Robot Rights, an Approach Appealing to Animal Rights Theory

A thesis submitted in partial fulfilment of the requirements for the degree of

MASTER OF ARTS

of

RHODES UNIVERSITY

by

MURRAY DAVID MILLIN

March 2020

ABSTRACT

This thesis proposes that Peter Singer's theory of preference utilitarianism, which is designed to be universally applicable to humans and animals, can be applied to robots of a particular kind – such as those seen in Isaac Asimov's work. I shall do this by using Singer's conception of interests as a framework, and appealing to Daniel Dennett's intentional stance to deal with methodological issues about other minds. I shall then apply those theories to Isaac Asimov's *Sally* and *The Bicentennial Man*. These two narratives show the importance of the intentional stance as an ethical tool and provide an example of how we might talk about the interests of a robot. Sally's behaviour and ethical status is examined according to how she is perceived, and so I shall investigate how various persons engage with her and why they do so in those manners. This narrative demonstrates the value of the intentional and design stance as methods to approach other minds problems with regards to ethical status. *The Bicentennial Man*'s Andrew allows us to look for interests in a more concrete way. I look to see how he situates himself in his world, as well as investigate how and why he makes the demand to be morally considerable. This will be done by examining his creativity, personal development and drive for mortality throughout the narrative.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

My gratitude and thanks go to my supervisor Ward E. Jones for his insight, depth of knowledge and extraordinary patience throughout the process of writing this thesis. His input has been invaluable to this project, and his frank and honest opinion have always steered me correctly. I must also thank Keenan and Tessa for their assistance and support, which was always given in generous proportion. Finally, my mother and stepfather, for their support and love throughout this time. I could not have done this without you.

Introduction

In this thesis I will argue that Peter Singer's ethical theory includes Artificial Intelligence (AI) of a certain kind. I wish to do this by showing that interests, which Singer posits as the source of moral considerability, are possessed by certain kinds of AI. My basic argument is given below:

1. Humans and (some) animals have interests.
2. We are obligated to morally consider the interests of these creatures as they have interests.
3. Therefore, the interests of these creatures, Humans and (some) animals, obligates us to morally consider them.
4. There are possible AI that could behave in ways that indicate to us that they are interest bearing.
5. Therefore, the interests of these AI will obligate us to morally consider them.

This thesis states that if we have reason to believe that there are potential AI that are not different in any morally relevant way from other morally considerable creatures, then we have an obligation to hold to a similar set of attitudes, beliefs and actions, that we call moral behaviour, towards AI as we do humans and animals.

The line of reasoning that aims to assess animal rights theory as a means of arriving at AI rights is not unexplored, Coeckalbergh (2010) investigated the analogy of animal rights as an attempt to justify robot rights, he did this from various angles. Of particular interest to this thesis is his rejection of the bestowal of rights via analogues to animal rights, claiming that the giving of rights is a very strong moral stance to take up. In this I think he is correct. Further, Coeckalbergh makes several arguments that highlight flaws in the animal rights line of argument and shows that these flaws are perhaps even more salient when applying these arguments to AI. These are problems about the nature of the features that animal rights theorists attempt to claim are ethical properties. Although valid, I find that Coeckalbergh's arguments require too much from us as moral agents. It simply does not seem necessary to show that these features (and therefore ethical properties) are present in animals, or AI to the degree of certainty that Coeckalbergh demands if we wish to make moral claims rather than claims of legal rights. Instead, we should focus on whether AI may be the sorts of things that we should consider morally, as moral consideration would be a good reason to start to consider whether potential AI should be given legal protections.

'Morally relevant qualities' is a term that I use to describe the qualities that separate the morally considerable, and the morally inconsiderable. Morally considerable is a term that Coeckelbergh (2010) and Mark Rowlands (2009) use to describe an agent that is treated in a morally appropriate way. A stone is not a morally considerable object, nor is a flower. These are not things

that we think of as having moral status. We may protect a specific stone against harm, or a particular flower, be it the whole plant or the actual pollen bearing organ, or even species of flower, but when we do so we protect it from harm in respect of some other quality, for example some ecological impact of those plants in a specific area might protect it from damage by humans. Likewise, a stone might be protected for its historic value. In neither of these examples is it the case that we impose moral standards on the objects for the sake of the objects, but rather we do so in relation to a state of affairs that involves interests of morally considerable agents. We view the moral treatment of children in a very different way. Children are not morally respected in terms of relation. We do not advocate (at least in principle) treating only some children with moral respect, and others without. We advocate that children should be morally considerable regardless of the context. They have moral properties that make it the case that they are the sorts of things that we feel should be given moral considerability. As such, children, and humans, are objects that are morally considerable. It is an assumption of this paper that humans should be treated morally. Arguments that seek to show that our actions somehow counter this assumption will be assumed to miss the point of moral consideration, we *should* treat people morally, even if we do not do so all the time

'Morally relevant qualities' are what we might call 'moral properties'. If creature A and creature B are morally distinct objects, in that creature A is morally considerable and creature B is not, then I assume that there are one or more qualities that morally differentiates creature A from creature B. This must be the case, or else both objects would be treated with the same attitude, there must *necessarily* be some difference in some quality/qualities that make object A morally considerable and object B not. I shall attempt to clarify what moral properties are by appealing to arguments from animal rights theory. The main of which will be centred around Peter Singers' notion of an interest, which I argue is a moral property.

Singer's account situates correct moral treatment in terms of respecting interests. He argues that interests are a moral property, and that to unfairly deny those interests in a creature is to ethically wrong that creature. I shall draw on Gary Varner's (1998) theory of interests, and appeal to Tom Regan's (1983) notion of welfare to support that theory of interests. I will argue that interests are a subcategory of preferences, interests being those preferences that are within the welfare of the creature at hand.

In Chapter 1 I shall explain how Singer's preference utilitarianism is constructed, and I shall give an account of interests as a moral property. This moral property is what I will demonstrate to be applicable to Sally and Andrew in *Sally* and *The Bicentennial Man,* respectively.

In Section 1 I shall establish why preferences are important to Singer's theory. I shall do this by explaining how preferences are a moral property. A preference is a desire for some state of affairs to come about. As preferences are a moral property, it follows that they should be respected, and that

implies that we should universally respect other's preferences. Singer does not distinguish clearly between an interest and a preference beyond claiming that interests are more specifically good for us than preferences in general. I shall then address Singer's claim that consciousness is a requirement for having preferences, I will later appeal to Dennett (1998) to avoid the problem of other minds. Singer moves from respecting preferences to respecting interests specifically, rather than preferences in general. This is done as not all preferences are 'good' for a creature. Then I shall argue how we can sensibly talk about measuring moral harms that come about by denying interests unfairly. Finally, I shall argue that interests are a moral property, this explains why we do not morally consider a creature because of its physical makeup. We morally consider other creatures because of properties that they hold, which may – or may not – be related to biology of any kind. This broad ethical structure is an important ethical feature of Singer's preference utilitarianism. Therefore, it is sensible to talk about robots also having moral properties under the circumstances I will lay out in later chapters.

Section 2 will provide an account of interests that is to be used in to expand and define the notion of preference in Singer's ethical theory. First, I give a general outline of some common-sense understandings of interests and how they are ethically relevant. Singer thinks that interests are a specific kind of preference. I take this from his claims that certain preferences are in line with a creature's welfare. I shall then give a more detailed account of interests by appealing to Varner (1998). Varner's account does assume many of the same themes as Singer's. Therefore, it is a fitting theory to prop up exactly what Singer would want interests to look like. However, Varner's theory has some drawbacks. His theory posits all pursued preferences as interests which clearly cannot be the case. I appeal to Regan's (1983) notion of welfare, as being a good for the life of a creature, to distinguish further between preferences and interests.

