly counterintuitive characteristic of epiphenomenalism, there are some intriguing aspects to it.

In the seventeenth, eighteenth and nineteenth centuries, many philosophers (see Descartes 1641, Locke 1690**,** and Hume 1739) believed in some form of dualism, where mental activity was not reducible to physical states and processes. As science (and in particular neurophysiology) progressed however, it uncovered no causal link between our mental activity and the brain or body; humans, like all animals and inanimate objects, began to be seen as part of the physical world and as such, subject to the laws of physics and natural mechanics. But the laws of physics leave no room for non-physical causes and so, the non-physical nature of our mental events necessarily renders them causally inert. Where apparently non-physical, subjectively experienced, mental events manifest themselves in a world governed entirely by physical forces is anyone's guess and without psycho-physical laws [Fechner 1860/1912], the appeal of epiphenomenalism becomes apparent.

The first difficulty in placing mental states in a world of purely physical events comes from the 'Anomaly of the Mental', which states that the link between causes and effects must adhere to strict physical laws and mental events, if humans are to be considered truly autonomous, cannot be subject to deterministic laws of physics. Consequently, there can be no such thing as mental causation. In 'Mental Events', Donald Davidson (1970) offers a possible solution to this problem by way of a variety of property dualism, namely, anomalous monism. Davidson argues that mental and physical events can, and do, causally interact and, although events related by cause and effect are governed by strict laws, mental events are not subjected to such laws and so cannot be explained or predicted by them. Davidson resolves this apparent tension between the physical and the mental by arguing that mental events as *types* are anomalous and that causal relations only exist between physical and mental event *tokens*. Such token physicalism allows that the mental be supervenient on, without being entirely reducible to, the physical.

Support for the idea of epiphenomenalism comes in the form of Jaegwon Kim's (2008) Causal Exclusion argument, an attack on non-reductive physicalism, which Kim believes inevitably leads to epiphenomenalism. Kim argues that if mental states are supervenient on physical properties, as non-reductive physicalists believe that they are, then they must be causally irrelevant.

Secondly, at least some mental contents seem to be dependent upon certain, external conditions in the outside world; they are the relational properties expressed in the content of object X for agent Y, while causality seems to be a local phenomenon, with any particular system's behaviour depending upon its internal structure or make-up. Subsequently, any two systems with an identical internal make-up will behave in exactly the same way, irrespective of mental states that seem to be dependent on external circumstances out with the system's inner structure. The assumption that causation happens within a closed physical system, allied with the assertion that mental states supervene on conditions in the outside world, means that we are committed to the belief that consciousness can have no effect on the physical world and we are left with epiphenomenalism.

Sven Walter (2007), in the Internet Encyclopaedia of Philosophy, uses the analogy of inserting a counterfeit coin into a vending machine:

> ...the meaning or content of a mental state, being a relational property, threatens to be as irrelevant for our behaviour as the property of being a genuine dollar bill coin is for the behaviour of a vending machine.

The property of authenticity – as in authentic or matching behaviour – is a relational one, that is, it depends on other, external facts, and so the vending machine will behave in exactly the same manner regardless of whether the coin is genuine or not, so long as the fake coin has certain local, internal properties. Similarly, unless it can be shown that either externalism is wrong, or that relational mental properties can make a causal difference, then, like the irrelevance of authenticity of the dollar coin to the vending machine, our mental states are irrelevant to the behaviour of our own physical structure. This leaves us with two alternatives: either we abandon non-reductive physicalism in favour of outright reductionism, or we are compelled, however reluctantly, to accept epiphenomenalism.

The idea of mental causation then, is problematic for those who wish to contend that mental states can, and in fact do, influence the behaviour of humans and other conscious beings. If it can be shown that mental content has no such influence, then what is generally perceived to be the active motivation for all of our acts and behaviour, is no more than an inert, yet conscious, by-product of our inner physical composition and active engagement in the world. Our consciousness, since it is powerless to affect any aspect of our physical world, would have had no bearing on our evolution and would have been a mere spectator to the unfolding of the various stages of our species' development. Crucially, we as a species, in the absence of any conscious mental activity, would have reached the very same stage of our evolution at the very same time and consciousness would have been irrelevant.

(ii) Epiphenomenalism and the Creation of Synthetic Consciousness

One possibly interesting upshot of accepting epiphenomenalism is that it would raise questions about the legitimacy of any attempt to create a synthetic consciousness. If consciousness is causally ineffective, why spend time and money reproducing it in an artifact? Already we have the 'other minds problem', with regard to other members of our own species; a solution to this problem, it could be argued, would present an even greater challenge to those attempting to demonstrate the presence of conscious mental activity in an inorganic creation. If it can be shown, as epiphenomenalists are attempting to do, that consciousness has no functional role in our evolutionary development, but is a mere onlooker to this progression, then it would seem that we now not only have something whose existence is already very difficult to prove, but also has no effect on the vehicle with which it is associated. So why produce it?

In 1997 IBM developed a chess-playing computer named 'Deep Blue' which defeated the then world champion Gary Kasparov[8]. Although there were various accusations of cheating and human interference between games, but what is beyond doubt is that overall, the computer held its own against Kasparov (thanks in no small way to its ability to evaluate 200 million positions per second). Since 1997, there have been massive advancements in

[8] For more on this see: http://en.wikipedia.org/wiki/Deep_Blue_(chess_computer)

processing speeds, computer efficiency and a plethora of other technological advancements which would allow the creation of computers that would dwarf the capacities and capabilities of Deep Blue. However, in terms of its chess playing ability, Deep Blue reached a level of competence that very few, if indeed any, humans have attained: it defeated a world champion and grandmaster of chess and it did so in the absence of any apparent desires, fears, excitement, nerves or in fact any conscious experiences. In light of this, reductionists are now perfectly entitled to ask what improvements could conscious mental activity possibly have made to such a machine. Any answer to this question other than ‘none’ would, no doubt, prove very difficult to support.

But Chess is only one action or activity among the countless possibilities available to even the most humble of creatures in the animal kingdom. If we loosen the constraints of our imagination slightly and extrapolate the competences of Deep Blue into the future and to the creation of a ‘Mega-Robot’ that has been programmed to reach the accomplished level of proficiency attained by Deep Blue in all aspects of human activity – a task of undoubtedly gargantuan proportions – there still seems little or no reason for us to continue our quest for a manmade consciousness. Theoretically, there could exist in the future a machine which carries out all features of human activity to a level well beyond the capabilities of most humans, while all the time lacking in any conscious thought whatsoever. Again the reductionists can demand an answer to the question of why the additional capacity of consciousness is any way advantageous to such a supremely efficient machine. And if the creation of conscious mental activity is an end in itself, then such an end seems meagre in comparison to striving towards the goal of creating machines which, with the assistance of incredible levels of processing power, can attain a level of excellence in all aspects of human pursuit.

One reply to this could be that bestowing consciousness on our Mega-Robot would be seen in some respects as presenting it with a gift. In possessing the capacity for conscious mental activity, the robot could experience pleasant mental aspects of its existence, such as revelling in its successes, basking in the glory of its achievements and savouring its triumphs and conquests. Though there would also be the inevitable counterbalancing of the negative aspects of its conscious experience, such as the disappointment of failure and the ignominy of defeat. Further, even if the robot’s capabilities prevented these negative mental experiences from ever arising, then surely the dearth of challenges left for the robot to face would give birth to unbearable ennui. In any case, there seems nothing morally wrong in claiming that, while lacking in consciousness, the robot would be entirely unaware of its predicament (or of anything, for that matter) and so the decision to leave it without consciousness is ethically neutral. Another reply could be that, in creating a synthetic consciousness, we can gain a better understanding of our own consciousness, that the manufacturing of a thinking, conscious machine would provide us with a greater insight into the inner workings of our own mental processes and conscious thought. However, it would be easier to create another, inorganic consciousness, if we already knew how to satisfy the various conditions that give rise to consciousness in ourselves. Otherwise, we are relying on providence to endow us with the good fortune to stumble upon the creation of a consciously aware inorganic system. Further, unless we know what processes give rise to consciousness in organic beings, and we are able to replicate or simulate these processes, then it seems a solution to the ‘other minds problem’, and the ability to identify conscious mental activity in machines, recedes further into the distance. Put another way:

(1) We know P produce C in O, (2) we replicate P in artifact A to produce C in A. Now we know a priori that any O with P will be C, and we should know ceteris paribus that any A with P will be C, but all is not equal, that is, we can't state, let alone guarantee the ceteris paribus conditions.

Possibly a stronger reason for creating consciousness in a machine, is that the machine could then make a moral distinction between killing an enemy solder and killing an innocent child. Humans, when faced with moral decisions, choose to act in a particular way because they become consciously aware of the anticipated feelings associated with doing the right or wrong thing, feelings**,** which in part**,** are linked to the effects our actions would have on others and which, in part, have been instilled in us by our enculturation. The unpleasantness associated with acting in a morally reprehensible fashion functions as a restraint which (usually) prevents us from doing so. However, there is a difference between 'consciousness' and 'conscience'[9] and there is no logical reason to suppose that our super-efficient 'Mega-Robot' could not be programmed to make these distinctions without the necessity of subjective, negative conscious experiences**;** all of which implies that consciousness still makes no discernible difference in the robot's ability to act with moral propriety. Of course, it could be argued that consciousness would be of benefit in exceptional circumstances, where it may be morally permissible to kill the child rather than the soldier for example, and that the absence of consciousness increases the possibility of an agent's moral fallibility. Although it is true that consciousness may allow us to weigh up situations such as these to a greater degree, it does not guarantee that we will arrive at the morally correct decision (assuming there is such a thing). If it is moral infallibility that we seek, then we are almost certainly compelled to train our sight away from any direction that leads us to humanity. In any case, if epiphenomenalism holds, then conscience, being a mental quality and thus causally inert, would have no part to play in the robot's decision to kill either the child or the soldier.

Effectively then, although it may not be seen as a particularly key moral issue, any attempt to create something whose presence is very difficult to prove and which seems to impart no obvious functional or evolutional advantage upon the agent, is open to charges of being a futile and wasteful endeavour. It is important to show that consciousness is beneficial to the organism**;** a demonstration that must begin with proof that the there is such a thing as mental causation, that consciousness does indeed have a functional role and, ultimately, the rejection of the Epiphenomenalist position.

(iii) Replies to the Epiphenomenalist Position

That epiphenomenalism is counterintuitive does not provide any good reasons for its being false**.** The fact that it is unpalatable does not provide grounds for the dismissal of a theory, otherwise we would still be puzzling over the scientific anomalies that would inevitably occur whilst we ignorantly observed the Sun's orbit around the Earth. Further, there would seem to be little or no phenomenological difference in our experience of the world irrespective of whether epiphenomenalism is true or not. On what basis then, if any, can we dismiss epiphenomenalism?

[9] Though their etymology is the same, from Latin *conscientia* privity of knowledge (with another), knowledge within oneself, consciousness, conscience.

One possible objection comes from the suggestion that epiphenomenalism simply flies in the face of what we generally believe about our own minds as well as minds other than our own. In general, if we have a belief that X, then we act accordingly, and if we believe that 'not X' then we act differently. The fact that we can observe the acts and behaviours of others, which appear similar to our own, provides us with the justification to assign to others comparable, if not identical, inner mental experiences.[10] In order to make such a comparison, we must take the observable behaviour to be the effects of unobservable mental causes, an inference that the epiphenomenalist cannot make since he subscribes to an outright denial of any such thing as a mental cause. However, proponents of epiphenomenalism can simply say that the observable behaviour of others, if it shows anything, shows only that others have similar inner physical states to our own, states which cause mental states akin to ours, and that there is no need for the inference of mental causation**,** that the two are concomitant, not causally related.

Sven Walter (2007) cites Donald Davidson's 'reasons for/reasons for which' argument as a possible objection to epiphenomenalism. Davidson claims that although I may have a reason for performing a particular action, which I then go on to perform, it may be that it is not done for the original reason. According to Davidson, a distinction can be drawn between the reason for an action and the reason for which the action is performed: the action is caused by the reason for which the action is performed. Walter gives the following example, taken from Wilson (1997):

> Suppose, for instance, I want to meet my mistress and I believe that I can attain this goal by giving her a call; suppose I also have a second-order desire to get rid of my first-order desire and I believe that I can attain this goal by calling my psychiatrist. When I finally walk to the phone, it seems, I have a reason for doing so (my first-order desire plus my corresponding belief) which is not the reason for which I walk to the phone**.**
>
> *(*Wilson, 1997, p.72)

For the epiphenomenalist, there seems to be no explanation for the above scenario, since epiphenomenalism dictates that no action can ever be caused by a reason, in virtue of the fact that reasons are mental events. However, the epiphenomenalists can hold fast to their position simply by asserting that the reason for which the action is performed is a by-product of the physical events which give rise to the action.

A more significant problem for epiphenomenalism is that it makes the claim that each and every one of our actions, judgements, thoughts and beliefs are determined by physiological or neurological processes beyond our control; in other words, epiphenomenalism commits us to a position of determinism. Epiphenomenalism does not allow the possibility of having rational beliefs, since all of our beliefs are determined by our physical and neurological states at any particular time; mental activity is causally inert, even to the point of being powerless to cause other mental activity and so my belief that today is Wednesday cannot be rationally justified by my belief that yesterday was Tuesday, or by my belief that the day before was Monday. The belief I have now that today is Wednesday, for epiphenomenalists, comes

[10] There is, of course, the possibility that we are born with an endogenous grasp of other human beings, but in everyday living we assign certain mental states to others based on observable behaviour. If, for example, I see someone crying, then I tend to assume that they are in a state of sadness.

about by way of my inner physical states, which means that I would be entirely unable to provide any basis for judging my belief to be either true or false, since a deterministic doctrine would prohibit me from arriving at any other judgement. Further, this deterministic standpoint raises moral questions concerning responsibility and culpability.

Marian Stamp Dawkins (1993), in reference to Bernard Baars' '*A Cognitive Theory of Consciousness*' (1988), discusses another possible argument against the idea of a causally inert conscious mind:

> The very fact that we can identify cases where consciousness is a help and where it is a hindrance suggests that it has a definite function in some circumstances and is not just an 'extra' coming along for the ride.
>
> (Dawkins, 1993, p171)

The point made by Baars, which Stamp Dawkins picks up on, is that there are certain actions which we perform more effectively when we are not conscious of them, and there are those which we perform better while they are at the forefront of our conscious thought and which require our strict, undivided attention. An example of this may be driving. When we are first learning to drive, it is necessary for us to have each individual stage of the process at the forefront of our thoughts. We may think along the lines of: *press the clutch with left foot - move gear stick to the left and upwards into first gear - allow clutch to move slowly off the floor - listen out for change of engine noise - press the accelerator lightly with right foot - slowly allow the clutch to come up the rest of the way - keep pressure on the accelerator - listen out for engine revving loudly - press the clutch in with left foot - pull the gear stick downwards into second gear...* and so on. At this phase of our learning, when we consciously think about each stage of the process, we tend to make several mistakes and we may even regularly stall the engine or pull away too quickly. However, after driving for a few months or years, and having gone through this process many times, it becomes automated and non-attentional, freeing up conscious attention for other tasks. Indeed, so long as we are driving to a familiar destination on roads we know well, we can be barely conscious of anything driving-related at all and can quite happily think of other things or hold a conversation or sing along to our favourite songs. Only when something unexpected occurs, such as a new road layout, or an obstacle in the road, do our thoughts become refocused on some aspects of the driving process. This idea of consciousness being of great assistance holds true for most novel situations, when we come up against unpredictable events or even things which have to be re-evaluated or worked out again from the beginning. The important point to note is that there are times when consciousness can help us and times when it can hinder us, and the very fact that we can make such a distinction means that consciousness must have some function, which has resulted in its being selected during the evolutionary process. Of course, this theory does not provide us with a specific function of consciousness, but it does suggest that our development would not have been as successful without the capacity to deliberately focus our thoughts on the things that we have not yet learned, as well as being able to perform the things we have learned effectively and efficiently in the absence of focused, conscious attention. A lack of this first ability would have made learning far more difficult, if not impossible; a lack of the second would have made the performance of already learned actions infinitely more laboured and the learning of new tasks problematic, with the consequence of stunting our evolution and development as a species.

Fred Dretske (1997) adopts a similar approach and claims that consciousness allows us to do those things without which we could not do. Following Rosenthal (1991), Dretske makes the distinction between 'creature consciousness' (being conscious of things and facts), and 'state consciousness' (certain mental states which, although they themselves cannot be conscious of anything, can give rise to consciousness in us). The benefits of creature consciousness for living things, according to Dretske are obvious: it allows us to hunt, feed, avoid obstacles, survive, build homes, reproduce and escape predators, among many others. Without perception we are, according to Dretske, mere vegetables, and animals**,** for example**,** gazelles**,** lacking in such perception would be far more at risk of being caught by predators than their conscious, perceptive conspecifics. With respect to the benefits of 'state consciousness', Dretske adopts what he describes as an 'act conception of state consciousness' where a conscious state is one that makes an organism conscious and not, as claimed by Higher Order theorists, an awareness of the fact that we are having a conscious experience. For example, on seeing a lion, a gazelle's visual experience can be described as a conscious state because it makes the gazelle conscious of the lion whereas, on the Higher Order theory rejected by Dretske, the gazelle could see, hear, smell and fear the lion but would not be occupying any conscious state whatever. The advantage of 'state consciousness' then, is that it produces 'creature consciousness', which according to Dretske, is crucial to an organism's survival.

However, this leaves open the question of why phenomenal consciousness has evolved since, if conscious organisms are better equipped for survival, it seems that we only require awareness of the facts about objects and not of the objects themselves. Returning to the gazelle example employed by Dretske, for the gazelle to increase its chances of survival, it only has to be aware that there is a lion nearby, that the lion is approaching, and so on; there seems to be very little need for the gazelle to have a visual experience of the lion's appearance for its motivation to flee. If there are ways to receive information other than those from experience, then why do we need experience and so consciousness? In fact, Dretske refers to Humphrey (1970), who worked with a monkey, Helen, who had her visual cortex removed and with it, her capacity for normal vision. Despite initially not looking at things, Helen managed to regain some of her visual faculty. She was able to avoid obstacles while moving around a room, pick up objects from off the floor and even catch flies. Effectively, despite not being able to see the objects, Helen was able to form some representation of the location of the objects, that is, despite being unable to see the objects and so experiencing no visual awareness of them, she knew they were there. However, as a result of losing her visual awareness, Helen was unable to recognise or identify what the objects were and this, according to Dretske, is the possible function of phenomenal consciousness: to allow us to recognise and identify objects. This in turn allows us to perform actions which those who are unable to recognise and identify these same objects cannot. Further, humans who are afflicted in this way, do not regain their vision to the same extent that Helen did and so, despite possessing this 'blindsight', they find it much harder to adapt in the way that Helen did, meaning that the loss of phenomenal experience has a far greater negative impact on humans. Dretske puts it thus:

> If we assume (as it seems clear from these studies we have a right to assume) that there are many things people with experience can do that people without experience cannot do, then that is a perfectly good answer to questions about what the function of experience is.
>
> (Dretske, 1997, http://evans-experientialism.freewebspace.com)

In essence, this is not too dissimilar to Baars' claim that there are things that we do better when consciously focusing on them; both Baars and Dretske agree that consciousness did indeed play a significant role in our evolution and continues to be vital in our goal for survival.

If our rejection of epiphenomenalism rests upon the claim that there is a difference between the actions we perform consciously (which usually fall under the category of 'novel' or 'unpredictable') and those we do without our consciously focused attention, then a similar argument could be used in the case of animals and synthetically conscious machines. If it can be shown – and I, along with Baars (1988) and Stamp Dawkins (1993), believe that it can – that there is a distinct disparity between the efficiency with which we perform actions which require our undivided, focused attention and those which we perform (usually after some repetition) subconsciously, then a similar disparity in non-human agents would provide strong evidence that they do in fact experience conscious mental activity. Indeed, it is not difficult to find cases where non-human animals have been presented with novel situations where they may take a while to come up with the correct course of action but, after a while, begin to perform the action with a high degree of competence. Instances such as these could always be seen as no more than the development of muscle memory, which enables the body, through repetition, to act in a kinaesthetically adept manner under a given set of circumstances; but, reaching a point where the muscles 'remember' requires a conscious effort at the beginning to make them move in a manner appropriate to the conditions and so some level of focused, conscious thought would have been required even if the 'muscle memory'[11] theory holds true. In any case, there is evidence of animals displaying behaviour which cannot be accounted for by muscle memory.[12]

We'll now shift our focus to how significant the attribution of consciousness would be in the formation of our moral attitudes towards non-human agents.

(iv) The Moral Significance of Consciousness

In any society, there are any number of varying moral attitudes towards a range of issues, whether they are about the environment, various aspects of crime, civil rights, international affairs, raising children, abortion or euthanasia, and then there is the overarching issue concerning how we treat others. In general terms, the way we treat others often centres on how we would or do make them feel by acting in a particular manner towards them, and our conduct towards other people can provoke an inestimable number of feelings, ranging from blind rage to elated exultancy and everything in between. Generally, we are judged morally on how our actions and conduct affect the feelings of others, feelings which can only be experienced by conscious and sentient beings. Even our treatment of inorganic material objects is usually deemed moral or immoral depending on how it affects the feelings of others. For example, if I, out of a feeling of boredom, grab a baseball bat, go outside and

[11] For a more in-depth look at muscle memory see Julian Dow's article at: http://jeb.biologists.org/cgi/content/full/207/1/11]

[12] One striking example of this is that of the Japanese crows, who strategically use pedestrian crossings and car wheels to crush food which they find too difficult to eat (See http://www.youtube.com/watch?v=BGPGknpq3e0). Here there seems strong evidence for the presence of conceptualisation and forward planning, in the absence of any repetition of muscle movement.

begin to destroy the car across the road, then I will inevitably be placed under moral scrutiny. From the car there will be no screams, no howls of pain and no attempt to avoid the vicious and destructive swings of the bat, which disfigure its body more and more severely with each crashing thud. Any moral judgements that arise from this behaviour are based either on the feelings of the car's owner, or on my vicious and destructive personality and not on the experience of the car itself. If however, the car has been built in such a way that has given rise to the capacity for conscious thought (in a similar fashion to KITT, the conscious and solicitous hero of the 1980's 'Knight Rider' television series), as well as emotions and the ability to feel pain both physically and psychologically, how far would our moral attitudes shift away from the human owner's feelings and move towards the predicament of the car itself? I contend that although the general, middle of the road moral stance would move towards the predicament of the car, it would only do so as a result of the car manifesting what we perceive to be human or, at the very least, sentient responses.

This contention is based on the argument I used earlier, that which gives us the belief that 'dolphin friendly' tuna is morally preferable to other kinds of tuna: our inability to remove ourselves completely from our subjective, anthropocentric standpoint and our propensity to reserve our sympathy and empathy for things which seem to possess traits and characteristics closer to our own. Audiences felt a connection to the car in Knight Rider not only because it could talk, but also because it displayed other human-like characteristics such as compassion, a concern for human life, as well as a dry, deadpan and often sardonic sense of humour. In the absence of any human owners, a car like this would clearly draw more sympathy if it were being mercilessly attacked with a baseball bat, than the car across the street ever would. Even if both cars were identical in shape, constituent parts and appearance, the car which could respond in a conscious manner, that is, the car with which we would adopt an intentional stance [Dennett 1989] would demand far more compassion than its lifeless, unconscious doppelgänger. However, if a similar, equally capable, consciousness were placed in a vessel which also had the physical appearance of a human, such as C3-PO in *Star Wars*, Data from *Star Trek*, or Pinocchio, then I contend that these, and not KITT, would provoke our strongest inclination to protect them from harm or destruction. As with animals, the greater the likeness to ourselves we perceive in an artifact, the greater the moral regard we feel towards it. Does this mean however, that if animals, like the car across the street, were found not to be in possession of consciousness that we can be entirely dismissive of anyone who campaigns for their wellbeing and ethical treatment?

The answer to this question should be 'yes' - but we probably wouldn't. The infliction of emotional distress to the owner aside, there would be no moral difference between inflicting physical damage to the family pet lacking in any consciousness and belligerently hammering away at the inanimate car across the street. The screams and howls of a non-conscious dog would be analogous to the piercing racket of the car's alarm system, the bruises and bumps on the dog's body no different to the scrapes, dents and bashes on the car's chassis. Generally speaking however, we would still perceive a moral difference in attacking the car and beating the dog because the dog, unlike the car and despite lacking in any conscious activity, still has the capacity to act in a manner reminiscent of certain human behaviours, and to perform actions capable of tugging on our anthropocentric heart strings.

It can be argued however, that these claims about anthropocentricity and the moral difference of consciousness do not always hold, since there are those in the world who would, with relative indifference, sacrifice someone else's dog to protect their new automotive pride and joy, and those who would, given the choice, donate their last remaining coin to cure their dog from disease rather than to a human stranger with the same affliction. There have even been cases of people leaving millions of pounds to their dog or cat rather than to their relatives or to some human charity or other, and there seems to be no moral contradiction in them doing so, after all cats and dogs don't behave with the self-seeking avarice displayed by some relatives**.** Indeed, any attempt to persuade the affluent pet owners to deny these opulent lifestyles to their dogs or cats and bequeath the money to a human, or a cause benefiting humans, could leave us open to accusations of species chauvinism [Ryder (2001) and Singer (2002)]. However, most moral judgements, as well as the laws of the land to which they give rise are, rightly or wrongly, based on objective uninvolved attitudes (the legal system might look very different today if, for example, the penalty for stealing was set by victims of burglary) and although there will always be those who feel it is morally justifiable to prioritise the care of a car over that of a dog, or even a human, those who have neither pets nor cars will, perhaps quite naturally, have higher moral regard for a dog than a car, and for a fellow human than for a dog. This inclination is formed in no small part on the approximation to which the agent resembles, in appearance and action, a human. Feelings and emotions play a major part not only in our moral judgements and our decision to treat others in a particular manner, but also, consequently, in our ability to assign consciousness to that agent. Effectively, our moral judgements are often based, whether directly or indirectly, on the feelings and consequent emotional states produced in others, feelings and states which, because of their felt nature, can only be experienced by conscious beings. And even when we are unable to assign feelings, emotions and consciousness to things we will tend to favour that which can perform in the way most resembling human actions. From a moral standpoint then, the perceived possession and manifestation of consciousness most certainly does matter. The intentional stance becomes a moral stance.

(v) *Emotions, Beliefs, Desires and Temporality*

The capacity to experience pain alone is not a sufficient condition for an agent to be granted moral consideration. There are innumerable occasions where pain can be inflicted without any moral or ethical impropriety. For example, there can be no doubt that young children experience pain when they are given injections in order to build up their immunity to a variety of illnesses, and there are similar reasons for putting non-human animals through such discomfort. Only when there is a continued, sustained application of pain, whether it be physical or psychological, does the organism enter the realms of what we call suffering. Pain, even extreme and unnecessary pain, does not by itself carry the same moral weight to that of suffering since suffering not only involves more than minimal or mild pain (a point made by DeGrazia, 1998), but involves pain that continues over a sustained period of time:

> Suffering is not the same as pain. Pain without suffering can be caused by an ordinary hand pinch, for example…Suffering is a highly unpleasant emotional state associated with more-than-minimal pain or distress.
>
> (DeGrazia 1998, p116)

Although there is no definite length of time at which pain turns to suffering, it seems clear, for example, that a sharp, painful pin prick, where the pain only lasts a matter of seconds and happens only once, cannot really be seen as causing suffering to any great degree. If however, this procedure is repeated at regular intervals over a sustained period of time and the subject begins to feel fear or dread at the anticipation of being subjected to it, then we can say that he or she is suffering. We must be careful however, not to fall into the trap of using the term 'suffering' in the casual manner in which it is used in common, everyday language. DeGrazia claims that for suffering '*we need more than minimal pain or distress.*' I agree that in order to suffer, we need to go beyond things like mild irritation or discomfort, but we also need to involve perpetuation and/or frequent recurrence.

The same criteria can be applied to an animal who is subjected to a certain amount of pain in a medical research laboratory. A substance can be applied to the animal's skin which causes the animal extreme discomfort, but only when this process either causes long-term irritation or damage**,** or is repeated at regular intervals does the animal begin to *suffer*. Of course, there will be those who argue that the infliction of even the slightest amount of pain on the animal is unjustified, but generally, it seems that experiencing acute, short-lasting pain does not cause a being to suffer; it would have to experience pain of a chronic nature, or acute pain that was experienced at frequent intervals, before it could be said to be suffering. I propose then, that suffering be defined as *pain or distress which is more than minimal and which persists over a sustained period of time or recurs at frequent and unpredictable intervals*.

DeGrazia (1998) talks about other important concepts, which I would argue are necessary for any kind of pain to be termed 'suffering': some basic sense of temporality, beliefs and desires, an interest in one's own continued existence, and emotions.

(a) *The Manifestation of Pain*

Pain and/or suffering, looked at from our anthropocentric perspective is, I believe, the ultimate determinant of moral consideration, since the level of pain or suffering an organism can experience has an effect on a great many issues. The extent to which any conscious entitiy can enter into states which we can term 'painful' or 'full of suffering' is a major factor in the level of moral consideration we grant it. For even the having of interests is only of any significance if the denial of the pursuit of these interests in some way causes distress to the organism in question and, of course, distress is a form of pain (whether it be physical or mental) or, in the longer term, suffering. Failure to permit agents to satisfy these desires will result in unpleasant physical and psychological states such as frustration or boredom, which themselves are forms of suffering. In addition to these interests, the agents would also seem to have an interest in pain avoidance or to live pain-free lives, and so if it is shown that a conscious agent does have such interests, then by being subjected to pain or suffering, they are being denied the right to pursue at least one of them.

Ethically, it is important to establish which non-human beings have the capacity to experience pain, and to what extent. Establishing the extent to which an agent can experience pain provides us with further evidence of the extent to which that agent is phenomenally conscious. Further, we identify an agent's capacity for feeling pain through observation of its behaviour and the physiological alterations that it undergoes; we then make moral judgements based on what we observe and the extent to which we, as humans, can identify with them. This idea is underlined by the fact that a higher degree of empathy would be felt towards a creature which suffered a relatively innocuous injury yet exhibited more 'pain-like' behavior; that is, an apparently unmoving caterpillar starving to death would provoke less sympathy than a beetle with a limp.

Of course, pain behaviour is certainly not the only, or perhaps even the best, method of detecting the level of pain a creature is experiencing since there can be massive differences in physiology, as well as in the resultant pain behaviour across species. The observation of physiological changes in an organism may provide us with a better idea of the extent to which it can experience pain. Phenomena like increased heart rate, respiration, and perspiration, and the release of pain reducing chemicals such as adrenaline or opiates could provide us with a better idea of the level of pain experienced by an organism. Neurological activity and alterations observed when non-human animals are subjected to conditions that would normally be considered painful or stressful in humans may also be of more worth than merely observing behaviour. The difficulty with these tests is that they are not directly and readily available to the majority of us and so the attributing of pain to an organism is derived, perhaps unjustly, from the behaviour it exhibits and whether or not we can recognise it as pain indicating behaviour. Further, we make these observations from behind our subjective, anthropocentric veil (about which I will say more in Section 4) and so, when available to us, compare the physiological changes of the organism to those humans who are subjected to pain.

Naturally, the level of sympathy that we feel has a major effect on the level of moral consideration we are prepared to give, and so it follows that an organism's ability not only to experience, but to display the effects of physical pain goes a long way in the determination of moral consideration. So, physical pain and an agent's reactions to it are important determinants in our moral attitudes towards both human and non-human beings.

(b) Temporality and Memory

Although it is not necessary for an agent to have the sophisticated and complex concept of time possessed by humans - such as the use of seconds, minutes, hours and days, or the complex use of a calendar system - it does seem to be the case that some, even sensorimotor grasp of the passing of time is essential if there is to be evidence of suffering. If there is no experience of the passing of time and only an awareness of the current and present moment, then talk of suffering, at least in the definition postulated earlier, is nonsensical. Living from one painful moment to the next without any recollection of what went before does seem like an extremely unpalatable existence (living like that even without the pain seems unpalatable). However, by allowing any being in constant pain to become aware of the fact it is constantly in pain, and will continue to remain so, compounds the unpleasantness of the experience.

For a human, being conscious of the fact that the painful experience has persisted for some time and is likely to continue beyond the current moment seems to add an additional stress or painful element to the situation, an additional stress that does not accompany short-lived, temporary though acute pain. Put another way, having the idea of previous experience (memory) and some concept of continuation or the future, is the difference between acute, temporary pain – which does not come with the anticipation of more pain – and chronic pain, where only the latter, at a particular threshold level and with a degree of persistence, can be worthy of being defined as suffering. There is no reason to suppose that a non-human being, at least one that has at least some sense of the passing of time, would not experience this additional stress at the experience of persisting and unrelenting pain. So the extent to which a non-human consciousness has the capacity to experience at least some kind of temporality has a bearing on the level of moral consideration afforded to it, since the absence of such a sense means that only acute, temporary pain can be experienced and does not extend itself to suffering.

Of course experiencing the passing of time can have the opposite effect of allowing us some comfort rather than increasing the stress we experience. For example, if I fall from a tree and break my arm, despite the fact that I will experience terrible pain – and maybe even to some extent suffer – for a period of time, having the ability to look to the future means that I am able to console myself with the thought that after a few months my arm will heal itself and the pain will subside and eventually go. Whether or not it is possible for non-human organisms to have a concept of the future and, if so, to what extent, is a matter of some debate.

It is not only physical suffering on which temporality has a bearing. Emotional or psychological suffering can also be a consequence of being able to draw a meaningful distinction between past, present and future times. Having a sense of the passing of time can have a significant influence over our emotional and psychological wellbeing. People often find justification for keeping goldfish in small bags filled with water by referring to the controversial claim that goldfish have very poor memories[13]. This lack of a memory, so some would have us believe, means that the fish are unable to recollect anything that happened only seconds before, and this prevents the arising of any negative emotions such as boredom, frustration or loneliness. However, our sense of the passing of time comes from our observation of change in the world around us; why would it be any different for other animals? If they're able to perceive change, they must be able to have some sense of time passing.

Similarly, the absence of a temporarily enduring experience and the assumption that animals only live in the present is one way of justifying their subjection to severe mistreatment. Since they are not aware of any time passing, so the argument goes, they are unable to feel anything like boredom, frustration, isolation, loneliness or suffering; something one might find very hard to believe if visiting some of the poorer zoos with limited space and facilities.

Singer (2002) describes experiments at a United States Air Force base in Texas, where monkeys were trained to operate a kind of flight simulator, which could pitch and roll like an aeroplane and was kept horizontal by way of a control stick. The training involved giving the monkeys electric shocks each time they allowed the simulator to deviate from the horizontal, thus teaching them to return the platform to the correct position. When the monkeys had learned to operate these devices correctly, they were then administered various drugs, radiation and chemical warfare agents (as well as the electric shocks to encourage them to keep the platform straight) in order to observe the effects of these agents on the monkeys' ability to operate the simulators. Naturally, the monkeys became frightened, sick, uncoordinated (for which they were given more electric shocks) and most eventually died. They also exhibited 'pain behaviour' such as struggling, screaming and apparent anxiety. In spite of this, the monkeys, while it was physically possible for them to do so, made every effort to keep the platform in a horizontal position and thus avoid the pain that accompanied the shocks dealt out as a punishment for failure. Apart from the moral objections to subjecting the monkeys to these harrowing experiments, the experiments themselves reveal a great many things about the monkeys' behaviour and the possibility of their entering into certain psychological states. They also reveal a great deal about the mental states of human beings who are willing to run the tests!

[13] See Gee (2003) for a refutation of this claim

The monkeys, after a fairly short period of time, were able to make the connections between certain bodily movements (reaching out for the control stick, moving the stick in a particular way**,** and so on), the angle of the platform**,** and the cessation of the electric shocks. Consequently, they began to attempt to move their bodies in such a way as to keep the platform horizontal in order to keep the painful shocks at bay. In order for the monkeys to have made this connection - since it seems highly unlikely that the monkeys had a natural instinct for flight simulation – they must have had the capacity for at least a minimal memory, which itself is the retaining, accessing and synthesis of past events. Also, their behaviour shows that the monkeys preferred not having the shocks to having them, otherwise why else would they always have attempted to steady the platform each time a shock was applied (and, after a while, *before* the shock was applied). This is a crucial point since anticipation demonstrates the presence of temporal experience**.** If the monkeys had no preference, the number of attempts to straighten the platform would have been expected to be arbitrary. So not only did the monkeys form beliefs and memories about the effects of their bodily movements and the electric shocks, but they also had a desire not to be given the shocks. This further strengthens the claim that the shocks produce more than mere physical reactions. (Singer notes that the monkeys struggled and had to be restrained in the chair, which suggests that, had they been given the option, they would not have taken control of the simulators voluntarily).

Solely from these experiments, there already emerges fairly strong evidence of the monkeys being able to remember, to form beliefs about the future, and this results from the possession of desires or preferences. Whether the monkeys had similar phenomenological experience of the pain to that of humans placed under similar circumstances remains unclear, though the best guess, given the physiological and neurophysiological similarities amongst mammals is that they do**.**

Clearly an awareness of the passing of time can have an impact on moral consideration since it can move an agent's experience of pain to one of sustained and continuous suffering. It does not seem necessary however, that the agent possesses any sort of detailed or complex idea of objective time in order to experience negative emotions such as frustration or boredom. Husserl's (1927) notion of an 'inner time consciousness', in which moments we would term 'recently past' are not, strictly speaking, in the past at all, but are 'retentive', meaning that they resonate in the present. This would seem sufficient for the generation of such emotions. Further, Husserl's idea of our experience being 'protentive' - where there is an expectation, based on what's gone before, of what will happen in the immediate future – seems sufficient for the creation of worry or anxiety. For Husserl, there are no isolated and individual moments; each 'now' moment contains traces of past moments as well as a pre-noetic expectation of what's to come. Such a notion of time would allow an agent to be free of being 'stuck in the moment' without the need for any complex concept of time. Indeed, this notion is what frees most human beings since very few of them actually possess a complex concept of time.

(c) Beliefs and Desires

Desires are an important aspect in the determination of moral consideration since they have significant implications for what has been described as non-physical pain and, as a possible consequence, physical pain. It is very difficult to enter into any moral or ethical debate about the treatment or predicament of any human or non-human consciousness if it can be shown that the being in question is entirely indifferent towards its circumstances: there seems no obvious reason for us to bestow any ethical weight upon any situation in which an organism is unmoved by its own plight. In order for the predicament to become a moral issue, it has to be shown that the conscious entity *desires* to be in a situation other than the one in which it finds itself, although, I contend, it is not necessary that be in a position to express these desires propositionally. Consequently, any artificially created consciousness must be endowed with desires, if it is to be given any moral consideration, even if these desires are of a basic nature such as the desire not to feel physical pain or the desire to roam unconstrained**.**

For Descartes, and others who followed, including Davidson (1975) and Frey (1980)**,** language is what most distinguishes humans from non-human animals and, by virtue of its absence, shows that animals are not conscious and so should be afforded no moral consideration:

> For it is highly deserving of remark, that there are no men so dull and stupid, not even idiots, as to be incapable of joining together different words, and thereby constructing a declaration by which to make their thoughts understood; and that on the other hand, there is no other animal, however perfect or happily circumstanced, which can do the like. Nor does this ability arise from want of organs: for we observe that magpies and parrots can utter words like ourselves, and are yet unable to speak as we do, that is, so as to show that they understand what they say; in place of which men born deaf and dumb, and thus not less, but rather more than the brutes, destitute of the organs which others use in speaking, are in the habit of spontaneously inventing certain signs by which they discover their thoughts to those who, being usually in their company, have leisure to learn their language. And this proves not only that the brutes have less reason than man, but that they have none at all…
>
> (Descartes, 1637/1997, p43)

In short, the absence of language proves an absence of consciousness. Descartes goes on to say that since animals have no reason and therefore are not conscious, they are mere automata and can be treated in any manner in which humans see fit, even to the point of the most extreme forms of physical abuse. Frey (1980) maintains that in order to have a belief, we must believe that a particular sentence is either true or false and animals, since they lack language, are unable to determine the truth or falsity of a sentence, therefore they cannot possess beliefs. Further, Frey goes on to argue that all desires are underpinned by beliefs and so animals, in light of the fact that they cannot form beliefs, are also incapable of possessing desires. Presumably this would include the desire to avoid pain and suffering and so again we have, for slightly different reasons, the result that there is no moral obligation on our part not to mistreat animals or, presumably, any agent not in possession of the ability to utilise language.