In Chapter 2 I will show that Dennett's intentional stance is a good a good account of what is going on when we speak of other creatures as minded. I shall show that this approach is useful in that it gives good reasons to hold that some other creature is a thing that can prefer, although it does not definitively prove the existence of preferences in other creatures. First, I shall explain that we need some theory of other minds to make sense of AI having preferences, and why Singer's evolutionary and behavioural argument about animals do not suffice for showing AI to have preferences. Singer's behavioural argument is like the intentional stance, in that it derives preference from behaviour, however Singer's argument it is a shallower instantiation of that kind of theory. I shall then explain Dennett's (1998) physical, design, and intentional stances. I shall focus heavily on the design and intentional stance as they generate normative conceptions of health, and welfare in objects and creatures. The intentional stance is a method that gives good reason to assume that the creature you are observing has beliefs and desires. Preferences are a form of desires, therefore successfully taking

up the intentional stance towards specific AI will be good reason to assume them to have preferences and interests.

Chapter 3 will discuss two narratives, *The Bicentennial Man* and *Sally,* both written by Isaac Asimov (2018). The focus of this chapter will be on exploring how interests are manifested in Andrew and Sally, the most prominent non-humans in the narratives, and how effective those narratives are in portraying creatures that are morally considerable. I will show that *Sally* highlights the implications of how the mode of perception influences the attribution of interests – with regards to the intentional stance – very clearly. *The Bicentennial Man* portrays a non-human robot that seems to have interests that would qualify him for being the sort of thing that is morally considerable.

Sally and Andrew are the foremost robot characters in the narratives *Sally* and *The Bicentennial Man* respectively. Sally is an automatic AI car that is cared for by a man named Jake Folkers. Folkers is visited by a man who wishes to transplant Folkers' cars into newer models. Folkers rejects this offer on the ground that Sally would be harmed. Andrew is a humanoid robot AI who pursues personhood.

In section 1, I shall look at Sally, explaining how Sally's actions can be interpreted from the design stance and so the notion of being well-functioning or dysfunctional can be applied to her. Then I will discuss how Jacob Folkers and Raymond Gellhorn differ in how they view Sally and explain what the repercussions of those views are for the moral considerability of Sally. In section 2 I shall look at Andrew and his aim at becoming a person. I shall discuss how the intentional stance is taken up towards him, by referencing how other persons engage with him in a way that assumes that Andrew has intentional and preferential states. I shall also discuss how the intentional stance is taken up by the reader towards him, and how his creativity is a signal of personhood. I will show that taking up the intentional stance towards Andrew while denying his interests – indicating Andrew as being morally considerable – is an instance of Singer's 'speciesism'. Finally, I shall discuss Andrew's interests as morally considerable, specifically how his eventual death is an indicator of his advanced understanding of personhood.

Chapter 1: Peter Singer's Ethics, and Interests

Section 1: Singer's Preference Utilitarian Ethics.

Subsection 1: Preferences, and why they should be equally considered.

I interpret Singer's account of interests (2011) as being an account of ethical properties. The fundamental claim that Singer puts forward is that the ethical status of a creature is determined by the presence of interests in that creature. Although I am unwilling to commit to the claim that all ethical properties will be interest-based, it is certain that to say that a creature has interests is also to say that that creature has ethical properties and is therefore morally considerable. This section will explain how Singer comes to formulate his preference-based utilitarian ethics.

Singer posits that we should adopt an attitude of equality in respect to preferences. He states that "[i]f I take one of these [preferences] but can offer no reason for holding it, other than the fact that *I* prefer it… then my preference must be weighed against the contrary preferences of others." (Singer, 2011, p. 13) His argument is summarised formally stated as follows:

1. If I value my preferences and ethics must be universal, then I must value other's preferences equally to my own.
2. I value my preferences.
3. Ethics must be universal.
4. I must value other's preferences equally to my own.

The first premise states that if I value my preferences and ethics must be universal then I must value others' interests equally to my own. For this to be true it must be the case that preferences play a role in ethics. Singer's claims about ethics appear to be naturalist in origin. He does not rely on notions of the state of the soul – or virtue – to describe why we should act in an ethical sense. Instead, his argument concerning how to decide whether a given action is ethical or not is linked to what we think that action will do for ourselves and others. In this case Singer points out that a general assumption that other's preferences are, at least in principle, equally important to them as our own preferences are to us.

Singer uses the word "preference" to mean a collection of "needs, wants and desires" (Singer, 2011, p. 12) and other such like qualities. He does not spend much time on the nature of preferences, but it will be an important aspect of this paper, and as such I will begin to provide some arguments to explain the importance of them. It is clear that beliefs, coupled with desires and so forth, are going to be foundational to holding judgements about any given issue, particularly judgements about morality

and ethics. When asked whether someone supports capital punishment, we hope the response comes from a careful and logical examination of whether capital punishment fits with their sets of beliefs and desires, their preferences. If they do not produce arguments that are grounded by their preferences, we may well say that those people's arguments are not sound in terms of their beliefs or that these people are not serious in their conviction of their position. It may, of course, be possible that their preferences do not correlate well together, I may believe that slavery is a great harm but still buy clothing from clothing lines that partake in that practice; but that is a problem of inconsistency of beliefs and judgements rather than of preferences not being grounds for our moral judgements. This is also not to say that any particular person's views (formulated by preferences) are to be considered ethically 'correct' if they can validly draw out that view from their preferences, merely that when asked for justification for those moral views it is normal to proceed in this way. It is always possible that they are not thinking through the holding of any particular preference, or that their preference is ethically unsound in principle. For instance, someone may prefer that all drivers should never drive drunk, but also that they should prefer that they be an exception to that rule. In this case, that person is holding conflicting preferences, and so cannot soundly draw out ethical arguments. Further, the elitism of such a preference renders the universalisation of that preference to be moot.

It follows from the above point that our own preferences are quite important to us. Our preferences shape the way in which we make decisions in our life, and so are a part of determining how we want to live. Galen Strawson (1989) makes the argument that our preferences are one of two factors that determine what is and is not important to us. He points out that our preferences and our environmental influences[1] are the only factors that will affect the choices we make and beliefs that we hold. I will not commit myself to such a strong claim, but Strawson is correct to give weight to preferences. There are few matters to which someone can claim a serious opinion that will not be heavily influenced by their preferences. Singer explicitly states this to be true, "perhaps the only thing... that would be relevant [for choosing my action] would be how the choice ... will affect my preferences" (Singer, 2011, p. 12).

Having shown, briefly, the importance of the role that preferences play in shaping our actions and beliefs, we can proceed with Singer's argument. Singer attempts to make the considerability of preferences universal. He assumes that there are good and bad reasons for any given action, and further assumes that good reasons appeal to a wider range of preference holding creatures than bad reasons. Singer states:

[1] Strawson includes things such as genetic influences in his definition of environment.

> For instance, a justification in terms of self interests[2] will not do. When Macbeth, contemplating the death of Duncan, admits that only 'vaulting ambition' drives him to it, he is admitting that the act cannot be justified ethically. (Singer, 2011, p. 10)

The universal nature of ethics is his underlying assumption and not argued for formally, though he does argue for how that universality is applied. The following two arguments provide an outline of that argument:

1. If {A consults their preferences when A decides on what is right for A and B will consult B's preferences when deciding what is right for B} and {action x being right for A is not to say that action x is right for B} then A's preferences hold no more weight than B's.
2. A consults their preferences when A decides on what is right for A and B will consult B's preferences when deciding what is right for B.
3. Action *x* being right for A is not to say that action *x* is right for B and vice versa.
4. A's preferences hold no more weight than B's.

And:

1. If A's preferences hold no more weight than B's and vice versa then A's preferences should be respected equally to B's.
2. A's preferences hold no more weight than B's and vice versa.
3. A's preferences should be respected equally to B's.

We have already seen that our preferences will be a large influence in our decision-making processes. Singer claims that at some point we naturally begin to universalise these preferences; we come to not only recognise the preferences of others, but also to respect the sovereignty of other's preferences (Singer, 2011). This sovereignty – the realisation that the preferences held by others are equally as important to that other as our own are to ourselves – is the realisation that one's own preferences are not 'worth' more than any other preferring thing's interests are. Therefore, it follows quite clearly that one's own preferences hold no more ethical worth, or weight, than any other creature's preferences. If it is the case that both creatures have preferences, then it follows from that point that all preferences should be respected equally, at least in principle.

[2] Self-interest, in this case, indicates a lack of care for the interests of other creatures. Having a good ethical theory that involves interests will include our own interests, but also other persons.

The important point to be taken here is that Singer claims that our ethical judgements are judgements that are aimed at being generally, and universally, practiced. For instance, in some countries adult humans can drink alcohol, and they can drive (under particular circumstances such as having a license to drive etc.,) but are expected to refrain from doing both simultaneously. It is universally expected that all adults refrain from doing both at the same time because that society views not drinking and driving as being generally better, in a moral sense, than drinking and driving. One reason for this is the increasing rates at which accidents occur when a driver is under the influence, and with the increase of accidents there is also an increase in numbers of death of those involved in the accident. Thus, anyone that holds the belief that causing death is a morally dubious thing to do should also hold the belief that drinking and driving is a morally dubious action for anyone to do. Therefore, any person who holds the belief that causing death is a morally dubious thing to do should also hold the preference that nobody should drink alcohol and drive simultaneously, if they are to act morally.

Subsection 2: Singer's Interests

At this stage of Singers argument, he stops using the term 'preferences' and begins to use the term "interests" (Singer, 2011, p. 20). Singer expresses some doubts about the term preferences, his main line of argument being that sometimes our preferences might not be in our best interests. His example is that many people have the preference to win a national lottery, however research indicates that people who win such lotteries are "not … significantly happier than they were before" (Singer, 2011, p. 14). In such a case it is not clear that winning the lottery is really within any person's interests. There are also examples of counter-productive preferences such as a smoker's preference to continue smoking, or an alcoholic's preference to continue drinking. These preferences are clearly not in the best interests of those who hold that preference. As such, preferences are not always strictly good for us, or may reduce our quality of life in a substantial manner. A more thorough and robust term for the aspects that are 'best' for us are 'interests'. I shall show in Chapter 1, Section 2, that interests are a subcategory of preferences that are related to our well-being. As such I shall leave that topic for the following section and continue to explore Singer's position.

Singer states that objects require sentience or consciousness to have interests, "In the case of bivalves … a capacity for pain or any other form of consciousness seems unlikely, and if that is so, the principle of equal consideration of interests will not apply to them." (Singer, 2011, p. 60) This line of reasoning will be hotly debated, and it is not clear that it can even be solved. The problem of

other minds is a far too broad in scope for this thesis. Instead, I shall appeal to Daniel Dennett[3] to provide grounds to assume that certain non-human creatures or objects are of the sort that can be thought of as having preferences without needing to be concerned about their consciousness.

Subsection 3: Equal Consideration of Interests, the ethical weight of interests, and interest-based calculations.

As I shall explore what an interest is composed of more fully in the Chapter 1, Section 2, I shall now outline Singer's argument for the principle of equal consideration of interests. The principle is that "we give equal weight in our moral deliberations to the like interests of all those affected by our actions" (Singer, 2011, p. 20). I have shown how Singer formally explains the principle of equal consideration of preferences, in that Singer's use of the term 'interests' is merely a substitution of interests for preferences. Therefore, we can comfortably claim that A's interests should be *respected* equally to B's. In making moral judgements affecting the interests of A and of B, one must consider who has the most to lose in terms of their interests. This is clear and well, however one might wonder what this has to do with moral and ethical well-being.

Singer's utilitarian account posits suffering as an ethical harm. Suffering, for Singer, is in the denial of an interest. He states "The capacity for suffering and enjoying things is a prerequisite for having interests at all, a condition that must be satisfied before we can speak of interests in any meaningful way." (Singer, 2011, p. 50) This follows quite clearly from Singers initial assumptions about the development of ethical structures involving preferences. We act wrongly when we act as though our preferences count for more than any others, and perhaps especially wrongly when we act as though others do not have any preferences. Drawing from his quote it seems that the pair of other harms that can be done is the denial that 1) a creature has interests when it is the case that they do, and 2) a creature's interests are not relevant when considering our own. In order to cause suffering, then, at least one of these two harms must have been met, in that either a creature is acted upon without consideration for the creature's interests or a creature's interests are not considered to be important when considering our own interests. I will explore broadly how these harms manifest in Chapter 1, Section 2, and specifically in Subsection 3.

Putting aside harms, interests do appear to be comparable, or capable of being morally weighted against each other. For Singer that weight is found in the extent of suffering caused by the denial of interests. His example being "X's pain might be more undesirable than Y's pain because it is more painful" (Singer, 2011, p. 20). A concrete example is that it is almost always within my

[3] See Chapter 2

interests to avoid pain. It is clear that anyone that inflicts pain upon me is acting against my interest, but my interests might come to be at odds. If I were to have intense chronic pain of some kind, as sometimes happens to knees, necks, or spines, it would be within my interests to find a way to reduce that pain. Perhaps I might be able to have a surgery that would prevent pain in the future, but would have to undergo extremely painful rehabilitation therapy, and post-operative pain for some weeks or months. In this situation my interest in avoiding pain will be under tension and either action would result in suffering of some kind. The obvious solution to this problem is balancing these interests in some way. I must evaluate the quality or kind of suffering I will endure in each scenario, having lifelong pain or shorter-term but more intense pain with a probability of relative painlessness in the future. In this example Singer would be unhappy to call the surgery a morally harmful action, even though I would be suffering at someone else's hand (the surgeon or the physical therapist). It would be in my interests to minimize the total possible suffering. Therefore, I should take the action that would cause me the least possible suffering.

Singer states that moral weight is determined by attributes associated with suffering, (Singer, 2011). Attributes such as intensity, length, and how much one creature stands to lose as opposed to another are good examples. That is not an exhaustive list, but it gives an indication of the kinds of attributes that are to be considered. In minimizing suffering these attributes must be accounted for, especially when minimizing suffering between two or more objects that are suffering. Singer makes some interesting points about this aspect of his theory, particularly how in some cases particular interests should be given more weight than others. For example, in the immediate aftermath of some natural disaster, leaving many in a state of suffering, it would be wise to privilege the reduction of suffering of those who have some medical training so that those people could then possibly help reduce the suffering of others. In this case the privileging of those with medical training is in line with the aim of generally reducing suffering. This careful weighing of interests is not dependent on the kind of thing that holds the interest.