Aristotle, in *De Anima iii 10*, argues that desire, working in tandem with practical reason, is the main cause of animal movement and action. Desire alone cannot initiate movement or action since, for example, there are those who have base and depraved desires and yet do not act upon them. Practical reason is therefore required since it allows us to identify an object as something desirable and effectively, we desire objects or occurrences of events which seem to us to be good or desirable. David Hume (1739/1985) makes the claim that reason is inert and by itself cannot motivate us to act. Peter Singer puts it thus:

> ...our beliefs tell us how the world is...An action is thus the product of these two forces: a desire representing the way the world is to be and a belief telling us how the world has to be changed so as to make it that way.
>
> (Singer, 1993, p 401)

It is not clear that beliefs do in fact tell us how the world is, rather than how it appears to us, but I agree that both belief and desire are necessary components in our decisions to act. However, in order for an agent to be considered morally relevant, it is not necessary, as assumed by Descartes, Frey and Carruthers, that these beliefs and desires need to be propositional: it is not necessary for the agent to express in language the desire *that* X occurs or to believe *that* in order to attain X, it must do Y. Robinson (1995, 2004, 2005), proposes a non-cognitive theory of emotions which states that judgements and propositional attitudes are not part of the emotion process. Take, for example, a hungry, crying infant. Although it can be said that**,** in any propositional sense**,** the baby cries because it is hungry, it is too far-fetched to suggest that the baby is crying because it has a desire for food and *believes that* by crying this desire will be satisfied. Indeed, it is even more of a stretch to suggest that, as Aristotle would, that the baby's faculty of practical reason allows it to see its foods as an object of desire. Crucially, the baby's inability to express its desire for food propositionally in no way eradicates, or even lessens, the pain or discomfort brought about by its hunger.

Marian Stamp Dawkins (1993) describes experiments carried out by Norma Bubier, of Oxford University, on battery hens. The point of the experiments was to determine the lengths the hens would go to in order to obtain something which they might desire, for example, food, comfortable nestboxes, more comfortable flooring, perches and companionship. In order to obtain these things, the hens had to pass through a gap, which varied in size from a large gap that the hens could pass through easily, to a very small gap which the hens had considerable difficulty in squeezing through. Only if it mattered to the hens would they make the extra effort of squeezing through the smallest gap to get to food, nestboxes and comfortable flooring. In particular, the hens made the greatest effort to get to nestboxes when it was time to lay their eggs; they would abandon everything else and put all of their energies into getting to the nestbox. This, according to Stamp Dawkins, suggests that millions of battery hens experience frustration at the inability to find a comfortable place in which to lay their eggs. The truth of this claim[14] is not crucial for our purposes. What is important, is that the hens' behaviour suggests that they possess desires for things (nestboxes, foods, comfortable flooring) as well as beliefs about the location of these things and how to get them. These desires and beliefs however, are expressed through the hens' behaviour and not in any way that could be described as propositional.

Going a step further than the non-cognitivists, William James (1884), argues for a somatic feedback theory, which claims that each emotion has a unique and corresponding bodily response meaning that there will, for example, be a different and unique set of bodily changes for desire, anger, fear and sadness. Whenever one of these bodily responses occurs, the mind registers the response and the resulting mental state is the emotion. Consequently, there can be no such thing as an emotion in the absence of these bodily responses. Damasio's (1994, 2001) account differs slightly from James' in that the process through which emotions arise does include cognitive evaluations. For Damasio, we form thoughts and evaluations of the circumstances presented to us, which in turn causes bodily responses; it is this process that

[14] Being able to see the nestboxes and being obstructed in their efforts to get to them may have caused more frustration than for hens that simply have never seen or experienced them.

Damasio defines as an emotion. The bodily response is then produced in the somatosensory cortices of the brain to produce what Damasio defines as the feelings, which are essential to our decision-making process.

Essentially, although humans have developed complex, linguistic, propositional methods of expressing their desires, there is no reason to suppose that this in any way increases the pain or discomfort experienced at the inability to satisfy these desires. To be morally considerable, it is not necessary for a synthetically created consciousness to express its desires propositionally; it is only important that if the agent experiences circumstances, which are detrimental to its overall wellbeing, that there exists a corresponding desire to re-establish the conditions conducive to its wellbeing. Any moralising then concerns the extent to which we assist or hinder the agent in its attempt to attain these conditions in the same way we are morally judged on the lengths we go to in our attempts to satisfy the baby's desire for food, hence alleviating the pain or discomfort we assume on good grounds that it experiences.

As far as belief is concerned, I concur with Aristotle (1985), Hume (1739) and Frey (1980) that desires must be accompanied by beliefs, though, as already written, I reject the idea that these beliefs need to be propositional. Suppose, for example, that I desire to write a note. So long as the pen I wish to write with is within reaching distance, I pick it up and begin to write. Although it is fairly clear that I must, in some way, believe that the pen is something that I can write with and that I believe that I am able to pick it up without leaving the comfort of my chair, these beliefs are expressed in a behavioural, rather than a propositional, manner: the desire I have is 'saturated' in the beliefs I have about the pen. Effectively, for any synthetic consciousness to be morally considerable, it requires the possession of desires, which in turn must be accompanied by belief, which, contrary to cognitive theories of emotion, need not be expressed propositionally.

It is worth mentioning that the allowing of an organism to act on its desires is not always a positive thing. Allowing someone with the intention to murder or maim to fulfil or satisfy their desires can rarely, if ever, be morally justified, and so there seems to be an argument for placing less moral emphasis on desire than I have done. To this I offer two replies. Firstly, although the thought of allowing a murderer freedom to act upon his desires is abhorrent, it still seems to be the case that by being denied this freedom to act, the murderer will experience unpleasant states such as frustration or anger. Secondly, I do not contend that each and every conscious organism should be granted the freedom to act upon every desire, but only that the absence of any desire renders any moral debate insignificant, since how else can we make moral judgements on pain or suffering if the organism in question has no desire *not* to find themselves in these states? For if there is no desire to alleviate one's own pain or suffering[15] then only one of three conclusions can be drawn: either we are not conscious of the pain or suffering; we are not troubled by it; or it is something which we have deliberately brought upon ourselves (we may, for example, have masochistic tendencies).

I contend that in the first two cases what is experienced *ma*y in some ways be painful, but cannot in any significant way be termed as 'suffering', and in the case of the masochistic tendencies, alleviation of the pain or suffering might lead to a more significant pain or suffering. For example, Silas, the Opus Dei monk from Dan Brown's '*The Da Vinci Code*' inflicts upon himself pain and suffering by denying himself any physical pleasure or by

[15] This includes the suffering we experience from the observation of the suffering of others

applying direct physical pain to his own body (predominantly through the wearing of a cilice). Silas would argue that the alleviation of these pains would lead to other, more intolerable pains such as a feeling of failure in the eyes of God or a lack of spiritual fulfilment. Whatever the reason, Silas has expressed a preference and a *desire* to experience one type of pain or suffering over another. Under these definitions desires, and hence beliefs, then are crucial factors in the determining of how far a being is worthy of moral consideration.

(d) Emotions

One of the major aspects of being human which distinguishes us most decisively from non-humans is the range of emotions that we can experience, or at least our ability to describe them. Although there has been some documented evidence of higher level primates being embarrassed or bashful, animal emotions tend to be viewed as more primitive than some of those experienced by humans. Although they appear to enter states that would be regarded as first-order emotions such as fear or anger, it is harder to show that any non-human animals have the capacity to feel second order emotions such as pride or shame. One reason for this is offered by Adam Zeman:

> Second-order emotions require a sense of self as the object of others' attention, or as a moral agent subject to praise and blame.
>
> (Zeman, 2004)

Here, the significance of any non-human being possessing the capacity for experiencing second-order emotions is that not only does it seem to attribute at least minimal self-awareness and hence a higher level of consciousness, but it draws such beings further into our moral universe or, as Zeman claims, makes us more able to define the being as a *'moral agent subject to praise and blame'*.

When a dog soils a new carpet, although we may express our anger at the dog by raising our voice or through physical punishment, we do not do so in an attempt to shame the dog into never repeating such behaviour. Rather, our aim is to make the dog *fear* the consequences of soiling the carpet, a fear that we hope acts as a deterrent to any repeat behaviour and as an incentive to do its soiling somewhere other than the carpet. On the other hand, one method employed when attempting to potty train a toddler, is to appeal to these second order emotions. We may praise her when she successfully uses the potty, but may say things that are designed to cause mild embarrassment or shame when she fails to do so, such as telling her that she is getting too old to wear nappies or that all of her friends use the potty. Humphrey (1977) argues that 'abusing' children can have a positive impact since it can '*educate children in the knowledge of disagreeable feelings*'. Humphrey argues that with all social animals, parents, particularly when weaning, become more hard-hearted towards their offspring in order to give them first-hand experience of negative emotions. Crucially, although the aim is roughly the same with the dog and the toddler, we appeal to very different emotions in order to achieve this aim, typically because animals are not reckoned to be capable of experiencing these second-order emotions - not to mention the fact that the causing of fear in a toddler would be seen as tantamount to cruelty, certainly more so than creating fear in the dog.[16] There may be those who argue that the dog, after being scolded,

[16] The fear created in the toddler would be less of a 'physical' than a social fear, which would still require the ability to experience second-order emotions.

hanging its head and looking sorrowfully towards its owner, is in fact experiencing shame, but this may be nothing more than the dog experiencing a form of sorrow, which, according to Alexander Faulkner Shand (1914/2008), is a primary emotion. The idea that a synthetic consciousness would be capable of experiencing these higher order emotions would have fairly far reaching consequences for our moral attitudes towards them. Second-order emotions could, at least according to Zeman's conditions, promote animals and other non-human consciousnesses from being subjects worthy of moral consideration, to moral agents; beings who can, in some sense at least, be praised or blamed for their actions and intentions.

A second consequence of having second order emotions, if we accept Zeman's criteria for the possession of such emotions, is that any being in possession of them would by definition have some sense of self, or self-awareness. Many commentators, including Derbyshire (1992), believe that self-awareness is a crucial element in the attribution of consciousness to any non-human animal (which they take to be impossible since animals lack such awareness). Whether or not non-human animals do possess the necessary level of self-awareness to facilitate second-order emotions is a matter of some debate; but there is no doubting the moral implications that would ensue should it be proven that any consciousness other than humans do in fact experience them.

Even if second-order emotions cannot be attributed to non-humans, we are still left with emotions of the first-order. So how significant are the so called first-order emotions in the determination of moral consideration? Only from our subjective human perspective can we attach meaning to any emotion - whether it is first or second order - experienced by a member of another species. As with the moral consideration attached to an animal in physical pain, we can only reflect on how such experiences affect us. When a typically active and playful animal suddenly becomes lethargic, we automatically attribute to it either sickness or an unpleasant emotional state. Similarly, when a normally placid animal suddenly becomes aggressive, we go through a similar process; either the animal has become sick or, for whatever reason, something has angered or upset it. In the first instance, we may spontaneously wonder what had happened to it to have caused its aggression, but if prompted for the possible causes of its behaviour we would invariably refer to the things that would trigger aggression in us. Therefore, at the very least, these first-order emotions are morally significant insofar as they appear to reflect the level of pain or suffering being experienced by the agent. If the agent exhibits behaviour that cannot be attributed to humans, for example, a dog wagging its tail, then we arrive at the conclusion that the dog is having a pleasurable experience because it wags its tail whenever it is subjected to other things that we, as humans, would consider pleasurable. From then on, any time a dog wags its tail we take this to be a manifestation of a pleasurable emotion.

First-order emotions**,** therefore, are a fairly significant aspect of assigning moral consideration to members of other species since we take emotions to be either pleasant or painful, whether the pleasure or pain is of a physical or non-physical nature. If a synthetically created being were able to experience distress, fear or sadness**,** then we would be inclined to internalise this, recognise them as painful or unpleasant and this, in most cases, would trigger a moral obligation to alleviate these feelings. Of course, we would have a similar moral obligation, by way of the same process, should it be shown that the being can experience negative second-order emotions. Emotions then, as states of being which can either be painful or pleasant, play a substantial part in the determining of moral consideration.

(vi) Conclusions

Based on Baars' conscious/subconscious distinction, and the fact that epiphenomenalism commits us to determinism, I will maintain that conscious mental activity does have a functional, causal role and so has played a part in our evolution. If there is no such distinction then there is a strong case for arguing that we, as a species, would not have reached this level of complexity in terms of our organic structure, our active engagement with the world, and our social and cultural sophistication. There would be no discernible difference between the efficiency with which we perform actions while consciously focusing our attention upon them, and those performed subconsciously; an idea that is demonstrably false. This is an important point because, although it does not identify the specific causal or functional role that consciousness plays, it does suggest that consciousness has played an important part in our evolution and the evolution of other animals. It could also, even in some small way, add to our justification for our endeavours to create a synthetic consciousness since it allows us to make claims about the performance of a machine being enhanced by this quality, which is, although still not a particularly morally compelling one, a better reason than creating consciousness merely for the sake of it.

From a moral perspective, consciousness is an important quality since we, as a species, generally have a strong tendency to show higher moral regard for organisms and agents which we perceive to be acting in a way which reminds us of ourselves[17]. Further, many of our moral views centre on how our actions and words affect others, and in order to experience the associated feelings and emotions, it is necessary for an agent to be conscious. It stands to reason then, that consciousness, and hence a capacity to experience feelings and emotions, does have a major impact on the moral consideration we afford any human or non-human agent. For those who wish to protect the wellbeing of animals or even of inorganic, synthetically conscious agents, the attribution of consciousness is of the utmost importance**;** without it, the screams, howls and cries of animals in apparent agony are of no more moral significance than the piercing wail of the dented car's security system.

4. Species Chauvinism: The Necessity and Impossibility of a Moral Code.

There are two main reasons for why we should give ethical consideration to the construction of a synthetic consciousness. Firstly, we must protect ourselves from potentially malicious conscious creations, which could seek to harm or even destroy us. Secondly, we must ensure that our treatment of these beings does not result in the infliction of unnecessary pain or suffering. Although I am more concerned with the moral justification, or lack thereof, of bringing such beings into existence, it is also important to look at the ways in which we could be harmed in order to ensure that we do not partake in similar acts against our conscious creations. The reasons for attempting to give machines a moral code, from an anthropocentric position, are fairly obvious: without such a code, we could be placing ourselves in grave danger. Firstly however, I will briefly examine some of the reasons

[17] This is a claim about the human species in general, while bearing in mind that there are those among us who do prefer animals to people.

members of the human race give for placing themselves on a plateau, well above all other species.

(i) *The Human/Non-Human Distinction*

The drawing out of the distinction between humans and non-humans is central to the debate surrounding our treatment of other beings. If for example, monkeys used in medical or cosmetic experiments have the capacity for experiencing the negative effects associated with pain, both on a physical and mental or conscious level, then where is the moral justification in experimenting on these animals and not on members of our own species? Why, if it is shown that they are capable of experiencing both the physical and emotional effects of being subjected to pain, is it morally justifiable to conduct painful experiments on rabbits or rats, when similar treatment of humans would provoke a moral outcry? Further, even if subjecting animals to pain and suffering is proven to be unethical and iniquitous, why do so many believe that the instantaneous and painless killing of a being for food, clothing or convenience is morally permissible, but the idea of this practice being mirrored with the use of human subjects appals the biggest majority of us? This last question is not easily answered, particularly when we consider that if humans did partake in cannibalism, for example, they would not be alone in the animal kingdom, since many rats, primates and fish have been known to eat members of their own species, including their own young. It is not just the repulsion we feel at the thought of consuming members of our own species that prevents us from doing so, because even the thought of others partaking in acts of cannibalism seems to spark a sense of moral impropriety in us.[18] Further, the idea of human flesh being fed to animals is met with no less, and possibly more, repugnance.

So long as we accept that animals are capable of experiencing physical and psychological pain, then there is no need for a synthetically created consciousness to reach ‘human level’ before it becomes morally considerable. However, those not of a vegetarian persuasion can still hold that, since it is morally permissible to bring an animal consciousness to a premature end, there would be nothing ethically aberrant in doing the same to a conscious machine, given that it is not human. But, as with the slaughtering of animals, it is not enough to say the termination of a conscious machine does not constitute a moral transgression because it is not human; such a claim needs to be justified..

One attempt at finding this justification is through an appeal to the human intellect. However, it then becomes a question of why intellect is placed on such a pedestal. Even if a being does not possess our powers of reasoning or complex thought, why does our ability to think logically and reason about, for example, the pain we are in make our suffering any more of a negative experience than a non-human’s? If we accept intellect as our criterion, then there appears to be nothing preventing us from killing and eating, or subjecting to painful medical experiments, humans who are incapable of such reasoning, inevitably this would include babies. If any non-human conscious agent has the capacity to suffer in either a physical or psychological manner then their capacity to mentally analyse their predicament seems largely irrelevant. Indeed, it is not even certain that, under extreme and excruciating pain, we are capable of such analysis. A prisoner of war who is subjected to the most terrible physical torture is unlikely to stop to think about his predicament; all that is experienced is

[18] Cannibalism is used as the ultimate and most severe form of revenge in the story of Prokne and Philomela, where Procne, as revenge for Tereus raping her sister, serves him his own son baked in a pie. In Shakespeare’s *Titus Andronicus*, Titus takes revenge on Tamora in a similar manner after she has his sons framed, and beheaded, for murder.

the all-consuming pain that seems to require no analysing, thinking or reflection whatever. It is difficult to see why, then, such emphasis is placed on the intellect, though perhaps the simplest explanation is that it's a throwback to Aristotle's elevation of reason in man as its distinguishing criterion, and Descartes' elevation of language**.**

Another attempt at justification makes an appeal to the complexity and richness of the human existence. We, the argument goes, are capable of reading books, creating and appreciating works of art and music, of experiencing and expressing an almost infinite range of first-and second-order emotions, taking pleasure in culinary delights, of forming very complex relationships, of making complex plans for the future and many other things that enrich and enhance our lives. Animals, on the other hand, do not experience such enrichment and so their lives are of lesser value[19]. Again, however, this fails to explain why we feel morally justified in killing and eating animals but feel morally repulsed at the thought of doing the same to humans. Firstly, the things mentioned above rely on the fact that humans possess a greater intellect, which I have already argued does not provide sufficient justification for killing. Secondly, much of the pleasure we take from things is purely subjective. How can we be certain that a chimpanzee does not derive the same degree of pleasure and enrichment from gazing at the horizon, cracking open and eating nuts and swinging from trees that humans do from all of their cultural, social, artistic and culinary activities?[20] With regard to a man-made consciousness, there is no contradiction in suggesting that such a being may derive endless pleasure from activities other than those typically enjoyed by humans. If any being is capable of experiencing physical and mental distress, the complexity and richness of its existence are entirely irrelevant, just as they would be for humans who led lives bereft of enrichment. Ultimately, proponents of animal testing, non-vegetarians and anyone who wishes to declare that a synthetically conscious life is less valuable than a human's, must work very hard to identify the deciding and definitive criteria which places our species above all others.