Weighing a creature's interests against another creature's interests has only to do with the kind of interests being held by each creature. Above I have shown that preferences, and therefore interests, are what are measured against each other in moral judgements. Above I have also shown that Singer claims it is a mistake to claim that a human's interests are more weighted than an animal's merely because one is a human and the other an animal, or vice versa. The overall reduction of *neglect of interests* is the prime directive of preference utilitarianism, not of *human specific* interests, or zebra specific interests. Therefore, to morally weigh one kind of creature's interests more heavily than another's is to commit a moral wrong. Singer refers to this kind of wrong as a case of "speciesism" (Singer, 2011, p. 48), as a form of bigotry akin to racism or sexism. I will not discuss whether

speciesism is as odious either racism or sexism, but I will point out that under Singer's theory it does follow that speciesism is a morally dubious position to hold.

Subsection 4: Ethical properties

Singer's account provides a simple locus for ethical consideration, a simple moral property. In having interests, one is the sort of thing that is morally considerable. The account of an interest-based ethics is really an account of interests as a moral property, moral properties being those features that we use to support our claims on the duties of others. Singer's account is an attempt to defend the notion that interests are moral properties, in that the interest-bearing creature is to be considered a morally considerable creature. That is not to say that all interests should be weighed equally against one another, but that they are all weighed as interests. The interests that a dog has in being provided food is not less weighty than my own simply because it is a dog, and I am a human. Conversely my interest in being fed copious amounts of chocolate (if such a thing is an interest) is not weighted equally to a dog's interest in having access to food in general. So, I mean to say that we should weigh interests as interests, but we should also be aware that some interests are ethically more weighty than others. I shall not go to far into this point, although it is quite interesting. This thesis concerns itself primarily with a very human-like example, and so questions about whether dogs can suffer from expectations in the future are not relevant enough to pursue.[4]

Moral properties are those things that differentiate between morally considerable objects and morally inconsiderable objects. This claim rests on the assumption that there are certain classes of objects in our world that we would want to treat in moral ways – these we refer to as morally considerable and morally inconsiderable objects. Given that we believe that there are morally considerable and morally inconsiderable objects, and we believe that classifying objects into morally considerable and morally inconsiderable is rational and grounded in the objects in some way, then there must be some difference(s) between these two categories of objects. These differences are moral properties.

We require reasons for saying A is an object of moral consideration because of reason M, where M is a property of object A. The reason we consider property M a moral property is explored by Singer, Varner and Regan in Chapter 1, Section 2. However, property M is somehow tied to object A. I do not mean to imply some particular metaphysics about the nature or source of property, I am

[4] By this I mean that it might be the case that human interests in eating could be more weighted by the concern that I will not have food tomorrow. If a dog cannot have that concern, then its interests in having access to food might not be as weighted as a human's. Alternatively, it might be the case that these are two entirely different interests to do with food and are therefore not comparable in the way described.

not too concerned as to whether property *y* truly is a necessary aspect of object A or is accidental or any other statement of that kind. I simply mean to say that when referring to object A with the intention to claim it morally in/considerable we must point to some feature that sits in it, in some sense of that phrase.

Moreover, the evaluation of objects as morally considerable should be rational and consistent, in the sense that all objects that bear property M should be morally considerable. If object A and object B both bear feature M and there is some moral discrepancy between objects bearing the same feature M, where it is claimed that feature M is a moral property, then feature M is not the feature that bears moral weight for that statement of differentiation. If there is no reason for saying that object A is morally considerable, and object B is not, then the ethical theory is defective. Singer argues, and I agree with his assertion, that we must be accountable for why we say certain objects are morally considerable and others are not (Singer, 2011). It is a core feature of ethics that moral considerability be rationally supported.

Moral properties come in degrees. It might be immoral to take a friend's book without asking and to never return it, that is theft. However, it does not seem immoral to throw out, or give away, old toys that no longer suit the tastes our toddlers. Likewise, there are things that would certainly be immoral if we did them to our friends or our toddlers e.g., intentionally starving either would be a deep violation of our moral duties towards both. In these cases, both 'objects' (our friends and our toddlers) are morally considerable, they share some moral properties, but they do not share all of those properties. That is not to say that one is *more* morally considerable than the other. Rather that when morally considering we must take care to consider them according to their moral properties.

Singer claims that interests are a moral property. The account of interests that I have provided is an account of a particular kind of moral property, one that appeals to welfare and preference. In Chapter 3, Section 1, I discuss how Asimov's Sally might or might not have interests depending on the stance taken up towards her. The question at hand when discussing Sally is does she prefer or is she the sort of thing that acts out her designed features. As I will demonstrate, in the first case she is a morally considerable creature, in the second she is merely a tool and not morally considerable.

Section 2: An Account of Interests

Subsection 1: Singer's approach to preferences

This section will show that an interest is a preference that leads to the benefit of the creature. Therefore, interests are a subset of preferences. The distinction between having a preference-interest and a mere preference is that an interest is a preference that is aimed at the welfare of the creature.

To be an interested creature is to be either of the following: being *interested in* something or something *being within one's* intersts (Frey, 1977). I shall refer to the being 'interested in' kind as interests$_1$ and to the 'within one's interests' kind as interests$_2$. The first meaning of 'interest' is to be interested$_1$ in, in that my attention is concentrated on an object of interest. This interest$_1$, or attention, does not need to be occurring at one specific moment (Frey, 1977). I may be interested$_1$ in my food tonight, having my attention drawn to my food at that time. I may also be generally interested$_1$ in food as a chef, not being interested$_1$ at all times, but in being willing to have my attention drawn towards food over a long period of time.

The second meaning of the term 'interest', interest$_2$, is the claim that there are things that are important to a creature's welfare in general (Frey, 1977). Some examples of this might be a person's need for food or water, both are within the interests$_2$ of a person. Being interested$_2$ implies some kind of good-for-being; water and food are both clearly vital for the continued welfare of a person. Although the term interested$_2$ is often the intended manner of reading 'interests', it also often happens that the first meaning of the term is implicitly tied up in the notion of 'interests' as well. This is the case with Singer's notion of interests. An account of interests that covers both meanings of the term, as far as Singer's ethics is concerned, will be a better account of interests than a theory that covers only one account. However, I will first explain how Singer posits preferences and interests which relies on an interests$_1$ sense of interests.

Singer claims that he uses the term 'preferences' to explain that being interested stems from having preferences. Singer also claims that the term 'preferences' is useful as "[w]e all know what preferences are" (Singer, 2011, p. 14). However, it is not clear that we do all have intuitive knowledge about what it means to prefer, nor is it clear how preferences are related to interests. To prefer X is to want that X occurs rather than not. In humans this is initially uncomplicated as a phenomenon but becomes more complicated when considering other kinds of animals. I can clearly see that my friend Joey prefers peanut butter to strawberry jam, in that he will always act such that when he eats bread it is covered in peanut butter rather than jam. Joey's interests are also fairly easy to determine in that he can declare them to me, or I can assume shared notions of value for things like water and food. I can assume that if I have an interest in having adequate amounts of water supplied to me, Joey would also share this interest. I can see something similar with my dog, Erin, in that she prefers steak to kibble. If Erin could, I am sure she would arrange things such that she always ate steak over kibble. But it is not clear that preferring to eat steak, rather than kibble is in Erin's interests. Singer does not provide a great deal of explanation other than the implication, never formally stated, that some preferences are interests.

We can be reasonably certain that Singer does think of interests as being closely related to preferences. Singer's theoretical framework relies on having preferences, and having preferences is

implied to be important to having interests. In Singer's ethical theory there is a clear distinction between things that do prefer, and have interests, and things that do not prefer and so do not have interests. This is because not all things are morally considerable. Given that Singer's use of the term interests depends upon having preferences, his use of the term 'interests' corresponds to the first meaning of the term, interests$_1$. As such, things that we would usually consider having interests, but not preferring, are not going to be things that are morally considerable. Things of this nature are plants and hammers and so on.

Singer then provides arguments for the case of animal preferences. Singer argues that certain animals likely have interests due to shared biological history with humans, that "the … nervous system evolved in more distant ancestors and so is common to all of the other 'higher' animals, including humans. This … makes it likely that the capacity of vertebrate animals to feel is similar to our own" (Singer, 2011, p. 60). This is good evidence that these 'higher' animals likely experience a range of physical sensations, and therefore preferences, in a similar way to humans. As animals have preferences, they will also have interests and so should be morally considerable. However, his development of preferences to interests is not made clear beyond claiming that interests are somehow 'better' for our happiness than preferences.

Both Regan and Varner provide arguments as to why this crude preference-based conception of interests should not be the whole of the case. Varner attempts to sketch out a more detailed theory of preference-based interests that appeals to a broad definition of desire. Regan provides a differing notion of what an interest stems from, although his theory also appeals to the idea that to be interested is to be desirous of some state of affairs, and his claim includes a Singer-like notion of a good in itself. Both these interpretations have a strong correlation with how Singer has situated his theory of preference. Both Varner and Regan think of interest as including an internal drive to move from a state of not Y to a state of Y, that indicates a preference for the state of affairs being the case that Y. Given this strong correlation I think it is fair to use these two theories to bolster Singer's argument by giving a fuller account of interests.

Subsection 2: Varner

In claiming "The paradigm case of an interested being is a desiring creature" (Varner, 1998, p. 26), Varner states that interests come about from being desirous of some state. I shall use the term desire, in this section, to refer to all species of wanting. If I wish to speak about the specific phenomenon of 'desire' I shall use the capitalised form of the word Desire. All species of wanting include notions such as 'brute wants', needs, desires, longings and so on. Included in this broad form of desire are preferences. Varner gives a formal definition of being desirous composing of three parts:

1) A is disposed to pursue X.
2) A pursues X in the way it does because A previously engaged or concurrently engages in practical reasoning about X or objects like X, where engaging in practical reasoning includes both drawing inferences from beliefs of the form "Y is a means to X", and hypothesis testing by which such beliefs are required and revised.
3) This practical reasoning is at least potentially conscious. (Varner, 1998, p. 28)

The first criterion situates desire as something that manifests itself through behaviour rather than something that happens to a creature. A is disposed to pursue X, and so these criteria do not strictly require that all desires be pursued, merely that if there were a path of pursuit for X the creature would attempt to pursue X. The implication of desire being action orientated and interests being derived from desire is that interests are also action orientated, in that it is what we do that makes something interest bearing.

This does present some problems despite Varner's qualification "disposed". One such problem is that of persons who can no longer be disposed towards action, such as persons who are in vegetative states due to injury or disease. Varner's position is that in such cases that if A can pursue X, they shall pursue X. In one sense of that statement, call it a generous interpretation, Varner's claims can be rephrased as 'in the best-case scenario, wherein our disposition to pursue X is clear, we have a viable manner of obtaining said desire, and we are capable of acting on the said desire, we would pursue X'. However, the 'is' in criterion 1 implies something stronger, that we must be disposed to pursue X, which is not the case in many vegetative states, where there is no disposition at all. I am no expert in vegetative conditions, but I take it at least some vegetative states come undone and people return to the waking world, as does happen to some persons in comatose states. In such a case it seems clear that criteria 1 still holds as they are now disposed to pursue X. However, there are some imagined scenarios wherein it is certain, for whatever reasons the reader can conjure, that the subject in a vegetative state will never return to consciousness; in such a case they can never have a disposition to pursue any X. However, such a state requires that the body be given nutrition through intravenous methods, or something to that effect, and it is clear that someone still has an interest in being sustained in a vegetative state. We would think it cruel if we simply starved such persons to death, rather than ending their suffering outright. This might also be against the interests of the vegetative patients, that is not clear[5], but to starve them while maintaining their life support system seems to be needlessly

[5] Regan's sense of welfare in the next subsection does account for the interests of people in vegetative states.

cruel. In this scenario criterion 1 seems to take away from the credence that interests derive from desires, and therefore preferences.

The second criterion is perhaps more convincing, if we assume that the first criterion holds water. It is convincing as it excludes many things that we would not consider desirous such as plants and tools, I do not think it is a contentious claim to state that plants are not the sorts of things that desire.

The division between animals-as-preferring and plants-as-not-preferring is a fair basic normative assumption and we should aim to develop a view in line with these basic normative assumptions about interests[6], even if the possibility of extending the greater normative conceptions about interests is left open for the moment. The evolutionary argument made in Chapter 1, Section 2, Subsection 1, by Singer gives good reason to extend our normative conceptions of interests (as a vague concept related to biology rather than Varner or Regan's specific theories of interests) to include animals of the kind that have similar nervous systems, due to shared evolutionary heritage. However, the basic assumption that interests have to do with how the nervous system functions does not extend our normative conceptions about what things have interests beyond animals of a particular kind, 'higher' animals, as such organisms such as plants and insects are excluded.

The second criterion argues that a particular kind of mental capacity for change is required to separate desire states from mere autonomous actions. A mental capacity explanation that appeals to reflexive action is not sufficient to explain how the desirous thing 'aims' (in the deliberate sense, as opposed to the habitual or reflexive sense) at producing the desired state. This kind of mental capacity seems to be tied up in being able to learn, or be trained, in particular ways (Varner, 1998). However, it is not enough that animals should be trained to react to certain stimulus. When that stimulus is changed the animal must be capable of making some change to their behaviour that will consistently bring about the desired state, without dedicated intervention by external participants. Varner provides examples of this by pointing out that Martin Bitterman's (1965) experiments on mammals, fowl, fish, and reptiles show that fish and reptiles struggle to abstractedly learn to pursue objectives, suggesting that their behaviour is better explained by reflex or habit; whereas in mammals and birds the behaviour is much easier to train without direct intervention by experimenters, suggesting abstract learning.

The distinction of action as reflex and action as deliberate also ties into the third criterion. Varner posits the notion that consciousness is required for desires. By this Varner means that there must be deliberate action taken towards the fulfilment of the desire, and not merely action taken by automation. This is again an attempt to distinguish between things like plants from animals like

[6] That interests are important for ethical treatment of persons but not of plants, assuming plants to have interests which is not the case under Varner, as they are not capable of pursuit.

humans. Varner's reasoning for this seems to be something like Singer's in that Varner assumes that interests are specific to things that have nervous systems of the same sort that humans do. That seems to me to have to do with the broadening of the idea of interests from a normative conception of interests about humans rather than a reformulation of prescriptive interests in general. Put simply, Singer and Varner both claim that humans have interests due to their nervous systems (Varner, 1998), thus anything else with a nervous system (nervous system here referring to organs such as the brain, as well as the systems along which movement, pain and other kinds of information of this sort are transmitted) of a similar kind must have interests as well. Varner basis his notion of interests on the notion of desire, working backwards he seems to think that desire requires a nervous system of a special kind as well. I think this is in the sense of desire being an active wanting of some mental state of pain or pleasure. That pain, or pleasure, is described as coming from the nervous system, as opposed to desire being abstractedly related to pleasures, such as walking on the moon.[7] However this approach begs the question as to whether interests are derived from nervous systems at all.