(ii) Self-Preservation and the Need for a Moral Code

At present, there exist machines which, in certain tasks, can outperform even the most gifted and capable of humans. Machines have been designed with massive memories and the ability to retain and process vast amounts of information, some can calculate the most complex of equations in a matter of seconds, and others have even reached the status of 'grand master' in the chess world, apparently 'out-thinking' the best our species has to offer. As time goes by and technology advances there is a very good chance that any manmade consciousness will reside in a highly efficient technologically advanced machine, physiologically superior to us in every way. Potentially - if we allow our imaginations to roam unbridled for a moment - we could create beings who are not only vastly superior to us in every physical way, but who are infinitely more intelligent. If we then grant such beings consciousness, we also endow them with the potential for developing beliefs and desires which could lead them to act in ways that are detrimental to our own species, and in the absence of any moral code set in

[19] For Heidegger the claim that animals are poor in their world (that the stone is worldless and man is world-forming) is an important theme in The Fundamental Concepts of Metaphysics: World, Finitude, Solitude [translated 1995 by McNeill & Walker, Indiana University Press, Bloomington and Indianapolis – see p.185 for the claim "the stone is worldless, the animal is poor in world, man is world-forming".]

[20] In fact, having met some very cheerlessly dour, yet highly intelligent people in my time, it is not clear that the chimp's life is not more enriched by its fairly humble pleasures than some humans are by theirs.

place to constrain these desires, we will have created the potential for disaster. Storrs Hall (2001) puts it more succinctly by remarking that ‘building superhuman sociopaths is a blatantly stupid thing to do’. Indeed he’s right, the creation of beings who are far superior to us both physically and intellectually, and who have the potential for developing malevolent aims, would be ‘blatantly stupid’, and so it is essential for the preservation of our own way of life and our species that we do not allow such creations to come into existence. But any attempt at such prevention is burdened by its own difficulties.

Could we ever be morally justified in creating a conscious, thinking being and then denying it the right to make its own judgements, form its own desires and make its own mistakes? Certainly, we pass laws and make rules which prevent humans from acting upon any desires or judgements which may result in the harming, oppressing or killing of others; very few would deny that we are, in general, morally justified in doing so. However, there is a considerable difference in preventing someone from acting upon their malicious desires because we wish to protect others and altering them in such a way as to render them incapable of ever experiencing these desires. This is, incidentally, before we even look to the agreed conditions for such filtering[21], conditions that would almost certainly vary depending on, among other things, religious beliefs, personal preferences and the society into which it was being implemented.

There is little doubt that we can make a strong moral claim in preventing organisations such as the Ku Klux Klan (see Chalmers 1987, MacLean 1995, Wade, 1998) partaking in behaviour which might satiate their racist desires, but it is not so obvious that we could find equal justification in preventing these desires from arising, even if such prevention were implemented at a genetic level or at the embryonic stage of development. Similarly, if we are to create a sentient, intelligent and conscious being, it is not necessarily the case that we would be morally justified in constructing a barrier which prohibits desires that we humans, in general, find morally abhorrent. In any case, after the inception of these superhuman beings, it is perfectly conceivable that they will eventually form the desire to procreate in the absence of any human intervention, and they may simply reject the idea that it is ethically permissible to disallow a conscious, thinking being from forming its own conscience, judgements and desires.

This last concern has a related point, which should give us reason for reflection. By creating super intelligent beings, we are potentially leaving ourselves exposed to the possibility that such beings would be capable of uncovering loopholes in even the most stringent and rigorous of moral codes assembled by us lesser beings. How could we ever be sure that any moral code is watertight against the wiles of hyper intelligent, conscious machines with the ability to process millions of alternative courses of action per second? Further, there would be little purpose in any appeal to the moral impropriety of attempting to wriggle free from any moral code, since many members of our own humble species earn their daily bread from doing exactly this within our legal system. Of course, Rule 1 of the moral code could simply be ‘never break the moral code’, but this gives rise to further problems which will be examined later [See Section (iv)].

There is also the difficulty in preventing the manufacturing of these beings by those with less than honourable intentions. An enormous negative aspect of technological advancement is the possibility of its misuse; we do not have to look far for examples to illustrate this point. Most

[21] By ‘filtering’ I mean the allowing or disallowing of an agent to possess certain desires.

recently, the internet has given paedophiles and child molesters direct and easy access to children through chat rooms and social networks. There is an almost endless list of the more malevolent members of our species employing technology to accomplish their vicious-goals, including the misuse of gunpowder, manned flight, nuclear energy and telecommunications. There is simply no guarantee that the technology and know-how to construct powerful, super beings would not fall into the hands of people of this ilk and with potentially catastrophic consequences. It is almost impossible to imagine the ramifications that would ensue in the event of this technology being employed by, for example, neo-Nazis or, for that matter, any organisation which desires the obliteration of any race of people or of any particular nation. Before proceeding with any attempt to create conscious, intelligent beings, it is crucial to our survival that we do not allow the realisation of the above scenario, but this, I fear, may be an insuperable task. It may be argued that the wicked intent of insidious groups or individuals has never before stunted our striving for progress and, in fact, it is this very thing that has advanced us as a species and allowed our technology to move forward, yet we ought to bear in mind that we have never before attempted to create something which superseded its creators so decisively in every possible way.

Before continuing, it is important to note at this juncture that the discussion above is looking at the doomsday scenario and might be, it could be argued, skirting on the verge of the fantastic. Selmer Bringsjord (2006), in a response to Billy Joy's (2000) ominously pessimistic '*Why the Future Doesn't Need Us'*, explains why he believes any discussion involving conscious machines with malevolent intentions blasts us into the realm of fantasy:

> …Turing in 1950 predicted with great confidence that by the year 2000 his test would be passed by our computing machines…five years into the new millennium a moderately sharp toddler can outthink the smartest of our machines.

Bringsjord certainly has a point. For all the technological advances that have been made in the last fifty or sixty years, including laptop computers more powerful than those that sent Neil Armstrong and his colleagues to the moon, we have yet to create a machine which possesses all of these characteristics[22] necessary for any claims about it being conscious. However, this is not to say that in 100, 1000 or even 10,000 years the secret to creating consciousness in powerful non-biological entities will not be uncovered and it can only serve our interests to ensure that we prepare ourselves for all conceivable eventualities, even if it means boldly going into the realm of science fiction occasionally.

Essentially, it is in our interest to ensure that, where possible, conscious machines are regulated by a code of ethics for the same reason that human societies have - albeit varying -

[22] There is considerable debate surrounding the necessary criteria for consciousness, however, Stuart (2007) suggests that an agent should: situated and embodied, have multiple goals, exist in an environment which allows complex responses, possess a rich sensory interface which facilitates an inner representation of its world and the ability to interact with its environment in ways which bring about significant changes. Igor Aleksander and Barry Dunmall (2003) claim that the agent should '*sense its environment, have a purpose, plan according to that purpose and then choose to act purposefully*'. They define consciousness as '*having a private sense of an "out there" world, of a self, of contemplative planning and of the determination of whether, when and how to act*.' They detail five axioms for consciousness which are: depiction, imagination, attention, planning and emotion, which is '*closely linked to the perception-planning-action link.*' What is perhaps furthest from realisation, is a machine-centred phenomenology – a synthetic phenomenology is still synthetic (see Chrisley, 2009, pp53-70).

laws, rules and regulations for protection against harm. In the absence of such a code, there is nothing to hinder our conscious creations acting in ways which are both contrary and dangerous to our own way of life, and our very existence, particularly if they evolve to surpass us in every aspect of intelligence, efficiency and physical performance.

Apart from the difficulty in the generation of an infallible moral code, a difficulty which is examined below, there are a number of other problems which should be addressed; problems which we ignore at our peril. Not least is the possibility of Storrs Hall's (2001) 'superhuman sociopaths' and it seems futile to even look for arguments that might counter this assertion. However, there does seem to be a question over the justification for implementing some type of 'moral filter', which stops malevolent desires surfacing. Apart from the debatable moral justification in doing this, there is another consideration: the accepted type and level of filtering would vary from society to society, group to group or even from family to family. Further, even if we could reach a consensus on the implementation of such filtering, there is no way of guaranteeing the infallibility of any moral code when interpreted by supremely intelligent beings with their own social and moral agenda. Even if solutions to these challenges can be found, there appears to be no practical way to prevent active, conscious, thinking beings from being manipulated and deployed as a means to malicious and pernicious human ends. These solutions, designed to protect us from our superhuman creations, may never be uncovered, especially when it is so apparent that we are nowhere near the point of having a universally accepted code which protects us even from our fellow man.

(iii) Anthropocentricism and The Inevitability of Inequality: Protecting them from Us

Nagel (1974) assumes that there is something it is like to be a bat or, for that matter, any human or non-human animal; McGinn (1989) argues that we are unable to understand our own consciousness since we are trapped conceptually inside it and unable to look at it in a way that allows us useful analysis. For similar reasons, it is impossible for us to remove ourselves from our subjective human standpoint and attribute to non-human beings anything other than traits, experiences and characteristics that we ascribe to ourselves and other humans. Take, as an example, the 'dolphin friendly' tuna mentioned earlier and the reasons for the feeling of moral superiority felt by those who buy it. From where, exactly, does such a feeling arise? Why do people feel that the killing of the tuna pales into insignificance when compared to the welfare of the dolphins? Would equivalent moral satisfaction be felt at the sale of 'shark friendly' dolphin meat? It seems unlikely – at least if we avoid straying into the realms of the absolutist[23] views – but why not?

The answer lies in the fact that, as well as being mammals, dolphins, more than either the shark or the tuna, have traits and characteristics that correspond to those that we identify as 'human'. Dolphins appear to have a higher intelligence than either of the other two creatures; they display characteristics that seem very similar to human emotions; they are very friendly and social animals, who are protective of their young and each other, again behaviour we identify as distinctly human. There are even records of dolphins partaking in apparently altruistic behaviour, guiding drowning humans safely to shore and providing protection from sharks:

[23] The view that the value of any living organism's life is inferior to that of any other (see, for example, Taylor (1986).

> In 2004, a group of swimmers were confronted by a ten-foot great white shark off the northern coast of New Zealand. A pod of dolphins "herded" them together, circling them until the great white fled. There are several other examples from the area of Australia of similar incidences.
>
> (*www.dolphins-world.com*)

A shark on the other hand is seen as predatory, ruthless, aggressive, a killer – all traits viewed, at best, as the worst of human nature and, at worst, non-human.

High intelligence, altruistic tendencies, sociability and protection of the young are all considered traits that are most prominent in humans and a dolphin's ability to exhibit similar traits places it in closer proximity to humans than either the shark or the tuna. It is this proximity and for these reasons that we afford the dolphin a higher degree of moral consideration than either of the other two.[24] A similar process takes place with all moral convictions regarding non-human organisms and, in certain circumstances, other humans. Take for example the following extract, from www.factoryfarming.org:-

> Can you conceive the mentality that looked at restlessly strutting creatures such as chickens… and decided to cram them five to a wire cage no bigger than a microwave oven? Then they piled thousands of cages one on top of another…so many that their bones break involuntarily from osteoporosis, the calcium leached to provide egg shells…they're cruelly gassed with CO2 or crushed to death…the fate of those chickens selected to provide meat is little better. As many as 50,000 or more are crammed into a single shed to stand in their own excreta for the six weeks of their obscenely short lives. Huge, waddling babies, forced to grow unnaturally fast - so fast that their hearts can't cope and many die. Legs give way and break under their ballooning weight… What sane person would look at highly-intelligent animals such as pigs and force them into crowded, concrete cells? No bedding, no enrichment, filth and squalor and absolutely nothing to do - unable to fulfil even their most basic natural instincts. And as a bonus, cut off their tails and crush their teeth without anaesthetic in an attempt to control the resulting aggression.

The extract above is designed to provoke an emotional reaction within and from the reader; its objective is to draw sympathy and move the reader to action. But how else can we arrive at a moral decision other than by appealing to our own subjective experiences of the circumstances described above? Assuming that animals are conscious and can experience pain and suffering, we can only sympathise – if we sympathise at all – through taking the described experiences, imagining them, and reflecting on how painful or unpleasant they were, or would be, for *us*.

Equally, the emotional experiences of animals due to some of the conditions described above, such as a lack of mental stimulation, separation from family, living in squalor, or loss of

[24] One point to note however, is that intelligence may be less of a consideration than some of the other characteristics mentioned (and possibly even some that have not). An octopus, for example, is considered to be one of the most intelligent non-human creatures on earth (see http://www.slate.com/id/2192211/ for more on this) and yet is widely available on restaurant menus throughout the world, whereas a rigorous search would be required to find cooked dolphin.

freedom are things that we can only understand as experiences that *we* have gone through or imagine *we could* go through. As with the physical aspect of the apparent suffering, we only have the benefit of our own experiences or the imaginative felt anticipation, and must refer to times when *we* were parted from loved ones, when *we* suffered from boredom or lacked any mental stimulation, or when *we* felt trapped or imprisoned. Indeed, this could also be said for our sympathies towards other humans since it still seems true to say that we reflect on our own subjective experiences to form moral or pragmatic judgements regarding the other person's predicament. How else could I sympathise with somebody's migraine if I had never experienced pain of my own? It could be argued that my sympathy would arise from seeing the other person's sadness or discomfort but, again, how could I sympathise to any great degree with this if I had never experienced sadness or discomfort? Humphrey hints towards something similar in his paper, 'Nature's Psychologists':

> I believe it could be shown that members of a society who have, for example, been put through a brutal initiation ceremony make better introspective psychologists than others who lack the experience.
>
> (Humphrey, 1977, p 68)

Humphrey argues that in order to make good use of introspection, which allows us a greater insight into the behaviour of our fellow men, we need a 'broad range of inner experiences.' In other words, we understand the predicaments of others by looking inwards and making a judgement about how their experience would affect us if it were we who were having it directly. There is no reason to suppose that a similar process is not present when ascertaining the level of pain or suffering that animals are subjected to. A similar point is made by Alvin Goldman:

> Our default procedure is to mindread in a fundamentally biased, egocentric fashion. We project our own conceptual, combinatorial, and ontological dispositions onto others… Moreover, they are initially projected not only onto people but also onto animals or anything else we mindread.
>
> (Goldman, 2008, pp.177-9)

As with holding dolphins in higher moral regard than sharks or tuna, we reflect upon and then project our own experiences – real or imagined - traits and characteristics onto any given situation before arriving at a moral decision. Of course, by adopting an extreme stance and denying any conscious experience to non-human organisms, no moral decision is left to make. But allowing that non-human consciousness, and so pain and suffering, are possible, our moral consideration can only arise through the reflection and projection of our own subjective, human experiences and in this case we are more inclined to extrapolate conscious states to dolphins.

A further hypothetical experiment[25] may throw up an interesting moral question regarding our moral attitudes to other species. Imagine a gorilla – we'll call him Rocky - who has reached the threshold of human intelligence. Rocky has, among other things, become very adept at sign language, can recognize and use certain words to form fairly complex sentences

[25] I only use this for what Daniel Dennett would describe as an 'intuition pump', a thought experiment designed to elicit intuitive responses, I offer no solid conclusions, or 'right' or 'wrong' answers.

linking his inner states to his outer world, has proven himself to have an excellent memory, can solve fairly complex and novel puzzles and is even able to communicate a vast range of emotions, using the words and sentences he has accumulated and which he can produce in novel ways. Further, Rocky is very protective of both his human and primate friends and, on occasion, has even been seen to chastise the other gorillas for acting in an inappropriate manner towards the humans. In the enclosure next to Rocky is Mickey, a feral child similar in nature to Genie, the child described by Michael Newton (see also Rymer 1994 and Candland 1993):

> She could only make strange sounds in her throat; language was beyond her…She had fallen into the pit of the other than human…she spat continually…she would just store the food in her mouth, waiting for the saliva to break it down, often spitting the unmasticated goo onto the plate or her table…she took people's things willfully, pulling on their clothes, invading their space.
>
> (Newton, 2002, p 214)

Mickey then, has failed to reach the level of consciousness that Rocky seems to have attained and does not exhibit behaviour as close to the human behaviour displayed by the domesticated gorilla. In a third enclosure, there is a wild, untrained gorilla, Apollo, who shows none of the abilities or characteristics shown by Rocky or any of the human physical attributes (e.g. being hairless, facial characteristics, proportionately long arms etc.) possessed by Mickey. One day it is discovered that each has contracted a type of fatal disease that affects both humans and primates, and can only be cured using an extremely rare antidote. Further, it is discovered that there is only enough of the potion to cure one of the enclosure's inhabitants and so a decision must be made to save either: the highly intelligent and sociable gorilla, his wild and untamed conspecific, or the feral, unsocial child. Again, for those who deny the possibility of non-human consciousness there is no dilemma; the outcome would not be in doubt since the death of an animal would be no different to the destruction of any inanimate object.

A strong case for the saving of Apollo could really only be made via some sort of species protection argument. If the species to which Apollo belongs is an endangered one, then Apollo seems a more obvious choice than Rocky since Apollo is still wild, untamed and more suited to release back into the wild where he can help repopulate the species. It seems that without this added premise, only an animal rights absolutist would argue for the saving of Apollo. Putting a case for Rocky as opposed to Apollo seems to be based on the attributing of human traits and characteristics, since there are no other discernible differences between the two primates.

However, I contend that the general moral consensus would to save the child, despite the fact that we would describe Rocky as characteristically 'more human' and despite it seeming easier to imagine ourselves as a human consciousness in a non-human body than a non-human consciousness in a human body, including our own. For although Nagel (1974) was right in his assertion that I can never truly know what it's like to be another animal, I can imagine the various alterations in my perception of the world that would occur by having my consciousness relocated to a non-human's body. For example, I, to some extent at least, can imagine things such as only seeing things in black and white or with altered dimensions, having a heightened sense of smell or sound, being able to run at great speeds, having the ability to view the world from the air and even making use of things like echolocation to find my way in the world[26]. Basically, each of these things involves imagining another body with 'me' inside. On the other hand, I find it more difficult to imagine myself without my level of understanding, my capacity for thought, feelings and emotions, my ability to form detailed memories, my beliefs, my hopes and fears for the future, my concerns and my inner self in general, irrespective of the physical form I attempt to hypothetically inhabit.

We now seem to have a contradiction. On one hand it seems that we are more able to identify with the domesticated and communicative gorilla and yet would still opt to save the child whose perspective of the world we find it impossible to internalise. In general, we, as a species have a natural tendency towards conspecific bias. It may be argued that such a conclusion leads us down a path towards what Ryder (2001) and Singer (2002) would call '*speciesism*', where all human interests are placed before those of animals. I do not believe**,** however, that we are necessarily forced down such a path. I do concede that there is a small element of what could be described as species chauvinism, but no more than exists among other groups of conspecifics. Would a dolphin, for example, be more likely to save a human being from a predator at the expense of one of its own?[27] There will always be a tendency to favour members of our own species, but this is not the same as putting every minor human whim above the most needs of non-humans. A similar argument can be made for the creation of a synthetic consciousness.

At the risk of stating the obvious, if and when we create a conscious, thinking being, we will be left with a conscious thinking being. This is not the same as creating a very clever non-conscious computer with the capacity to act like, or give the impression of being a feeling, thinking conscious agent which can, without any moral concern, be switched off at the end of every working day. This point may be obvious, but it is by no means insignificant. As soon as we have created something which has the capacity for a conscious experience of the world, we are morally bound to ensure that it is maintained in a manner which protects it from pain, suffering or harm. If a being is capable of experiencing unpleasant or painful sensations, then we should do all that we reasonably can to ensure that no unnecessary harm should befall it. This moral duty extends to any consciousness regardless of its mode of existence, its origins, or the form of embodiment in which it finds itself. Whether we create super intelligent, physiologically advanced robots, talking and thinking automobiles, or a consciousness which exists only in a box connected to a mains socket in the wall, we are morally obliged to ensure, at the very least, that no unnecessary pain or suffering befalls our creation. However, it is important that we examine some commonly held assumptions that we have about things we create.

There is a tendency to assume that our creations and inventions belong to us and that they are ours to do with as we please, even to the point of destroying them when they are no longer useful to us or begin to become a burden. The only creation that we bring into existence which is not subject to these assumptions is that which results from natural reproduction: the creation of another human being.

[26] There have been instances of such things in humans, including the case of Ben Underwood, a teenage boy who lost his eyes at the age of three and has developed the ability to use echolocation to navigate his way around the world (http://www.youtube.com/results?search_query=echolocation+boy&search_type=&aq=1&oq=echoloc)

[27] This, as with the previous example involving the gorillas, has the status of an 'intuition pump' and any experiment aimed at finding conclusive proof would be impractical, not to mention dangerous.

When a child is brought into the world, although the parents may refer to him or her as ‘our son’ or ‘our daughter’, this in no way suggests that the baby belongs to the parents in the same way the family car or the garden shed does (see Montgomery 1988 and Feser 2004). And it certainly does not suggest that the parents can do as they please with the baby. Given that babies are made entirely from things donated by their parents, and despite the fact that the parents are held responsible for the child’s misdeeds, what else distinguishes them so decisively from our other creations? The difference is that the child is conscious and the artifacts we create are not, and although the infant may not have a fully developed conscious mind, a baby will have the capacity for experiences such as tiredness, restlessness, discomfort, hunger, thirst and physical pain. And even at the very beginning of its life, when its beliefs, desires, self-awareness, emotions, physical interactions and its understanding of the world are at their most basic, there is a strong compulsion to protect the baby from any physical or mental distress. Indeed, for most of us, the very thought of deliberately inflicting any such distress on the child would be very disturbing. If we create a human consciousness of even the most basic kind, it is likely that it will, like the baby, have more than phenomenal consciousness. It will have the capacity for primary emotions and rudimentary desires, which we, as its parents, would be morally obliged to satisfy. It would be reprehensible to assume that, because a thinking feeling consciousness was not located in a vehicle resembling human form, that the unpleasantness associated with the failure to satisfy its needs and desires would be lessened in any way.[28] Of course it is true that a robot possessing ‘baby consciousness’ may not have the desire for food or milk in the way a human baby might, but this is only to say that the robot has different desires, and in no way suggests that the failure to satisfy these desires would cause any less distress. However, I have already stated my contention that, we, as humans, have an innate disposition to sympathise more fully with members of our own species than with members of another and we would probably, even if unintentionally, offer far less consideration to a ‘baby conscious’, non-human machine than we would to a human infant.[29]

The consequences of this attitude could be quite severe and not unlike the situation we have at present with non-human animals, although more folk today acknowledge that animals can feel pain and can suffer. What is important here, is that our inability to shed our biased anthropocentric tendency towards favouring beings most resembling ourselves, would lead us to give less consideration to fully conscious non-human machines than we would to other humans, even in cases where the machine has an equal, or even stronger, moral case. This could lead to their mistreatment, their being used merely as tools for human ends, in a similar way to which animals are used for human consumption and were once used in farming to plough and grind grain. The morality of robot servitude will be examined later [see Section 5 (ii)], but for now the important point is that beings with similar wants, needs and desires to our own would lead cheerless and desolate existences if these needs and desires were not met; and this, I maintain, would be very likely if they were not afforded the same consideration given to a conscious being fortunate enough to exist in a form consisting of flesh, blood and bones. Of course, there is always the possibility that we could change the aims of our mission to include the creation of vehicles biologically identical to the human body, thus negating our propensity to look more favourably on members of our own species. However, we already have a method of doing this with negligible costs (at least prior to birth) and which takes only nine months.

[28] I take this assumption to be equally reprehensible when, as is often the case, it is taken with respect to animals.
[29] Unless they were indistinguishable from each other.

Unless we can, as a species, fully liberate ourselves from our subjective, anthropocentric veil – and I contend we cannot – then the pain, suffering and neglect which could ensue provides us with reason enough to halt our quest to create a thinking, feeling consciousness. There is, of course, the possibility that a synthetically created consciousness could be made to look, feel and smell like humans, but even the smallest identifiable, distinguishing feature would drop our anthropocentric and biased veil once more.

Effectively, as Nagel (1974) asserts, humans cannot go beyond their own subjective conscious experiences in imagining how things in the world affect other species. Consequently, we reflect upon and then project our own past experiences onto any given situation to imagine what *our* experiences would be like under the same or similar circumstances. Although this may be viewed as an overly cognitive and reflective account of how we come to empathise with others, and there is the argument that our empathy comes in a more immediate, spontaneous manner (Stuart, 2009), I maintain that our past experiences are essential for the capacity to feel empathy. It is from these 'reflected experiences', whether it be confinement, physical pain, emotional suffering or a sustained lack of mental stimulation, that we form our moral judgments[30] and separate the things that are morally relevant from the things that are not. Maybe however, this is all too reflective and self-conscious. An alternative may be some version of the theory-theory (see Stich & Nichols, 1992 and Stich and Ravenscroft, 1994) where we gain an understanding of others' mental states through a 'folk' psychology', which arises through the experiences and perceptions of our daily lives. If someone goes to the fridge for a drink, I attribute this behavior to the fact that they were thirsty and so had a *desire* to satiate the thirst, and that they *believed* there was something in the fridge which could help them do so. However, even if this is the case, I am only able to theorise in this manner as a result of my own similar, past experiences, i.e., being thirsty, drinking to alleviate the thirst, finding drinks in fridges**,** and so on**.** So whether I attribute mental states in a simulationist manner as postulated by Goldman (2006), or by way of a folk psychology, it is clear that two major aspects of determining moral consideration are: our inability to go beyond our own subjective, human conscious experiences, and the extent to which we can assign human-like traits and characteristics to the non-human being and its predicament. Forming moral judgements in this way is not the result of deliberative practice and overt species chauvinism, since it would be very difficult for us to form such judgements in any other meaningful way. It could**,** however, lead to a great deal of physical and emotional anguish for any inorganic consciousness.

(iv) Moral Facts and The Incomputable Nature of Ethics

Elsewhere I have made reference to the question of whether or not we are morally justified in creating a being with a set of pre-programmed moral rules that would prevent the arising of desires which ran contrary to human interests, but this is only part of the problem. The other part lies in the difficulties we face in selecting a robust and infallible code for moral living. Anyone with even a passing interest in moral philosophy will be well aware of the plethora of moral and ethical theories, some dating back hundreds, or even thousands of years, and each advocating a sometimes subtly different path to leading a morally good life. The one thing these theories share is that each of them is flawed.

[30] For an alternative to this view, see Gallagher (2007) who argues that we're endogenously moral.

In Kantian ethics (1785/2008) the crucial, determining moral factor in any act is motive, acting from a sense of duty as opposed to any felt emotion or sympathy towards our moral subjects. However, two conflicting duties (e.g. ‘never lie’ and ‘always protect young children’) could result in the possibility of an act being deemed moral in spite of having negative and unpalatable consequences. On being asked by a gun-wielding maniac about the whereabouts of a small child he intends to kill, Kantians would be forced to argue that we ought not to lie but to inform the maniac of the child’s hiding place.

Rights theories, postulated by Hobbes (1651), Locke (1690) and Paine (1791) stipulate that we come into existence with certain basic, inalienable rights which are irrevocable by any person, group or government. One problem with such rights is that proponents of these theories owe us an explanation as to how we acquire such rights.[31] A possible answer is that they were given to us by a creator, but the existence of any such creator is by no means certain and, even if it were, we are still owed an explanation as to how we became aware of His or Her divine intentions.

Another difficulty with Natural Rights theories is conflicting rights. Philip Montague offers the following dilemma:

> You and I are neighbours, with our houses situated closely together. You lead a group of rock musicians who can practice only in the evenings in your backyard; while I, on the other hand, enjoy nothing more than quiet evenings spent on my porch accompanied by the sounds of frogs and crickets. Presumably, you have a right to pursue your musical career, and I have a right quietly to enjoy my property. If we do indeed have these rights, however, then they seem to conflict with each other, in that your exercising your right is incompatible with my exercising mine.
>
> (Montague, 2001, http://journals.cambridge.org)

The important point to note here is that occasions can arise where one person’s seemingly ‘inalienable right’ conflicts with another’s, and proponents of a Natural Rights Theory have a difficult time in resolving such conflicts, particularly if both rights have been ‘given by God’.

Utilitarian theorists, including Bentham (1789), Mill (1863) and Singer (1993), also fail to provide an infallible moral theory. According to these theories, the moral worth of an act is determined by how much utility it provides or, in other words, the net pleasure brought to those affected by the act. One major difficulty with Utilitarianism comes from John Rawls (1971) who argues that there is a difficulty in calculating the ‘net pleasure’ or ‘net happiness’. Utilitarianism, according to Rawls, fails to acknowledge each person as a single, distinct and unique conscious agent and, instead treats a group of individuals as a single conscious entity[32] with no variation in preferences, desires, or routes to happiness. A further problem with a Utilitarian account of ethics is that, in contrast to Natural Rights theories, it goes too far the other way and completely disregards the rights of the individual, meaning, for example, that an individual’s life could be justifiably sacrificed to protect the lives of one hundred or even just two others. Finally, as a consequentialist theory looking only at the results of actions, utilitarianism leaves no room in our moral sphere for intentions, meaning that an act of

[31] Locke gives us a long discourse on the matter in The Two Treatise on Government.

[32] Ryder (2001) describes this as the ‘boundary of the individual’.

intended maliciousness which inadvertently brings about good consequences would be considered a moral act.

Essentially, it is this which makes the idea of a programmable ethics so problematic: if we do not yet have an infallible ethical theory of our own, how can we ever be justified in programming artificially conscious machines to live their lives? It is this question that I will examine in the remainder of this chapter.

Moral facts are strange things. On the one hand, in a bid to maintain our freedom of thought and individuality, we often wish to deny their existence, and on the other, there are times when the denial of their existence fills us with moral revulsion. There are occasions when the non-realist approach (see Mackie 1990, Ayer 1990, Harman & Thompson 1995) seems most appropriate, when our moral disagreements with others do not compel us into direct conflict, or to regard them as morally aberrant. There are those, for example, who believe in a person's right to die when faced with the prospect of living with the pain and indignity of a debilitating and incurable disease, and so advocate the legalisation of euthanasia, and there are those who oppose the idea of assisted suicide and argue that, by taking the Hippocratic oath, doctors are swearing only to endeavour to save lives and not, as proponents of euthanasia would have them do, to end them. In this case, although we have a clear moral dispute, it is rare for an individual on one side to hold someone on the other in complete moral contempt and, more often than not, proponents and opponents of euthanasia live side by side in reasonable harmony. In cases such as this, most people have a willingness to allow others to have a point of view and to put the disagreement down to a divergence in moral values, as opposed to an ignorance of any 'moral facts'. Indeed, non-realist theories have a strong appeal here since it allows that I have my moral standpoint, you have yours, nobody is right or wrong, and we can all live in a relatively peaceful and tolerant accord. That is, it is argued, until such theories are put into practice.

Suppose we base our moral code on a non-realist theory where moral facts are no more real than fairy dust; what do we then say of those who advocate rape, paedophilia or genocide as lifestyle choices? Do we really want to say that such people simply have different moral values? Do we content ourselves with the thought that these individuals simply have divergent ethical attitudes? The reality is that we absolutely do not want to allow that such people merely have differing moral attitudes to our own, or that the only reason to deny them the right to act upon these attitudes is the protection of others. Rather, we often feel compelled to go further than this: we want to adopt a moral realist stance (see G.E. Moore, 1903) and say that these people are simply *wrong*. At best, we label them as victims of an ethical aberration or grotesque mental deficiency and, at worst, as iniquitous and soulless fiends unworthy of the right to live in civilised society. We often want to say that it is a *fact* that these individuals are wrong, that it is a *fact* that their moral compass is either damaged or absent**,** and that it is a *fact* that they are not like us**,** and any denial of this places us in direct conflict with the 'moral majority'. The implication here is that even if we feel some discomfort at the idea of living within the confines of ethical absolutes, it is these very truths we reach for when anyone dares to transgress those moral values we ourselves hold dear - and sometimes unreflectingly so.

Nevertheless, in spite of our strong compulsion towards the use of moral facts in our condemnation of the most "deviant" members of our society, in over three thousand years of trying, we have yet to uncover even one provable moral fact.[33] Effectively, we have begun to

[33] One possible exception to this could be 'never cause unnecessary harm', however this, I contend leads to questions regarding the definition of 'necessary' and 'unnecessary', 'cause' and 'harm', and has all the

discuss the implementation of ethical values and moral virtues in powerful, intelligent machines when, because we have failed to reveal a single indubitable moral truth of our own, we don't really know what we're talking about or even whether there can be agreement. We do not even know if such truths exist and, even if they were somehow discovered, it is contestable as to whether or not we would adhere to them after living for so long in a moral world that, in general, permits the acceptance of (some) differing moral attitudes. I doubt that we could say with absolute certainty that, if it could be proven beyond any doubt that abortion is wrong, everybody's moral attitudes would suddenly shift towards the pro-life ethos.

Our attempts at programming a strong, failsafe ethical code have hit a problem. Is there any way then, that we can implement a moral code in the absence of moral facts?

(v) Hedonistic Act Utilitarianism

Storrs Hall (2001) argues that to simply give the machines a set of specific instructions to obey would be erroneous, since if the machines were to exist on a higher intellectual plane than humans, they will surely and quickly find ways in which to exploit any loopholes in these instructions, possibly even to the detriment of humanity. Therefore, he argues, it is far better to create a *'true, valid, universal ethics that will be as valuable to them as it is to us'*.[34] According to Storrs Hall, societies become far more prosperous under the guidance of moral codes since the members of a moral civilisation are more likely to cooperate and work together for the betterment of that particular society. With no ethical constraints on our actions, anarchy is likely to ensue, which would lead to the breakdown of society in general. This is similar to Hobbes' (1651/1985) 'state of nature' where life is 'solitary, poor, nasty, brutish and short' (Leviathan Ch XIII), although it should be noted that Hobbes' does not rely on morality to escape from this stark vision of society, but rather on mutually beneficial contracts between its citizens. Locke (1690/1997), perhaps in a manner closer to Storrs Hall's, attempts to prevent society from lapsing into a state of nature by appealing to reason as that which dictates natural law. Reason, according to Locke, shows us that we should never harm others by way of physical injury, impinging upon their liberty or by relieving them of their possessions Whether or not moral virtues enhance the survival and prosperity of a society is not an open and shut case, it could be argued that many species in the animal kingdom have survived well enough in the absence of any moral code. In any case, let us examine the idea of a '*true, valid, universal ethics*' and what it might be like.

The problems with the standard moral theories and the loopholes identified by us mere humans (described in the preceding section) would most certainly be discovered by conscious, powerful, super intelligent machines with an eye for detail. Any attempt, therefore, to implement any of the conventional moral theories fails to provide us – and the machines – with an infallible moral code, and this, I have already claimed, is crucial for our safety and perhaps even for our survival as a species. Anderson & Anderson (2007) however, suggest an adaptation of Utilitarianism – Hedonistic Act Utilitarianism – as a possibility.

potential, because a necessary harm in one culture might be an unnecessary one in another, to lead us into an inevitable moral relativism.

[34] This however, does not rule out their superior intellect seeing through our 'universal ethic' and discovering its flaws.

Crucially, hedonistic act utilitarianism (HAU) involves calculating which course of action allows the derivation of the greatest amount of net pleasure for each affected individual. The calculation of net pleasure is carried out in the following way:

> Total net pleasure = (intensity x duration x probability) for each affected individual.

Anderson & Anderson suggest a possible scale of 2 to -2 for each part of the equation, although the range appears to be largely irrelevant. They argue that this is a promising starting point for the implementation of a moral code in machines for the following three reasons. Firstly, as humanswe do not, when it comes to arriving at moral conclusions, carry out strict arithmetical or algorithmic processes, and so most of our moral judgements are based on estimates or ‘best guesses’. Machines, on the other hand, would be far more adept at carrying out the kind of calculation set out above, and so would arrive at the morally correct conclusion much more efficiently than we would. Secondly, humans have a propensity for partiality, which allows for the particularizing of our moral judgements by our personal preferences, while machines or robots, in virtue of being devoid of such bias, would be more likely to take a morally objective standpoints. Thirdly, due to a lack of processing power, or a lack of knowledge, or a combination of both, humans cannot, or do not, consider all of the possible consequences from all of the possible actions, something which powerful machines with an enormous database of relevant information could do with ease. Certainly, if morality can lend itself to such a calculation, then there is no doubt that machines or robots, due to their vastly superior processing power, would make far better moral judges than human-beings.

One immediate objection to HAU is its ethical relativism. The equation above relies on the fact that there is only one possible correct action in any particular moral dilemma, which is simply not right, particularly if we, as I do, reject the existence of absolute moral truths. There may, for example, be a case for arguing that in Country X one action is morally appropriate, but in Country Y it would be a different course of action. An example of this may be the variation in the age of consent from one country to another, where in Nigeria it is 13, in the UK 16 and in Yemen, to participate in intercourse, the couple must be married, although the legal marrying age in Yemen is 17.[35] Further, there may, on occasion, arise dilemmas to which there is no easy and clear answer, or where no one answer is better than any of the others; there may be numerous possible courses of action, each with a net pleasure equaling 2, and this leaves us no further forward in our attempt to ascertain the most appropriate manner in which to act. Anderson & Anderson’s response to this is that it may not be possible to find a moral code which allows machines to resolve the more difficult issues and that it is probably more prudent to shield the machines from the thornier moral problems. This may be fine at the most basic level, where machines are not conscious and are only designed for a limited number of purposes, such as doing laundry or cleaning the car, but when discussing conscious machines with a mental capacity which equals, or even surpasses, that of our own, this response is far from satisfactory. By Anderson & Anderson’s own admission, one advantage of creating ethical agents is that they can help us to further our knowledge and understanding of our own moral universe; it seems this benefit diminishes somewhat if we disallow these creations from examining our more challenging ethical conundrums.

[35] Figures taken from http://www.avert.org/age-of-consent.htm

Another problem with HAU is that, like many other ethical theories, there remains the possibility (indeed the probability) of stringent obedience leading to decisions that run contrary to our most strongly held moral beliefs. Bernard Williams (1973) argues that in order for a decision or action to be considered moral, it must allow us to *'preserve our identity and psychological integrity.'* Williams gives the example of a botanist, Jim, who is faced with the dilemma of shooting one captive in order to save the lives of the other nineteen, or walking away and allowing all twenty captives to be killed. From a utilitarian perspective there is no dilemma: one of the captives must be killed. Williams however, argues that such thinking strips us of our humanity and integrity, turning us into un-thinking machines that go through life bringing about consequences which by-pass the decision-making process completely. He argues that there is a crucial moral difference between a person being killed by me and being killed by someone else because of what I do, a difference that the utilitarian fails to recognise. Difficulties such as these are yet another example of a more general problem with the calculation of Bentham's (1789/2009) felicific principle, and we have seen with Rawls' (1971) criticism, regarding the difficulty in aggregating utility, that there is a very real difficulty in ascertaining the level of pleasure or happiness associated with each action, rule or policy.