It seems that at least some interests are. It is clear, at the very least, that interests that arise from the notion of mental states of pain and pleasure must necessarily come about due to the presence of the capacity to feel pain or pleasure – that is to say, in having a nervous system of a certain kind. Humans having the experience of pain or pleasure seems to be inextricably linked with having nervous systems a) that can generate the necessary electrical-chemical signals that act on said nervous systems, and b) that produce some kind of mental state phenomena related to those electro-chemical signals. It is not immediately clear that any object with something resembling a nervous system should also experience pain and pleasure in this way; Varner points out that there might be a difference in the body's response to pain in the terms 'pain' and "nociception" (Varner, 1998, p. 51). Nociception defined by Varner as the nervous systems reaction to pain stimulus without exhibiting pain phenomena (Varner, 1998). Some creature could then have a nervous system that 'processes' pain stimulus but does not have capacity to exhibit that pain as phenomena. The implication seems to be that it is the mental state of pain and pleasure that would lead to having interests related to those phenomena, whereas merely having nociception-like reactions is not sufficient for having interests. That being said those interests that stem from pleasure and pain are dependent on the nervous system being of a particular kind, one that can produce the mental state of pain or pleasure. This is only part of the story, we would not perform cruel experiments, say jabbing needles into the feet of someone

[7] To be sure, if my desire were to walk on the moon, I would feel pleasure if I were able to do it. It might also be extremely uncomfortable, the trip to the moon cramped with equipment and other persons, and I might be terrified the entire way and the entire duration of the walk. In this sense the pleasure of the moon walking is more abstract than the pleasure of drinking when very thirsty.

that cannot feel their feet, on unwilling persons merely because they could not feel the needles. In such a case there are higher order desires at work, for instance that of bodily autonomy.

Assuming that Varner and Singer are right that desire originates from the nervous system in some way, it follows (from the same reasoning as has been demonstrated for pain) that whatever shares that nervous system structure will also share the capacity for desire. This answer is touched on again below in the section on Regan, but I will be employing Dennett's theory of the intentional stance to show a more sensible approach to determining mental states in a foreign body or system. However, it is clear that desire does require some specific nervous system schematics, as desire is a function of the brain, whatever else is included in those requirements. Varner seems happy to conclude that at least some other animals can desire, in the broad sense of the word, with particular focus on animal groups such as mammals.

It is also clear that this approach is not sufficient for AI, as it would be unclear that they have any 'organs' that correlates well with human or animal nervous systems. However, the point is moot; the overall theory, that desire states require being deliberate as opposed to reflexive, still holds weight even if this approach is not applicable to AI in particular.

While the third criterion works well for defining the requirement for conscious awareness of the desire, when looking at interests as originating from desire it seems to fall short. Varner claims something like 'to suffer (in the sense that I suffer when I have broken a bone) I must have conscious awareness of that suffering[; i]f I am not conscious of it, I do not suffer.' Clearly, I can have interests that I do not desire, and I can consciously desire things not in my interests. The first case is one wherein I may have a deficiency in some sort of vitamin, I have a clear interest in not being deficient in that vitamin, but I do not desire vitamin supplements as I am unaware of that deficiency. The second case is one wherein I may desire to buy an expensive television unit, despite being in a financially precarious position. In such a case a new expensive television unit is not in my interests even were I to desire it. Neither of these kinds of interests seem to be explained by Varner's account. In this way the theory organises all desires that are pursued as an interest, which Singer will not accept.

Varner's account, although flawed, is useful as an expansion to the account that Singer produces based on preferences. Preferences seem to be intimately related to wanting, falling broadly under Varner's notion of desire. As such, Varner's account helps to elucidate the issue at hand, and gives a more nuanced account of Singer's notion of preferences. Although Varner's account is limited, I think that it could be improved by appealing to Regan's account of interests and using the two as a guideline for distinguishing between the notions of preferences and interests.

Subsection 3: Regan

Regan gives an account of interests that appeals to welfare rather than merely to an object's desire. Regan (1983) makes the distinction between having interests and being morally considerable; this is because he claims that it is sensible to talk about things, such as cars and plants as having interests. In these cases, plants and cars, we do not want to talk about them as being morally considerable for their own sake. Only in the case that objects are shown to have interests and are preferring, in that they have a welfare and would prefer to have that welfare fulfilled, can we sensibly talk about what our obligations towards such things are. I will not explore what Regan thinks needs to be added to interests to generate moral obligations, as it is the notion of interest-based preference utilitarianism, rather than Regan's ethical obligation, that is of concern to this thesis.

Regan defines two varieties of interests in animals, the first are preference-interests, the second are welfare-interests (Regan, The Case for Animal Rights, 1983). Preference-interests are those interests that have to do with what creatures (animals, people, or otherwise) want from the world, what they pursue and what they aim to achieve. This could be episodic, it may be that I desire a beer tonight, and having drunk my beer I no longer desire one. Alternatively, they may be chronic (not his term), in that I may describe myself as someone generally interested in beer. In the first case I describe a short-term desire, in the second a true statement about my general disposition. These kinds of interests seem to be similar to how Varner, and Frey (1977), theorise interests to be.

It is clear that desire is an important aspect of preference-interests, Regan spends some time arguing that the notion of animal desire is a sensible one, although for the purposes of this section it is not essential so I shall address it below. Regan does not seem overly concerned with the structure of desire, instead arguing -in the same vein as Singer, see Subsection 1 above- that the evolutionary history of mammalian animals implies that such creatures act based on their desires. Whether or not this is true is not important to this thesis, appeals to Dennett will show that mammalian animal behaviour is best understood when seen from the intentional stance. As such these animals should be assumed to be intentional, assumed that they act in such a way as to achieve intentions, implying desire[8]. Further, this thesis is concerned with AI, and only partly with animals, and it is intentional stance theory that will show that objects like AI have states of desire (or that we should treat them as such none-the-less), not evolutionary theory.

Regan appeals to the "cumulative argument" (Regan, The Case for Animal Rights, 1983, p. 29); in structure this is similar to Singer's evolutionary argument of how animals should be thought of as having the same feelings of pleasure and pain as our own[9]. Regan adds that the explanation of

[8] See Chapter 2, Section 4
[9] See Chapter 1, Section 2, Subsection 1

behaviour by appealing to desires is much more satisfying (in that it can explain far more behaviour) than mechanistic arguments. This approach, again, is better understood by looking at the intentional stance. As far as animal desire goes, it is sufficient for the argument, as will be applied to AI, that we accept that animals, at least mammalian animals, can desire. It is sufficient that if animals can desire then to deny them the capacity of interests, or deny them moral status despite having interests, is to commit speciesism[10]. The same argument will be shown to apply to AI.

Welfare-interests are those things that are 'good' for the object; here Regan points out that objects can be benefited or harmed in terms of what is within their welfare-interests. Benefits being those things that broaden the possibility of living in a "good life" (Regan, The Case for Animal Rights, 1983, p. 88) sense. Conversely, taking away of benefits leads to a harm, harms being either a deprivation or an infliction. I shall explain the harms below, but first I shall discuss the notion of a 'good' for being.