With respect to utilitarian robots, Chrisopher Grau (2005) postulates the idea of limiting the capacities in robots or conscious machines for the purpose of lessening the chance of an 'enslaved' machine becoming thoroughly dissatisfied with its lot in life. Grau suggests this as a reason for why utilitarian robots should not be created, both from the point of view of humans and the point of view of the conscious, sentient robots. From the robots' perspective, utilitarianism is not a viable option due to what Grau describes as 'the integrity objection', which draws a distinction between acting in accordance with the moral theory and acting in a way that we believe intuitively to be right. Grau uses Bernard Williams' example of a man saving his wife from drowning, rather than a stranger faced with the same fate. In this example, even if utilitarianism can provide justification for the man's action, it still demands that he alienates himself from 'his natural motives and feelings'. Utilitarianism is thus unable to account morally for the man's decision, despite the fact that most of us would agree that he has acted with complete moral propriety. Such dilemmas could leave the machine – or a human, for that matter – torn between its natural allegiance to the things it holds dear and a commitment to living in accordance with the utilitarian moral code. Grau's solution is to withhold certain traits from synthetically conscious machines:

> It may well be immoral of us to create a moral robot and then burden it with a life of projects and commitments that would have to be subsumed under the demands required by impartial utilitarian calculation. This leads me to the more general question of whether we may be morally obliged to limit the capacities of robots.[36]

Such traits may include sentience, consciousness and meaningful commitments that may *'conflict with the demands of morality'*. The second reason given by Grau for not creating utilitarian robots concerns the possible consequences for humans and runs along the lines of a standard objection to utilitarianism: that a utilitarian machine could still, with moral impunity, act unjustly by violating the rights of a few in order to maximise utility for the many.

[36] http://www.aaai.org/Papers/Symposia/Fall/2005/FS-05-06/FS05-06-007.pdf

Even a brief examination indicates that there is more to morality than mere calculation. We arrive at our moral judgements and our 'correct' courses of action, via our ability to sympathise and empathise with those affected by our actions, even if we are unable to label the rightness or wrongness of our actions as 'fact'. It is true, as Anderson & Anderson claim, that humans do find difficulty in making moral judgements from a point of impartiality, but there is no reason to suppose that this is an entirely negative aspect of our moral constitution. What better way of regulating our behaviour and actions towards others than the ability to understand the negative emotions associated with their mistreatment, or comprehending the joy and pleasure experienced by those we treat well? We feel compelled to act on behalf of the black man facing death because we know what it is to be scared or threatened and this, in turn, prompts us to want to alleviate these feelings in him. It is going too far to define our most strongly held moral convictions as 'facts'- although there is the argument that by permitting everyone the right to their own moral values, regardless of their perceived depravity, we go too far the other way - but our emotions and our capacity for sympathy, whether we believe that they arise from, as of yet, unattained moral truths, or from subjective moral attitudes play a significant role in guiding us along whichever moral path we choose.

Damasio (2005) argues that moral judgements, indeed all judgements, are emotion-based and emotions, rather than being consequences of judgements, are in fact bodily responses[37] resulting from a change in environment or predicament. This contention is partly based on results from research carried out by Damasio and others from the universities of Southern California, Harvard and Iowa[38], which showed that by 'shutting off' their emotions[39], people are more likely to adopt a 'colder' approach in their moral decision-making. Although not entirely bereft of the ability to solve moral problems, those suffering from ventromedial prefrontal cortex (VMPC) damage showed a greater propensity and willingness to harm an individual for 'the greater good'; a utilitarian motive. Here then, we have possible evidence of emotions being a significant factor in our moral judgements. Calculations however, deal only in facts and since, as of yet, we have no such things in our moral realm, we have no place for numbers, formulas or equations in this realm either. Further, like the situation described with the lynch mob, it is inconceivable that a program could ever be designed which allowed a machine to be in possession of all the relevant facts pertaining to any given circumstances. Without *all* of the facts, the machine would only be making 'best guesses', for which the HAU machines would, according to Anderson & Anderson, have been designed to eradicate.

In addition to the difficulty in calculating the net pleasure involved with moral decisions, HAU also suffers from a similar problem to that of Ryder's (2001) '*Painism*', in which the morally correct action is that which results in the least amount of pain, aggregated between all of the affected individuals. The difficulty with painism is that it assumes that an absence of pain equates to an absence of moral transgression, meaning that quietly and quickly killing a forest-dwelling hermit with no friends or family is a morally neutral, if not acceptable, course of action. Similarly, the killing of the hermit does not seem to give a negative rating –

[37] Such a theory was first postulated by William James and Carl Lange, independently of one another, in the 1880s.

[38] For a fuller description of this research see: University of Southern California (2007, March 22). Moral Judgment Fails Without Feelings. ScienceDaily.

[39] This 'blocking off' was a result of several of the subjects having damage to their ventromedial prefrontal cortex (VMPC), an area of the brain located in the prefrontal lobe which plays a part in our ability to make decisions.

or any rating at all for that matter - when placed in the ‘net pleasure’ equation, and so it seems that, according to HAU, no moral transgression has occurred in the hermit’s killing. Equally, if the victim of the lynching has no family or friends and no one ever becomes aware of his death, there is no negative rating other than that for the victim himself, which is minute in comparison to the net pleasure felt by the mob; would this make the lynching any more acceptable? It seems not, since we are inclined to take into account the feelings, hopes, desires and ambitions of anyone who finds themselves in such a predicament. Indeed, even if there is no mob, but only a single fanatical and bloodthirsty supremacist who is intent on carrying out the lynching and whose rating is equal and opposite to that of the victim’s, then we have a moral dilemma, which seems to contradict our attitudes towards morality.

Finally, there is no guarantee that a consciousintelligent machine would, as Anderson & Anderson argue, remain impartial with regard to moral decisions. There is every possibility that in the event of having to decide between saving a number of humans or one of its ‘own kind’, the machine will decide to save its robotic counterpart, even if the morality of doing so is questionable from our perspective. Such a course of action would hardly seem surprising given that humans would be likely to act in a similar manner. We could never be certain that conscious, intelligent machines would not possess a disposition analogous to the anthropocentric standpoint in humans, in which case humans could find themselves well down the hierarchy of morally considerable creatures.

Undoubtedly, if it were ever possible to reduce morality to numbers, formulas and calculations, then conscious intelligent machines would make by far the best moral judges. And certainly, with respect to moral dilemmas involving *humans* and their issues, it would be fair to say that ethical machines seem like the most impartial and objective of moral judges. However, the discovery of a programmable and infallible ethical code, at least at present, is as likely as unearthing the meaning of life itself. This is for two main reasons. Firstly, calculations work with facts, of which there is a definite dearth in our moral sphere or with estimates, which are perfectly acceptable in some fields, but not in conscious, intelligent super-beings with the potential to supersede humans as the dominant species and, potentially, to bring about our downfall. Secondly, moral facts seem dependent on other extenuating facts surrounding the circumstances of any moral dilemma and do not seem sustainable as facts in their own right. In the lynch mob example described above, there could, under certain circumstances, be justifiable reasons for the lynching, or it may be somewhat controversial, which would remain unsolved if any calculation resulted in an equal score for both sides. The likelihood of any agent gaining possession of every mitigating fact and every interest of all affected individuals is improbable and so even the quickest and most capable of machines would still, like us flawed and partial humans, be making ‘best guesses’ or estimates – a thoroughly unsatisfying outcome for proponents of calculable moral decisions.

5. The Killing, Rights and Ethical Treatment of a Synthetic Consciousness

(i) Why is Killing Wrong?

Few of us would condone the random and unnecessary killing of a fellow human being[40] and in fact most of us would consider murder among the most heinous of acts, but perhaps we do not often think of why we think this way. This judgement cannot rest on any pain or suffering felt by the victim so long as the murder is swift and the victim is unaware of his or her impending death since, being dead, the victim feels no physical pain and, so long as they are unaware of the act, no emotional distress. So the wrongness of the act must lie elsewhere. One obvious objection to murder is the pain and suffering endured by the people the victim leaves behind, such as a husband or wife, children, parents and close friends. Although the victim, being dead, cannot experience any negative emotions, the pain and suffering of the friends and relatives would appear to be very great and, since it is wrong to inflict unnecessary pain and suffering, murder is obviously wrong (this is the type of reasoning Ryder seems committed to). Another reason for deeming murder to be wrong could come from rights theorists who argue that we have an inalienable right to life, which should only be brought to an end through natural causes (or perhaps through our own volition in voluntary euthanasia). The argument here then is that the murderer denied his victim of this right and so committed a wrongful act. A third objection may be that the murder was simply a waste of life; that the victim (we will assume he is not on the verge of dying anyway) left behind unfinished projects, a young family he can no longer help raise, and so on and, by cutting this life short, the murderer has prevented the victim seeing through these projects and so has caused a ‘waste of life’. Between them reasons seem to provide very strong moral opposition to the unnecessary killing of a human being, but how many of these can be applied to the case of conscious to show that killing them would also be wrong? I contend that none of them can and that there may even be difficulties in maintaining that they show the killing of humans to be morally wrong.

(a) Pain and Suffering

The objection that referring to the pain and suffering of those the victim leaves behind appears to be a strong one, but even it fails to give a fully satisfactory reason as to why unnecessary killing is wrong. It seems fairly uncontroversial to assert that we should not deliberately inflict unnecessary pain or suffering on others, and by murdering a person we do inflict such pain on their loved ones and so have committed a wrongful act. However, there seems more to our condemnation of murder than merely the pain and suffering felt by the friends and relatives, since this suggests that we would be morally justified in killing a long-time hermit with no friends or family, but who leads a happy, pleasant and enjoyable life out in the wilderness among the trees and forest animals. Here there is apparently no pain or suffering felt by anyone left behind, for there is no-one left behind. But even if a response can be found to this, it is still not obvious that by killing a sheep, a cow, a fox, or a conscious machine we are leaving behind a herd, flock, skulk, or network of distressed or outraged mourners. In order for pain and suffering to be cited as a reason for the wrongness of killing a being, it must be shown that there is a strong and direct negative effect on who or what is left behind since, providing the killing is swift and painless, the slaughtered being itself cannot be said to have suffered.

The machines, by their very nature, would be ‘built from scratch’ and so would not come under the influence of the pain or suffering of leaving mourning relatives behind in the way humans, calves or puppies would. Finally, even if satisfactory answers are found to these

[40] At this point I wish to set to one side any discussions about the rights and wrongs of abortion, euthanasia, just or unjust wars and so on. These are beyond the scope of this thesis.

objections, and the pain and suffering of those left behind is worthy of moral consideration, such objections would still fail to explain why it is morally objectionable to kill beings that are not generally sociable, such as sharks, hermits or any other solitary creature. Of course, there is still an argument to be made as to why it might be less of a moral transgression to kill the shark than it would be to kill the hermit, but this requires that we establish a morally acceptable distinction between humans and non-humans discussed in the previous chapter.

(b) Desire for Continued Existence

One interesting topic that emanates from the having of desires is the question of whether or not non-human beings can have an interest in prolonging of their own lives. For even if it could be shown that non-humans can experience pain, or be subject to suffering, it does not necessarily follow that there is anything morally wrong in the swift and painless ending of such a life. A problem arises when we consider a conscious, feeling, sentient machine with memories and interests but which has outlived its usefulness and/or becomes a financial burden; would there be any moral transgression in simply switching it off? According to Carruthers (1992) for any being to have an interest in the prolonging and continuing of its own life, it would require the ability to focus its attention on its life as well as the capacity to desire its continuation beyond the present moment. In addition, the failure – or anticipation of the failure - of these things coming to pass must cause a negative or painful response in the being. Carruthers explains:

> …a desire for one's own future existence must involve concepts of oneself, the future and of existence. Moreover, possession of any given concept must involve, in addition, possession of its contrasting concepts…if any animal were really capable of conceptualising, and desiring, its own future existence, it would also have to be conceptualising non-existence.
>
> (Carruthers, 1992, p136)

Most humans possess at least a mild apprehension at the prospect of dying[41], but the prospect of being unable to partake in activities from which we derive pleasure plays only a minor part in the forming of this apprehension. I take myself to have a strong interest in the continuation of my own life, but the fact that I will never again get to eat hot chocolate fudge cake, lie on Daytona beach or partake in sexual activities – all things from which I derive immense pleasure – do not seem particularly significant when I question my desire to continue to exist. There seem to be other, stronger factors in play when I examine this desire, such as: a concern for those I would leave behind, the thought of missing out on things such as younger family members growing up, a fear of the unknown, the leaving behind of various unfinished projects, or even the worry that I would not be missed.[42] I have an interest in the continuation of my own life based not only on pleasurable experiences, but on a sense of my life as a whole (which includes past and future experiences), a consideration of other people, and the impact my death would have on them.

For Carruthers, in order for any being to have an interest in the continuation of its life, it must have the ability to focus its attention on its existence and desire that it continue, or the ability to focus its attention on 'non-existence' and a desire not to enter into this state. A wildebeest may witness a member of its herd being caught and killed by a predator. Consequently, the wildebeest may become aware of the fact that due to being caught by the predator, one of its conspecifics is no longer around. This however, is not the same as having a concept of its own mortality or experiencing sadness or fear at the thought of its impending expiration. At the very least, the surviving wildebeest must have some understanding of the fact that its herd-mate is now absent, and must desire not to enter into that same state. For this it must be

[41] The fears of those who believe in an afterlife are assuaged by the thought of continuing to exist in some form and self-help manuals have been written for centuries about this. See, for example, Jeremy Taylor's *The Rule and Exercise of Holy Dying* (1651).

[42] Perhaps asking a convict sitting on the electric chair about his reasons for desiring the perpetuation of his life would provide further proof of this.

capable of two things: it must remember the effect a predatory attack had on its conspecific; and it must also have the capacity to form some concept of the future in order to avoid a similar fate, a fate which it *desires* not to fall into. Otherwise, Carruthers asks, apart from 'pain avoidance', how else could we explain the utmost importance animals place on the avoidance of predators or any desperate attempt to locate food, water and shelter? In effect, having an interest in one's own life amounts to having a desire for continued existence or, put another way, a desire 'not to not exist'. Desire entails beliefs and so one must have a belief in the possibility of non-existence in order to have an interest in the continuation of one's life. Carruthers' argument is however, unsustainable.

Let's take the idea of having a desire 'not to not exist' and the process of autopoiesis which, roughly speaking, involves the constituent parts of a system interacting with each other in a manner which facilitates the sustenance of these parts and the relationships between them. Such a system attempts to maintain its organisation through the process of autopoiesis and can be identified with respect to the *process* as opposed to its constituent parts, which may or may not remain constant over time. Maturana (1980), one of the pioneers of this work, describes it as:

> …a network of processes of production (transformation and destruction) of components that produces the components which: (i) through their interactions and transformations continuously regenerate and realise the network of processes (relations) that produced them; and (ii) constitute it (the machine) as a concrete unity in the space in which they (components) exist by specifying the topological domain of its realisation as such a network.
>
> (www.cs.ucl.ac.uk)

Effectively, although the parts of an autopoietic system may change over time, the system maintains its structure through the dynamic processes of destruction and regeneration, which help to maintain the whole. Biological cells provide an example of this:

> The eukaryotic cell, for example, is made of various biochemical components such as nucleic acids and proteins, and is organized into bounded structures such as the cell nucleus, various organelles, a cell membrane and cytoskeleton. These structures, based on an external flow of molecules and energy, produce the components which, in turn, continue to maintain the organized bounded structure that gives rise to these components.
>
> (Wikipedia.org/wiki/Autopoiesis)

Here we seem to have an example of a system which, although it could not be considered conscious in any of the more commonplace uses of the term, seems to exhibit a deliberate attempt to continue to exist, presumably in the absence of any beliefs about death or nonexistence. Indeed, it seems that such a system concentrates all its efforts on existing, even though it does not appear to have any beliefs or concerns whatsoever about the future which could be said to be driving these efforts. There is then a case for saying that all living, conscious beings from the most basic to the most complex, possess an innate drive towards continued existence, a drive that is independent of any beliefs, desires, concepts, or concerns regarding the avoidance of pain, and with which we have a moral obligation not to interfere. So now we have another good reason to suppose that, in opposition to Carruthers' account, these instincts operate pre-conceptually and that there is no need, contrary to Carruthers' claim, to assert the possession of concepts as necessary for a desire for a continued existence.

Such a moral obligation would still require reasons for why interference in this process is unacceptable. But it may simply be the case that all living, conscious beings possess a similar 'striving' towards continued existence similar to that of the canonical autopoietic system, the biological cell, and any attempt to obstruct this process would cause resistance and, possibly negative or unpleasant sensations in any being complex enough to experience them. Naturally, varying levels of consciousness may have varying beliefs, if any, as well as various desires, and so could be said to have different reasons – if we can call them that – for their desire for survival. Human parents with young children may desire to stay alive not solely for the sake of being alive, but because of the concern they have for the wellbeing of their children and because they want the children to be given the love and care that only they, the parents, can provide. But an inability to form concepts and thoughts such as these seem insufficient for denying the desire for continued existence to other beings.

Another objection to Carruthers involves the example of the visual cliff illusion, created by Gibson and Walk (1960) to test the depth perception of a variety of species. For humans, it was found that at 6 months we are capable of perceiving a sharp drop and so pull away when we believe non-propositionally there to be the possibility of falling. This poses a problem for Carruthers since at 6 months old we are far from being able to conceptualise things in the way he demands of animals in order that they have a desire for continued existence.

Whether or not, and to what extent, a non-human consciousness can have an interest in the continuation of its own life is an important factor in the determination of moral consideration. Consequently, by showing that non-humans can have an interest in their own existence, and not just in the avoidance of pain, it would be difficult to deny the moral transgression involved in the killing (even humanely) of non-human beings. Further, contrary to Carruthers' claims, the ability to conceptualise things such as life, existence and non-existence are not necessary criteria for the possession of a desire to exist.

(c) Rights: Accountability, Ownership and Termination

(1) Rights

Some form of rights theory may be used to argue against killing and would be structured in such a way as to claim that all conscious living beings have an inalienable moral right to life and which should never be compromised or taken away, irrespective of the benefits to humanity. To the extreme form of animal rights, it is tempting to present ridiculously improbable scenarios designed to show the absurdity of their position, such as the moral impropriety in killing a house fly in order to save the entire human race from certain extinction; or from knowingly stepping on a group of beetles on the way into a burning building to save a baby and so on. However, I believe that there are very few**,** Jainists being the exception, who would hold us in moral contempt for killing the non-humans in these circumstances. What I want to examine is the less extreme claim that all conscious beings have the right not to be killed or slaughtered for reasons of human entertainment (hunting, shooting etc.), pleasure (enjoyable dining experiences, fashion) or convenience (robot slavery).

The term 'moral right' is, in itself, a problematic one since such rights can, and invariably do, vary with shifts in our moral attitudes. In most Western societies for example, women are seen as equal to men and are seen to be morally justified in demanding things such as equal pay, equal opportunity for advancement in the workplace, the right to vote**,** and so on. In some middle-eastern countries however, women are forbidden from attaining such things and are seen as subservient to men, even to the point of having to keep their entire bodies covered or remaining a certain distance behind their husbands whilst out in public. Clearly, these two types of societies have distinctly opposing views on the moral rights of women and from these stem legal rights dictating how women can and cannot be treated. Take for example the claim that all humans have a moral right to see out their lives, lives that should only be ended through natural causes or acts of nature. In the United States, there appears to be opposing views on this claim depending on which part of the country we look at. The states of Texas and California, among others, allow capital punishment for the most serious of crimes, whereas Alaska and Michigan maintain that all humans, irrespective of their transgressions, do in fact have a moral right to see out their natural lives without intervention, and this has led to the abolition of capital punishment. There are other issues such as abortion[43] and euthanasia, where one side makes a claim about a moral right and the other side rejects it. There seems then, to be no difference between deciding for oneself (or as a community) the rightness or wrongness of an issue, and the claiming of a moral right. Frey (1983) makes a similar point to this by arguing that talk of moral rights with regard to non-human animals, in particular food animals, takes us no further in the debate:

> …the reason is simply that moral rights get in the way. On the one hand, they make it appear that issues such as our treatment of food animals cannot be discussed or discussed properly unless animals are ceded rights, and this is quite false. Numerous, concerned laymen have argued for years over the morality of our treatment of animals, without the intervention of moral rights…and Britain has long been a leader in reform legislation without its legislators and, through them, its populace having first to cede animals moral rights in order then to be in a position to favour reform. In each case, people have thought they could condemn harsh treatment without having first to postulate moral rights in animals in order to do so.
>
> …it would be utterly perverse to point to a man writhing in agony as the result of being tortured by scalding and to claim that what was wrong there was that the man's moral rights were being violated…Moral rights are excess baggage here; for there is nothing required to mediate between the (extraordinarily) painful character of the act and its wrongness.
>
> (Frey, 1983, p37)

The point here is that we can, both as a society and as individuals, decide on how to act based on whether we judge something to be right or wrong, without first having to divert our attention to a set of rights, which are themselves, I contend, based on variable moral attitudes. Further, even it is shown that moral rights do have a role to play in the debate over the treatment of non-human beings, it is still to be shown that (a) non-humans do in fact possess these rights and (b) a right to life is one of them.

With regard to the first of these claims, that non-humans do possess such moral rights, Frey (1980) argues that animals do not in fact possess them and that, due to the fact that no animal can undergo the level of emotional or intellectual experiences undergone by humans, their lives do not have the same moral value to that of humans. An obvious initial response to

[43] See www.prolife.org.uk and www.prochoicemajority.org.uk for more on this

Frey would be to demand an explanation as to why intellect and emotional capacity are placed in such a prominent position. Surely an animal's capacity to have a conscious experience of pain and suffering is an equally important factor determining how an animal should be treated. If any organism has the capacity for experiencing pain and suffering, then the level of its intellect seems like a poor reason for subjecting it to such experiences. However, even the strongest of cases for a being's moral right not to be subjected to pain and suffering fails to explain why swift and painless (painless in both a physical and psychological sense) killing is morally objectionable. In order for the slaughter to be seen as a moral transgression, it must be shown that the slaughtered beings do in fact possess a moral right to life. It is incumbent upon proponents of an animal's moral right to life, to explain why, for example, after thousands of years of killing animals for food, we should bring an end to our carnivorous ways.

Similarly, before any synthetic consciousness could justifiably be placed in our moral world, it would have to be shown that such a conscious system possessed moral rights and that, once created, a right to continue living this life was one of them. To base such a decision on intellect alone would not be appropriate. To base the decision on whether the machine could be defined as a moral agent, given that certain rights are assigned to animals despite our tendency not to see them as moral agents, would also be inappropriate. But, given what we've said above, even if a machine did possess the capacity to experience the unpleasantness of pain and suffering, this alone does not prevent us from quickly and painlessly bringing its existence to an end by, for example, pulling the plug or cutting off its power supply.

Regan (1983) argues that all animals who are the 'subject of a life' have an 'inherent value' and, so, should not, for any reason, be treated as a means to an end. Such ends would, I presume, include food, fur coats and hunting for sport. According to Brennan and Lo (2008) to be a subject of a life '...involves, among other things, having sense-perceptions, beliefs, desires, motives, memory, a sense of the future, and a psychological identity over time'.[44] Importantly, any animal – and so, it might reasonably be presumed, any being – that can satisfy these conditions is a subject of a life and should never be used as a means to an end. There are others, including Paul Taylor (1986), who take this a stage further by extending this 'inherent value' to all living things. Taylor argues that all living things in nature are 'teleological-centres-of-life', possessing an equal moral right to consideration and respect; thus living things should never be used as a means to an end, but treated only as ends in themselves.

The problem with such a claim is that it is very difficult to maintain when applied to real life situations, and it reveals a number of problematic grey areas. Imagine, for example, a local council decide that there is a desperate need for an Accident and Emergency hospital which has the potential to save hundreds if not thousands of human lives. The site they decide to build on is covered with natural forest and the council promise to plant more trees than they cut down, in another area, in order to redress the balance. In the long run~~,~~ there claim is that these actions will actually benefit the environment. Taylor however, would be forced to say that, despite the fact that there are more trees now growing in the community and thousands of human lives are being saved, the council have still committed an act that is morally wrong because each living thing deserves an equal amount of consideration and killing one hundred

[44] [http://plato.stanford.edu/entries/ethics-environmental/, Stanford Encyclopaedia of Philosophy

trees to replace them with two hundred new ones is not assigning equal consideration to the trees that are felled.

Frey raises another interesting question: does a withered, dying tree or flower have the same inherent value and therefore deserve the same moral consideration as, for example, a healthy dog or even a newborn baby? If so, there seems no moral contradiction in the claim that our last drop of water should be given to the dying flower rather than to the baby or even the dog. Of course, Taylor can hold fast to his position and accept that this indeed throws up no moral contradiction at all; but it is worth asking pragmatically what he would do if confronted with the choice in real life. The example stated above does not, of course, prove that animals have no inherent value, or even that they have less value than humans; what it does show is that a claim made about every living organism being shown equal moral consideration is very hard to maintain and is unlikely to be accepted by the majority of us in our practical daily concerns.

Regan, since he does not go as far down the scale as Taylor, can respond to this by simply arguing that any conscious being is not replaceable in the way that trees, flowers or plants are. They, unlike trees and flowers, have thoughts, feelings, phenomenal sensations, emotions, memories, psychological identity, beliefs and desires. In other words, they are subjects-of-a-life in a singular way.. This is not unlike the idea we have of humans; it would seem odd not to grieve for the death of a friend on account of the fact that we can make a new one, even if the new one was an identical twin with no discernible differences in either appearance or personality. Indeed, it is similar to the way people feel about pets; slaughtering somebody’s dog and presenting them with a new one, even one of the same breed, size, colour and so on, would not be met with a favourable response. This is a fairly strong argument and relates to what I discussed earlier with regard to fungibility. If a synthetically created being lives a conscious life that is rich in mental activity, has a sense of identity and of its potential future, derives pleasures from life that it both remembers as pleasurable and seeks to enjoy again, has interests and things that it cares about, and holds a desire to survive beyond the present moment, then it is difficult to argue that it is morally permissible to kill such a being without conceding that it is equally permissible to do the same to a human being.

The biggest problem facing Regan is the stringency of his criteria for being a subject-of-a-life; he may find that, in attempting to identify beings who fully meet the criteria, the range of beings that we are not morally justified in killing reduces to include conscious machines. How narrow this range becomes is, at this stage, unclear. This is because, apart from the ongoing debate surrounding which animals have which capacities, Regan readily admits that his criteria may not be sufficient for a being to be considered the subject-of-a-life, and the addition of any further criteria make the conditions even more stringent and narrows the range still further.

An appeal to moral rights seems not to answer the question of why killing is wrong, even if we resist the extreme route taken by Frey in denying that moral rights have any place in ethical theory. Moral rights, by their very nature, are dependent on moral judgements, values and beliefs that can and do vary over time and societies. A blanket ban on the killing of any living being for any reason seems not to be a viable option since, in practice at least, it permits some very questionable decisions in the face of moral dilemmas. There does remain the possibility of machines that are subjects-of-a-life and would deserve a level of consideration equal to that of humans; but first Regan must show that there are beings other than humans who meet the rigorous requirements.

(2) Responsibility and Accountability

Earlier I discussed briefly the accountability parents have with regard to the actions of their offspring and how we often hold the parents responsible for the misdeeds of their children; this in spite of the fact that the parents cannot claim ownership of the child and nor do they have the moral (or legal) right to do what they will with the child. There may be many reasons why the parents do not own the child and so are not entitled to treat it as they would their possessions; these might be because the children have rights, that they have feelings and desires, that they can experience physical and emotional pain, that the propagation of our species relies on children being looked after, or simply that mistreating a child is just not nice. Whatever the reason(s), the most significant are that once the parents bring a conscious, thinking, feeling being into existence, they have a moral, legal, and civic duty to: (1) protect it from harm, (2) where possible, prevent it from harming others, (3) ensure the best life possible for it and, (4) accept responsibility for, and take appropriate action if, they fail to achieve any of the first three conditions. Clearly, there is far more to parenting than this narrow set of constraints, and there is an almost infinitely expansive array of conditions that would have to be met in order to satisfy (1), (2) and (3). However, for the purpose of this thesis, it is sufficient that the parents are morally obliged to ensure the child comes to no harm at the hands of others, that others come to no harm at the hands of the child, and that the child is given the best possible start in life. Failure to make a genuine effort towards satisfying these conditions is almost always seen as neglectful and so morally reprehensible. The moral, legal and civic duty to adhere to these criteria arises simply in virtue of the fact that the child is a sentient, conscious creation of the parents, but how far can these same criteria be applied to a non-biological creation where, instead of parents we have a bunch of machine engineers, and in the place of a child, we have a conscious, thinking, and feeling machine?

Firstly, the moral propriety of keeping a conscious machine, or any conscious being for that matter, from harm, seems obvious. There should be no disputing the claim that, if there is any being, organic or inorganic, capable of experiencing physical or emotional pain, then we have a moral duty to ensure that, where possible, we do not unnecessarily inflict such pain upon them[45]. In respect of the lengths we are morally obliged to go to in order to *protect* others from harm, this is something which cannot be so easily identified. In the case of our offspring, the legal onus to protect them from harm only lasts until adulthood, although most parents worry about the welfare of their children far beyond this point. However, failing to shield a young child from harm would, unsurprisngly, cause a parent to be held in far greater contempt than the failure to protect a 30 year-old son or daughter from harm.

The reason we are morally bound to ensure the safety of our children, is because children, especially at a very young age, are ill-equipped to deal safely with the contingencies of life and, in the absence of parental assistance, are far more likely to come to harm. It hardly seems fair~~,~~ to expect, for example, that a six year-old buy, prepare, cook, and serve their own dinner, because there are many aspects of this process which make it impossible or, at least,

[45] I do accept however, that this may be very wishful thinking on my behalf and that there are those who inflict unspeakable and unnecessary pain and suffering on animals and, in some cases, fellow humans.

dangerous for the child to undertake and so we have a moral responsibility to ensure that these things are done for her. Similarly, expecting a twelve year-old to go out and work in order to pay for their own food and clothing seems rather harsh, and so we also have a moral duty to feed and clothe children of this age. As children grow into adulthood the parents' responsibilities diminish as the child's ability to manage on its own increases. Eventually, as the child grows into adulthood, she becomes self-sufficient and the number of things that the parents are morally obliged to do diminishes. Beyond this point, any assistance offered by the parents arises not from moral obligation, but from a generosity of spirit, which might reasonably be shown to anyone from a close familial bond with the child. Therefore, the moral obligation on behalf of the parent to protect their conscious creation from harm depends on the extent to which the creation can function in everyday life, on its own, with a minimum risk of suffering injury. If we turn now to a conscious machine and the machine is created with a full working 'adult level' knowledge of the world, then its creator's obligations only extend as far as a parent's would to a grown son or daughter; if, on the other hand, the machine is constructed with all of its learning still to do, then it would be morally incumbent upon the creator to guide and care for the machine until, like the maturing of a human child, the machine becomes self-sufficient and competent enough to live autonomously, in the world. If the machine never reaches this level of competence and functionality, then the creator has a moral duty to care for it indefinitely, just as a parent would have a duty to care for a child born with learning difficulties.

With regard to the second part of a parent's responsibilities detailed above, ensuring that the child does not cause harm to others, the responsibilities alter slightly, especially if the machine is created with a full, working, adult consciousness. In creating a machine with a consciousness equivalent to that of a child, like all parents, the creator has a responsibility to raise his creation well and allow it to develop its own moral values. If the machine develops a vicious nature quite unlike his creator, then so long as the creator has tried to instil good moral values in his creation, the level of responsibility is lessened. There is a difference however, if the machine is created with a full working, adult consciousness and is sent out to live and function in the world immediately after its creation. The responsibility then must surely fall with those charged with the task of equipping the machine with its moral principles, which would surely reflect those of the creator(s), since there is a contradiction in assigning principles for decent moral living which run contrary to one's own[46]. There would be little point in apportioning blame to a machine that insulted or assaulted a black person if it has been programmed to revile any non-white human; in such cases, the culpability surely lies entirely with the bigoted and intolerant designer. It may be argued, that if it's a conscious machine then can be taught to behave with a greater racial toleration, it would be a conscious, responsive thing that can change its behaviour. However, as can be seen from the bigotry and religious intolerance currently rife in the West of Scotland, changing moral attitudes is very difficult and in some cases impossible. In any case, I have already discussed the difficulties of implementing a moral code in a machine, difficulties which should cause creators of conscious machines to think twice before placing such machines in homes and workplaces.

(vi) The Indefensibility of Robot Servitude

[46] I make this claim about how we believe how we ought to live and not how we actually live. For example, I believe I should do much more for charity, but rarely get round to doing more than making the occasional donation.

In section 2 (ii) I claimed that before any ambition to create a synthetic consciousness is realised, we have a moral obligation to examine our reasons for creating such a being, because little justification could be found for doing it 'just to see if we can'. Historically, technological advances have been sought for three main reasons: firstly, to increase our ability to defend ourselves through advancements in weaponry, secondly to improve our quality of life, whether they are advances in medicine, gadgetry, or travel and, thirdly, to increase our knowledge and understanding in a particular area or scientific field. I have already argued that an enhancement in our understanding of human consciousness will not necessarily be realised through the creation of a synthetic one and so, in this section, I intend to look at the moral consequences of creating conscious machines designed to lighten the burden of everyday human exertion. In particular, I aim to examine the moral indecency of what Stephen Petersen (2007) refers to as Engineered Robot Servitude (ERS).

Effectively, ERS is the creation of robots designed to carry out tasks that humans tend to dislike; these tasks could be anything from doing dishes, washing the car, vacuuming, or bathing the dog (the example used by Petersen is that of 'laundry bots' - robots designed to do laundry). It is important to note the 'engineered' part of the term, since Petersen does not advocate robot slavery, where conscious robots are forced to carry out tasks which they do not desire to do, nor does he support the modification of already born humans, by altering them in such a way as to give rise to desires for the performing of such tasks.

> ERS: The building and employment of non-human persons who desire, by design, to do tasks humans find unpleasant or inconvenient…the design must be "from scratch". I am not talking about what you might call post-identity modification – the manipulation of an already existent person's desires to new, servile desires that would have been against the pre-modified person's will. I take such cases to be uncontroversially wrong, whatever the material nature of the person so modified. Instead, I am thinking of cases where the person comes into being with the servile desires intact.

Essentially, Petersen argues that there is no moral impropriety in designing robots, or any conscious machine, with desires that suit human ends. The 'laundry bot' for example, would be designed with a desire to do laundry, which may come from sub-programs that cause the machine to react negatively to unclean laundry or to experience a pleasant reaction to the smell of freshly laundered clothes. Similarly, a 'vacu-bot' may be designed to react negatively to messy floors and may find clean, vacuumed carpets aesthetically pleasing; each of these would contribute to the machine's pre-programmed disposition to vacuum the floor. Indeed, Petersen believes that any moral impropriety would lie in prohibiting these machines from carrying out the tasks which would satiate their respective desires, even if such machines were in possession of consciousness and intelligence equivalent to that of humans. He goes on to argue that we already have instances of organisms acting upon desires which benefit humans without any apparent moral transgression. He offers Retriever Dogs as an example because they have an innate desire to fetch things and seem to derive great joy from doing so. Consequently, there is nothing unethical in allowing these dogs to act upon these desires in ways that benefit humans and, in fact, the immoral thing would be thwarting its desire. Furthermore, presenting Retrievers with human intelligence would only make them more resourceful and adept at retrieving; it in no way suggests that the basic, inherent desire to fetch would in any way subside.

This idea rests on the premise that it is possible to have a human level of consciousness and intelligence in the absence of human goals, aims and desires. Likewise, conscious, intelligent robots would have goals and desires of their own, which would very likely be far removed from those of humans, and only the prevention of satiating those desires would be seen as morally impermissible.

An obvious objection to the idea of ERS is that there seems no reason to assume that there is anything unethical in Engineered *Human* Servitude (EHS). If we are morally justified in implementing conscious machines with desires which serve human ends, then it seems that we would be perfectly justified in doing so, possibly at a genetic level, with conscious human beings. For Petersen EHS is not obviously wrong, but in any case is not analogous to ERS. A virtue ehticist would argue for the impropriety of EHS based on the fact that the wellbeing of human beings rests on their NOT having a particular function. Aristotle (1998), asks us to think what the ergon of man might be and he draws a blank; human beings are not tools which have a function, the telos of man, if one can be devised is to live a life in accordance with the mean, that is, a life of contentment. However, Petersen replies to this by insisting that such an argument 'completely severs the analogy with engineering robots'. Robots *are* designed to have a particular function and so would be pursuing the Aristotelian idea of *eudaimonia* in carrying out the tasks that they are designed to do, whether these be washing up, vacuuming or doing laundry.

Kantian ethics also fails to explain why ERS is wrong. Kant would argue that it is wrong to treat laundry bots or vacu-bots in such a way as to allow them to serve us, since this would have a detrimental effect on our moral attitudes towards other conscious, autonomous creatures (other humans). However, by allowing a laundry bot to carry out the task of doing laundry – which it has an inbuilt desire to do – then we are allowing it to experience joy, which if reflected in our treatment of others, is a positive thing. Kantians may also argue that the laundry bot is being abused for human ends, but such an argument requires identification of the moral aberrance in pursuing laundry as an end.

However, there is a deeper problem with the idea of creating obsequious, servile, yet highly intelligent, machines: happiness is not necessarily directly proportional to the level to which we are able to satiate our desires. Take for example a heroin addict. Heroin addicts, at times when they are not under the influence of heroin, experience pains and cravings that most of us may never understand; so much so, that they are driven to beg, steal or commit other criminal acts in order to get the money to acquire more heroin. When they eventually obtain the heroin that they so desire, their aches, pains and cravings may go, but it is hard to describe them in any meaningful way as 'happy'. Similarly, a smoker, in the midst of a battle to quit, will almost certainly have an intense desire for nicotine. After a while, or a particularly stressful day, the smoker may eventually give in and go outside for a cigarette. Although, the desire and craving have been sated, there will be other emotions accompanying the decision to capitulate, including guilt, frustration, embarrassment and possibly even low self-esteem at the thought of being so thoroughly ruled and controlled by little bits of crushed leaves and nasty chemicals wrapped in paper.

The point to note is that if we create conscious and intelligent machines with pre-programmed desires to carry out mundane and menial tasks, there is the possibility of the machine having an irrepressible desire to perpetually launder clothes, for example, but at the same time feeling thoroughly embarrassed, frustrated, ashamed or even guilty for having such desires. A similar point can be made for the war robots discussed earlier; the robot may have inbuilt desires to kill enemy soldiers which it cannot override, much in the same way the smoker cannot give up nicotine, but it may feel abject shame for not only acting upon these desires, but being cursed with the inability to desist. The example of the Retriever, used by Petersen, also throws up this possibility. It is true that the Retriever, after being elevated to the level of human consciousness, may still have the desire to fetch, but it may also become miserable at the realisation that it has no loftier desire than the one which compels it to spend its days, and ultimately its entire life, retrieving. Each of the above cases are examples of conscious, intelligent beings, whether human or non-human being fully permitted to act upon their desires and yet falling some way short in their quest for a happy existence. Consequently, equipping machines with pre-programmed desires in no way ensures that the machines will have, fulfilling existences acting upon these desires, particularly if they have an intelligence and consciousness equal or superior to that of humans.