What exactly constitutes a good life is only partially explored. A good life is the best kind of life that a thing can have, relative to its capacities and the broader outlines of its life e.g., the kind of society it lives in, what resources are available to it etc., but what constitutes the 'best' kind of thing for any object[11] is not fully explained. I shall leave the issue of the 'best' kind of life, and what that could mean for a moment to first describe benefits and harms more carefully.

Benefits in humans include, according to Regan, a degree of education relevant to the sorts of life that human will live, adequate nutrition, having access to certain material goods and the like (Regan, The Case for Animal Rights, 1983). He explains that the form of education will depend on where this person may live e.g., as a western focussed education may not be useful to a villager in remote Tibet. More general benefits that Regan thinks applies to most persons are things such as having access to water and food, shelter, social interactions and so on; lacking these sorts of things will often result in a narrowing of a person's opportunities to live well.

This narrowing of a person's opportunities is referred to as a harm of deprivation (Regan, The Case for Animal Rights, 1983). Regan gives the example of a father that gambles away his fortunes instead of providing care and education to his children, and thus the children are deprived of the benefit of that fortune. This can, perhaps, be seen as an inverse to benefits in general. In being harmed in this way, something that was to the benefit to that person is now taken away, decreasing the possibility of being able to live a good life.[12]

[10] See Chapter 1, Section 1, Subsection 3
[11] Regan thinks that (for example) a drill can be treated in variety of ways, one of which is 'best' for the drill.
[12] In Chapter 2, Section 2, I investigate how Andrew (an AI) draws interests from his welfare and pursues his own death. His death, he claims, is less of a harm to him , comparatively, than the his benefit in being a person.

The second variation of a harm is the infliction, or to induce suffering (Regan, The Case for Animal Rights, 1983). Singers sees the capacity to suffer as a prime indication of interests, and Varner also clearly posits desire as central to interests; but Regan seems to think that to suffer is only a particular element of the broader notion of harm. Regan states that to suffer is to experience a "prolonged pain of considerable intensity" (Regan, The Case for Animal Rights, 1983, p. 94). Regan does this to distinguish between suffering and mere pain. Although pain can sometimes become suffering, a person that is given a small static shock when touching a metal shelf is not suffering in the same way as someone with a chronic pain in their back. Regan then claims that this distinction between pain and suffering is a central notion – arguing that most people would agree that animals feel pain (Regan, The Case for Animal Rights, 1983). However, if animals can suffer it implies other important notions about them that are morally relevant. This distinction leads to ethical notions that will not be explored here[13]. In Chapter 3, Section 1, I investigate how Sally's (an AI) ethical status to Folkers (her caretaker) is established. The suffering that Sally would 'experience' is a powerfully motivating factor against her handler Folkers' agreeing to Gellhorn's deal.

Benefits and harms are useful for designating the distinction between Varner's interests and preferences. A simple explanation would be that those preferences that lead to a benefit are interests, as benefits derive from welfare. Alternatively, interests are those preferences that are within one's welfare. Benefits are metaphysically murky, relying on a difficult to define notion of an object's goodness. I have spent some time below highlighting some of these issues and providing a possible solution. However, it is worth noting that this approach, harms and benefits, shows a more distinct difference between interests and preferences than Varner's. Varner's account leads to making all preferences interests when pursued. This is plainly not the case, as I can have preferences that are not in my interests, such as preferring to buy new furniture, rather than paying my rent. It is not to my benefit to have new furniture but no home, and so it is not within my interests.

In this formulation it is still necessary that I prefer to rent a house for my having an interest in renting a house, and so the scope of having the moral property called an interest is still limited. Given both Varner and Regan's account, it is still unclear how I might have an interest in something of which I am not aware. One explanation might be that it would be in my interests to take up some particular benefit or avoid some harm if I were to know of them. This account is still very unsatisfying, as human infants surely have interests that they are not aware of and can surely be benefitted by certain routines such as playdates. In such a case it is unclear that they can even become aware of their

[13] These ethical notions are not important to this thesis, to be sure Regan's theory may well lead to similar outcomes as taking up Singer's theory. However, Singer's theory is far simpler and more easily applied to a wide variety of objects. The very presence of interests in a creature is sufficient reason to act with moral considerability towards that creature under Singer.

interests. They also have interests they are aware of, such as a desire to eat when hungry, but more complex interests (such as anxiety about the future, or the dread of inevitable death) are beyond their capacities. Regan's theory of animal rights might give an account if explained in full, but it is Singer with whom we are concerned. As such I leave the question unanswered. To my mind preference theory does a great deal of work towards a good reason to morally consider creatures with interests, even if there are some inadequacies. I shall now discuss what it means to talk about the goodness of an object according to Regan.

We all have a common sense understanding of what a good life *is not* (at least there are some kinds of living that we should not wish anyone to endure) even if that common-sense understanding does not outline what a good life *is.* For example, I may not wish anyone to be an addict, despite having no positive definition of a good life. Further, there does seem to be a sense in which a plant itself is made better by giving it adequate nutrition and water, and the same could be said of a dog, or a human. They are made better by being healthy. However, the question of what this 'better' or 'healthy' means, and how we should come to know these meanings remains largely unexplained.

In his response to Feinberg, Regan states that the source of the goodness of being (he specifically refers to the good for a plant) is not to be found in what we wish to get from the plant, or what we want the plant to do. It is found in the plant itself: "Good gardenias are desired because they are good; they do not become good by being desired" (Regan T. , 1976, p. 495). However, this does not seem to be true, to some degree we see the health of the plant as good for us, or at the very least we see a particular state of the plant as being 'better' than another. There is clearly an element of projection of goodness unto the plant in either case. Therefore, Regan's account raises difficult metaphysical issues about goodness in an object. It is unclear how Regan intended to avoid or explain these metaphysical issues, he specifically leaves it to "Aristotle and Aquinas" (1976, p. 495), but he seems satisfied that we can all agree that the plant must have some circumstances wherein it gets all the nutrients and space it requires to become 'good', whatever that good entails

It seems that a possible solution is to give up the idea of any creature having a metaphysical good in the way that Regan argues to Feinberg. It is enough to consider normative cultural values about goodness when dealing with the goods of things, if you are consistent with the kinds of things you are applying these normative values too. To ask if supplying a dog with adequate amounts of water is good for the dog, knowing what we do about dog physiology and our assumptions about dog preferences, it would be very strange to claim that it is not good, or neither good nor bad, for the dog. Water is plainly good for the dog, it makes the dog healthy. The same holds for a plant, but the locus of that 'goodness' must come from the observer. This answer derives from holding a good reason for the belief that dogs prefer water, which Singer and Varner both give, and that a dog needs water for

its body to adequately perform its functions. A similar argument can be made for the plant. A plant is at its 'best', in terms of our normative cultural values, when in particular states of growth.