Petersen attempts to get around this problem by disallowing robot slavery, in which case the laundry bot that no longer derives any pleasure from doing laundry would not be forced to do so and, in fact, should be set free immediately. But this response is thoroughly unsatisfactory. Firstly, there is a considerable experiential difference between not acting upon a desire which makes us unhappy and not having that desire at all. The idea of setting the laundry bot free from servility in no way brings emancipation from the desire to do laundry. Indeed, unlike the heroin addict who at least feels better while on heroin, the laundry bot could be miserable whilst doing laundry - in a similar way to the unhappy smoker with feelings of guilt, shame, embarrassment, and self-loathing – and miserable whilst not doing it as a result of its pre-programmed desire to launder clothes. The only way to ensure the avoidance of this misery, other than abandoning our attempts to create human consciousness, is to refrain from designing conscious machines with desires that we ourselves would find objectionable. But there is no way of guaranteeing this if we place a human consciousness – or any type of consciousness for that matter – in an inorganic vehicle. Secondly, the idea of setting the machines free after they have reached a state of dissatisfaction with their existence deals another major blow to anyone attempting to justify the creation of a human consciousness. If one of the reasons given for manufacturing consciousness is that it can be designed to make life easier for humans, then there seems little point in creating machines which can get up and leave any time they wish; apart from anything else, there's the economic issue: who would buy such a machine when this possibility exists? Thirdly, and related to the second, there is the question of ownership. If an individual pays to have a laundry bot or something similar in their homes, it seems that they would be very reluctant to watch their investment down tools and leave in the event of a waning desire to work. Even if it can be shown that conscious, intelligent machines with a disposition to serve humans could, with moral agreement, be paid for and kept in the home, it would seem unethical to retain machines without such a disposition against their will: to do so would be tantamount to slavery.

Perhaps a solution to the above problems can be found in refraining from placing a synthetic consciousness with a high level of intelligence in, for example, a laundry bot. Instead, by placing a lower level, or different degree, of consciousness in the robot, we reduce the risk of the unpleasantness associated with guilt, embarrassment or frustration. By designing the robot in a way that prevents it from experiencing these emotions, questioning the value of its existence or concerning itself with the merit of its circumstances, the robot will never feel a low sense of self-worth. Christopher Grau (2006) asks the question: "…whether we may be morally obliged to limit the capacities of robots". Although Grau asks this question in an attempt to show the inappropriateness of creating utilitarian robots, an idea that is examined elsewhere in this thesis, the same point can be applied here; if we are intent on creating conscious, sentient machines for the purpose of serving us, then we have a moral duty to ensure that the machine is incapable of experiencing any negative side effects as a result of its drudgery. According to Grau, there would be nothing wrong with Petersen's idea of ERS, so long as the robots do not have 'morally relevant features', such as sentience, autonomy or an 'identifiable self'. Grau's response only serves to add to the case against those who argue that there is justification in the effort to create conscious machines. It is true that the machine, if designed in the way described above, could not experience the negative emotions associated with unwanted desires, but it is also true that, by being unable to question the merit of its vocation, it will fail to derive any pleasure from it either. In fact, it will have no thoughts or attitudes to its work whatsoever. In this case, it seems that consciousness is a superfluous addition to the machine, which could operate with an equally high level of competence and efficiency in the absence of any conscious thought.

6. Conclusions

Unquestionably, machine ethics should be as important to the design, planning and implementation of a synthetic consciousness as the nuts, bolts and computer wizardry employed to make it a possibility. Although ethical consideration has never been at the forefront of our minds in most of our technological advancement in the past, the creation of conscious, thinking, active beings is, without question, our most ambitious project to date, but it also has the potential to be our most dangerous.

We have a moral duty to examine our reasons for creating machine consciousness and, through some focused consideration, we can see that identifying these reasons may not be as straightforward a task as it first appears. Invention for invention's sake is not always a good reason for bringing something into existence and the idea of machines being fungible fails to provide satisfactory justification, since consciousness may bring with it other aspects of mental activity, such as desires, emotions, memories and psychological distress, which render a being far from fungible. Nor can we appeal to the possibility of gaining a better understanding of our consciousness as justifiable reason for creating another. The fact is, that any synthetic consciousness, although it may produce the same results and effects, will do so in an entirely different manner to that of human consciousness and will shed little light on the inner workings of the latter; in the same way both fire and electricity emit heat, but the study of one gives little or no understanding of the other. Lessening the effort involved in our daily chores also fails to give sound reasons for the creation of super intelligent conscious robots, mainly because such beings would desire far more reputable endeavours than those which save humans from doing housework. And even if the machines are pre-programmed to desire the completion of these menial tasks, there is no guarantee that allowing them to satiate these desires will provide them with even a palatable existence, just as smoking a cigarette often fails to bring blissful, elated pleasure to the disheartened smoker.

There are too many incalculable variables and unanswered questions regarding the consequences of the creation of a synthetic consciousness to make it a morally justifiable project. It could be argued that the invention of electricity, quicker ways of travelling between distant places and advances in medicine, among others, heralded the dawn of cleaner, healthier and safer living, things which could only be seen as benefits.[47] Now, however, we are inventing things which are contributing to a lack of good health and social

[47] It must also be said however, that they've also polluted the planet and used up its resources.

skills and which remove, almost entirely, the effort from living. Before embarking upon a journey towards creating a synthetic consciousness, we are morally obliged to provide justifiable reasons for doing so and, as of yet, these reasons have not been forthcoming.

Due to the nature and level of intelligence, and the likely physical prowess of conscious machines, it is crucial that we identify a code of ethics which prevents our creations from turning upon us in a malevolent and harmful way. Further, we must also ensure that any conscious creations are treated in such a way that keeps them free from mental or physical suffering, otherwise we could, even inadvertently, inflict unnecessary pain on innocent victims. However, each of these conditions bring with them problems for which solutions cannot be found, so there is a further moral burden on us to delay or abandon our attempts to create conscious machines. In attempting to protect ourselves, it is vital that we implement a moral code which prohibits our creations from harming us. But developing such a code, especially one which prevents the arising of harmful desires and which is devoid of loopholes is extremely difficult given the fact that we do not yet have a firm enough understanding of our own moral world. It is also questionable as to whether we are morally permitted to 'force the machines to be good' through programming them in such a way as to disallow desires that we humans find objectionable.

But self-preservation should, at most, only form half of our concern. If history has taught us one thing, it is that we are not the most tolerant of species. Wars had to be fought and people had to die in order for civil human decency to be afforded to those with different skin colour and, even now, there are the intolerant among us who still refuse to do so. How long after the creation of conscious machines would humanity accept the claims of machines to be treated with dignity, decency and respect? Would humanity ever reach this acceptance? If not, the widespread pain and suffering inflicted on our conscious creations could potentially rival that of the misfortunate souls who, because of the colour of their skin, often endured unimaginable existences before, and for a long time after, the emancipation proclamation. Already, even before there is a glimmer of light at the end of the tunnel for those attempting to create a human consciousness, there are those, like Petersen, who are attempting to find justifiable reasons for placing conscious, thinking and intelligent beings into servitude. Not, I would suggest, the most promising of moral starts.

Even if all of these problems can be solved through the implementation of a robust, infallible moral code, the designing of such a code is, at present, well beyond our reach. This is in part because we do not yet have an infallible moral code of our own and in part because ethics does not lend itself to algorithmic calculations. Calculations rely on there being facts to calculate and so in order for a machine, or anyone for that matter, to perform a moral calculation, they would need to be in possession of moral facts, none of which we have been able to identify in thousands of years of trying. There is of course the possibility of using estimates or 'best guesses' rather than facts, but the whole point of being able to use moral calculations, as far as Anderson & Anderson are concerned, is to eradicate the indecisiveness associated with human moral judgements. In any case, the idea of super-intelligent, highly-powered machines 'getting it wrong' should surely fill us with a deep concern. Additionally, using numbers and formulas to ascertain moral propriety helps us very little with moral dilemmas if the two alternatives provide us with the same number or result.

Unless answers and solutions can be found to these questions and problems, we should not be attempting to build conscious machines which, in virtue of their being conscious, have the capacity to think, feel, remember and desire. In doing so, we are creating the potential for more pain, suffering and conflict in the world: things we already have in abundance.

Addendum

Throughout this thesis, I have made no allusion to the concepts of pleasure or happiness when dealing with the determination of moral consideration. This is because I believe that our moral obligation for the feeling of pleasure or happiness in other species (and possibly, it could be argued, in humans) only extends to that which is felt at the alleviation of their pain or suffering, or in being left unmolested in the pursuit of their own interests if, indeed, they have any. This obligation does not extend to the arousal of happiness in any other conscious being.

An obvious objection to this may be the keeping of pets and the fact that we are obligated to do more than just keeping them from coming to harm or suffering. I do concede that anybody who opts to have a pet takes the additional responsibility of making that pet as comfortable and happy as is feasible. However, the having of a dog, for example, is not a moral obligation, and the fact that other people have them in no way increases, in any way, the moral obligation that I have towards dogs.

It may be the case that I would be considered a nicer or more considerate person if I made it my aim to bring as much pleasure and happiness to every human and non-human that I came across. My morality however, would surely only be brought into question if I refused assistance to people or animals that I took to be in pain or suffering, or if I was the cause of them. It hardly seems justified to cast a person in a negative moral light for, example, failing to give help or assistance to any human or non-human that is not in obvious pain or state of suffering. The bringing of pleasure or happiness then are, I believe, insignificant in the determination of moral consideration.

<u>Bibliography</u>

Aleksander, I. & Dunmall, B. ‘Axioms and Tests for the Presence of Minimal Consciousness in an Agent’ *Journal of Consciousness Studies*, 10, 2003

Aristotle, *‘The Nicomachean Ethics’*, Oxford University Press, 1998

Aristotle, *‘De Anima’*, Penguin Books, 1986

Ayer, A.J ‘*Language, Truth and Logic’* (1936), Penguin Books, 1990

Bauby, J.D. (2008) ‘*The Diving-Bell and the Butterfly*’ HarperPerennial

Bentham, J. (1789) *An Introduction to the Principles of Morals and Legislation*,. Dover Publications, 2009

Bostock, S. *‘Adaptations, Exaptations, Spandrels and Altruism*’, lecture notes 28/11/08

Candland, D.K. (1993) ‘Feral Children and Clever Animals: Reflections on Human Nature’ Oxford University Press,

Carruthers, P. (1992) *‘The Animals Issue’*, Cambridge University Press,

Chalmers, D.M.(1987) ‘*Hooded Americanism: The History of the Ku Klux Klan*’ Duke University Press; 3rd Revised edition.

Chrisley, R. (2009) “Synthetic Phenomenology”, *International Journal of Machine Consciousness*, 1 (1), pp.53-70

Damasio, A. R. (1994) ‘*Descartes’ error: Emotion, reason, and the human brain*. New York: G. P. Putnam

Damasio, A. R.(2001) ‘Fundamental feelings’. *Nature*, *413*, 781

Damasio, A.R. ‘*Descartes’ Error: Emotion, Reason and the Human Brain’*, Vintage, 2006

Darwin, C. (1859)*‘The Origin of Species’* , Wordsworth Classics, 1998

DeGrazia, D. (1996) *‘Taking Animals Seriously’*, Cambridge University Press.

Dennett, D. C. (1989) *The Intentional Stance*, MIT Press.

Derbyshire, S. (1992) *‘Animal Experimentation’*,

Descartes, R. (1637) ‘Meditation VI’, in *Meditations on the First Philosophy*, Everyman, 1997

Faulkner Shand, A. ‘The Foundations of Character: Being a Study of the Tendencies of the Emotions and Sentiments (1914), Kessinger Publishing, 2008
Fodor, J. ‘Making Mind Matter More’, *Philosophical Topics* **17**, pp. 59-79, 1989

Franken-Paul, E. et (2001) *‘Why Animal Experimentation Matters: The Use of Animals in Medical Research’*, Transaction Publishers,.

Frey, R.G. (1980) *‘The Case Against Animals’* Oxford University Press,.

Frey, R.G. (1983) *‘Rights, Killing & Suffering’*, Basil Blackwell Publisher Limited.

Gallagher, S. ‘Moral Agency, Self-Consciousness and Practical Wisdom’. *Journal of Consciousness Studies* 14(5-6), 199-223

Goldman, A.I. (2008) ‘*Simulating Minds: The Philosophy, Psychology and Neuroscience of Mindreading*’, Oxford University Press.

Griffin, D.R. (1984) *‘Animal Minds’* The University of Chicago Press.

Harman, G.& Thompson, J. (1995) *‘Moral Relativism and Moral Objectivity’*, WileyBlackwell.

Hobbes, T. (1651) ‘*Leviathan*’, Penguin Classics (1985)

Hume, D. (1739)*‘A Treatise of Human Nature’* , Penguin Books, 1985

Humphrey, N.K. ‘*Nature’s Psychologist’*, Humphrey, N.K. ‘Nature’s Psychologist’ In ‘Consciousness and the Physical World’ ed. B Josephson and V. Ramachandran, pp. 57-75, 1985. http://www.humphrey.org.uk/papers/1980NaturesPsychologists.pdf

Husserl, E. (1927) *‘Introduction to Phenomenology’*, Trans. Palmer, R.E, Encyclopaedia Britannica.

Huxley, T.H. (1874) *‘On the Hypothesis that Animals are Automata, and its History’* in Method and Result: Collected Essays Part One, Kissinger Publishing, 2004

Kant, I. (1785) ‘*Groundwork of the Metaphysics of Morals’* Wilder Publications Limited, 2008.

Kim, J. (2008) ‘*Physicalism, or Something Near Enough*’, Princeton University Press.

Locke, J. (1690) ‘*Two Treatise of Government’*, Cambridge University Press, 1997

Locke, J. (1690) ‘*An Essay Concerning Human Understanding*’, Everyman, 1996

Mackie, J.L. (1990) ‘*Ethics: Inventing Right and Wrong’*, Penguin Books.

MacLean, N. (1995) ‘*Behind the Mask of Chivalry: The Making of the Second Ku Klux Klan’*, OUP USA; New edition,.

Mill, J.S. (1863) *‘Utilitarianism’ Hackett Publishing Company, 2001*

McGinn, C. '*Can We Solve the Mind-Body Problem*?' in 'Modern Philosophy of Mind', Lyons, W. (ed), Everyman Library, 1999. First published in *Mind*, Volume 98 (1989)

Montgomery, J. '*Children as Property?*' The Modern Law Review, Vol. 51, No. 3 (May, 1988), pp. 323-342

Moore, G.E. (1903) '*Principa Ethica'*, Cambridge University Press.

Nagel, T. (1974) *'What's it Like to be a Bat?'* in Modern Philosophy of Mind, Lyons, W. (ed), Everyman Library, 1999. First published in The Philosophical Review, Volume 83.

Newton, M. (2002) *'Savage Girls and Wild Boys: A History of Feral Children'* Faber and Faber Limited.

Nussbaum, M.' Emotions as judgements of value and importance' In R. C. Solomon (Ed.), *Thinking about feeling: Contemporary philosophers on emotions* (pp. 183–199). New York: Oxford University Press. (2004)

Paine, T. (1791) '*Rights of Man'*, Penguin Classics, 1985

Petersen, S. The Ethics of Robot Servitude. *Journal of Experimental and Theoretical Artificial Intelligence* 19 (1):43-54. (2007)

Rawls, J. (1971) '*A Theory Of Justice*' Harvard Univerity Press.

Regan, T. (1983) *'The Case for Animal Rights'*, London Routledge & Kegan Paul.

Robinson, J. Startle. *The Journal of Philosophy*, *92*, 53–74 (1995).

Robinson, J. (2004) Emotion: Biological fact or social construction?

Robinson, J. (2005) *Deeper than reason: Emotion and its role in literature, music, and art*. Oxford, UK: Oxford University Press .

Ryder, R.D. (2001) *'Painism: A Modern Morality'*, Centaur Press.

Rymer, R. (1994) '*Genie: A Scientific Tragedy'*, HarperPerennial.

Singer, P. (ed.) (1993)*'A Companion to Ethics'* Blackwell Publishing.

Singer, P. (2002) *'Animal Liberation (3rd edition)'*, HarperCollins Publishing.

Solomon, R. C. 'The Logic of Emotion. *Noûs*, *11*, 41–49 (1977)

Stamp Dawkins, M. (1993) *'Through Our Eyes Only'*, WH Freeman and Company.

Stich, S. & Nichols, S. "Folk Psychology: Simulation or Tacit Theory?" *Mind and Language* (1992)

Stich, S. & Ravenscroft, I. (1994) "What is Folk Psychology?" *Cognition*.

Storrs Hall, J. (2001) '*Ethics for Machines'* KurzweilAI.net, 2001

Stuart, S. "Machine Consciousness: Cognitive and Kinaesthetic Imagination", *Journal of Consciousness Studies*, Imprint Academic, (2007) **14** (7) pp.141-53

Stuart, S. (2009) "Conscious Machines: Memory, Melody and Muscular Imagination", Springer Science.

Taylor, J. (1651) *'The Rule and Exercises of Holy Dying*', BiblioBazaar Publishing, 2009

Taylor, P. (1986) *'The Ethics of Respect for Nature'*, Princeton University Press.

Wade, W.C. (1998) *'The Fiery Cross: The Ku Klux Klan in America'* Oxford University Press Inc; Reprint edition.

Wallach, W. & Allen, C. (2009) *'Moral Machines: Teaching Robots Right From Wrong'* Oxford University Press.

Williams, B. & Smart J.J.C. (1973) '*Utilitarianism: For and Against'*, Cambridge University Press.

Wilson, G. (1997) "Reasons as Causes *for* Action", *Contemporary Action Theory*, ed. G. Holmstrom-Hintikka & R. Tuomela. Dordrecht: Kluwer.

Zeman, A. (2002) *'Consciousness: A Users' Guide'*, Yale University Press.

Other Sources

www.animals.org/reasons.htm, *'Reasons for Cruelty Towards Animals'*

www.cs.ucl.ac.uk/staff/t.quick/autopoiesis.html *'Autopoiesis'* 16th May 2008

http://www.dolphins-world.com/Dolphins_Rescuing_Humans.html

Anderson, M., Anderson, S.L. 'Machine Ethics: Creating an Ethical Intelligent Agent (December 1, 2007. http://www.allbusiness.com/science-technology/computer-science-intelligent-agents/8893432-1.html

Allen, Colin, "Animal Consciousness", *The Stanford Encyclopedia of Philosophy (Winter 2006 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2006/entries/consciousness-animal/>.

Bringsjord, S. '*Ethical Robots: The Future Can Heed Us'* http://www.aaai.org/library/symposia/Fall/fs (2006)

Dretske, F. '*What Good is Consciousness?*' Canadian Journal of Philosophy *27 (1):1-15.* 1997 http://evans-experientialism.freewebspace.com/dretske.htm

Fechner, G. T. (1860/1912) Elements of Psychophysics, Sections VII & XVI, trans. Herbert Sidney Langfeld, URL: http://psychclassics.yorku.ca/Fechner/ accessed 25/10/2009

Feser, E. '*Self-Ownership, Abortion and the Rights of Children: Toward a More Conservative Libertarianism*' Journal of Libertarian Studies Volume 18, no. 3 (Summer 2004), pp. 91-114. Accessed 01/11/2009

Gee, P. (2003) "Teaching fish to tell the time!", http://www.plymouth.ac.uk/pages/view.asp?page=7705 Accessed 19/11/09

Grau, C. 'There is no "I" in Robot: Robotic Utilitarians and Utilitarian Robots http://www.aaai.org/Papers/Symposia/Fall/2005/FS-05-06/FS05-06-007.pdf (2005)

Johnson, G. '*Theories of Emotion'* http://www.iep.utm.edu/emotion/

Joy, B. '*Why The Future Doesn't Need Us'* (2000) http://www.wired.com/wired/archive/8.04/joy/html. 2000

Montague, P. 'When Rights Conflict' http://journals.cambridge.org/action/displayAbstract;jsessionid=416040CD55D5137FAE0D5DC5079554A4.tomcat1?fromPage=online&aid=100907

Slate Online Magazine http://www.slate.com/id/2192211/

University of Southern California (2007, March 22). Moral Judgment Fails Without Feelings. *ScienceDaily*. Retrieved October 13, 2009, from http://www.sciencedaily.com /releases/2007/03/070321181940.htm

Walter, S. '*Epiphenomenalism*' The Internet Encyclopedia of Philosophy, 2007http://www.iep.utm.edu/e/epipheno.htm

Wikipedia (Deep Blue) http://en.wikipedia.org/wiki/Deep_Blue_(chess_computer)

World Health Organisation: www.who.int/dietphysicalactivity/publications/facts/obesity/en/