I do not wish to imply that someone must be thinking about a creature for that creature to have a good, or welfare, that is plainly not the case. I argue that we can come to know something of that welfare by inspection and by making some assumptions about whether that creature has mental states that we would think of as 'preferring'. This is a simple answer, and loaded with its own problems, but for the purposes of this thesis I think it will suffice despite Regan's claim that "they do not become good by being desired." (1976, p. 495) I think this approach is sufficient for generating 'goodness values' about the state of an object. When dealing with things like animals and humans, things we consider to be an intentional thing[14], it works to describe how preference-interests can be differentiated from preferences in general. Under this conception an animal has as much claim to having a welfare as a human might, although far less subtly, as they do not have language, and so cannot finely distinguish their desires. In *The Bicentennial Man,* Andrew can make claims as to what his good is, and so we can take that seriously. Sally is the more interesting case as her good is not clearly defined by her, in the same way that animals do not define what is good for them.

A useful outcome from welfare-interests is the notion that things such as tools and plants can have interests, but not be morally considerable. I shall explore this notion further in Chapter 2, Section 2 but the claim is that water and nutrients are within the welfare of plants and oil and other lubricants are within the welfare of hinged workman's tools. Neither of these objects are things we want to think of as morally considerable, and Singer's theory will not include them. These things cannot prefer, and so do not have preference-interests. However, they do have a value in being kept a particular way. This is useful because there is a sense of the term "interests" that is applicable to things such as cars, plants, ovens and the like; however, they cannot be morally harmed in the same way that Singer talks about morally harming a creature. For moral harm to be done a preference for a particular benefit must be ignored or undervalued. Plants and tools can be diminished, they can be damaged so that they no longer function correctly – not providing water to plants or leaving iron tools in salt water will be a deprivation of some kind – but without a preference-interest to be denied, it is not a moral harm.

Regan's theory, when considered with Varner's theory provides a good working account of interests. The main advantage to using both theories is to cover both meanings of the term interests, both interests$_1$ and interests$_2$. Varner's account is something close enough to Singer's envisioning of his ethical theory to move from preferences to interests. Regan's account helps to fill in some of the issues that an action-based theory leads to, specifically how something can be a preference that is

[14] See Chapter 3, Section 1, Subsection 4

pursued but is not an interest. Pursuit is still an important aspect of interests, but the sense that interests are also something that have to do with being beneficial to a general well-being is covered. In this thesis I wish to proceed with the notion that preferences that lead to benefits, where benefits are those things that lead to well-being, are interests. As Singer discards the term preferences for interests but maintains that his theory is a utilitarian preference theory, I can only assume that interests are drawn from preferences and that this approach covers at least the large part of what we think interests should be.

Chapter Conclusion

Section 1 discusses Singer's utilitarian theory of ethics as being a theory that considers the weights of interests as being a moral property. I have shown that preferences and interests are linked to each other, and that equally considering them is a bedrock for the preference utilitarian ethical theory. I then explained how interests can be weighed against each other – and why that weighing is not morally dubious – by explaining that weighing interests is done by weighing interests specifically and not weighing interests based on what kind of thing is holding the interest. To weigh interests according to the kind of thing holding the interest is to commit an act of speciesism.

Section 2 discusses the notion of interests in detail. Singer's notion of interests is not explained enough to be of value in this thesis. As such I have drawn on Varner and Regan to a) provide a more detailed version of action-based interests; and b) provide a motivation for the distinction of preferences from interests. I do this by appealing to Regan's notion of welfare.

Chapter 2: The Intentional Stance and Non-Human Interests

Section 1: Other Minds and Preferences

Having outlined Singer's argument leading to non-human-centric ethical consideration in Chapter 1, the question arises as to how we should sensibly talk about AI having interests. In Chapter 1, **Section 2: An Account of Interests**, Subsection 1, I explained the two notions of having an interest. The first notion was that of being interested in something, and the second of something being good in a welfare sense. Singer's theory of interests is predicated on the notion of being a thing that prefers, and so this first sense of the term interests – that of being interested in – is a key aspect of being an interested creature (in the moral sense). Singer provides some arguments for why we should believe that animals are the sorts of things that can prefer in this way, and Regan and Varner make other arguments aimed at the same end. Specifically, Singer and Regan argue that there is evidence that can be drawn from evolutionary facts that makes it extremely likely that certain animals share mental capacities with humans, and likely the capacity to prefer is among those capacities; both also appeal to behavioural evidence.

There are a number of issues with the position predicated on drawing evidence from evolutionary facts for being a preferring creature. Firstly, it is difficult to apply the preference-interests account to non-biological things. Secondly, evolutionary arguments only do partial work. There is no necessary reason as to why AI should be the sorts of things that share our evolutionary history. It is sensible to talk about sharing evolutionary history with a gibbon in that a gibbon has (relatively) recent shared evolutionary ancestors with humans. It does not make the same kind of sense to describe a computer as being related to humans in the same way. A gibbon has an unbroken biological history that links the gibbon to any human by means of that evolutionary history. This unbroken evolutionary history suggests that a gibbon is of the same kind of thing as a human, although of a different variety[15]. We are connected by our biology in a way that is not the case with a computer. However, it is not clear from the fact that we have no shared biological history that we should treat a computer in an ethically inconsiderate manner. Shared evolutionary history is not itself a moral property.

One could describe a computer as being created by humans and therefore being related to them in a different way, in that humans have given the computer certain capacities that require certain

[15] I am not using the terms 'kind' and 'sort' in any conventional biological sense, this usage merely denotes that gibbons are biological creatures, as are humans; but not so different that they are of a different 'kind' of thing, merely of a different 'sort' of biological creature. Computers are not biological creatures and so they are a different 'kind' of thing, with their own hierarchy of 'sorts'.

'needs' be met. In an abstract sense we pass notions of need and desire onto artifacts. For instance, a computer requires electric power to perform its computations, that is something that engineers have organised the computer to need. Ruth Millikan describes a theory of representation based on the "histories of the items possessing [those histories]" (Millikan, 1995, p. 86). This approach is interesting and may describe functions as being interest bearing, where a function is defined by its historical use, or the desire for the product of that function. For instance, a heart has the function of pumping blood throughout the body and is excellent at being a heart when it performs this behaviour well. The standard by which we can define that function is found in the historical behaviour of the heart over a long period of evolutionary history, hearts have always pumped blood around the heart and the system that employs a heart tend to survive better when the heart is in this 'excellent condition'. However, this approach's usefulness is limited. It does not take us from defined and well-known capacities of mental states to artificial mental states in the same way that biological arguments would. That is not to say that Millikan does not think her argument applies to artefactual organs, but instead that the history which she attaches to those artefactual organs does not carry over to artefactual organs in the same way that a biological organ's history is carried over. I shall not pursue this further as it does not result in evidence for a mental state driven conception of preference; although it may be valuable in that it gives us reason to talk about the computer having interests in the second sense of the term used in Chapter 1, Section 2, that is, having an interest in electrical power by electrical power being good for it. For Singer, this approach will not suffice.

Secondly, evolutionary arguments only do partial work, and they need further evidence to substantiate those arguments. A clam shares biological history with us but that does not mean it can prefer in the way that humans do, so clearly more evidence must be given if we wish to think of the clam as a morally considerable creature. Singer (2011) and Regan (1983) provide supplementary evidence by explaining that there is behavioural evidence that can be drawn on for the claim that certain animals are things with mental states and preferences. For instance, a dog may yelp if its tail is stepped on. Although this may be good evidence for animals having interests, it is clearly not good evidence for AI having interests[16]. Further, it is only good evidence for a claim for a high likelihood of probability, and not a certainty, that animals have preferences. Thus, AI displaying behaviour that indicates the presence of pain provides little evidence for an argument of probability for AI having preferences, given that they have no shared biological history with humans, or other creatures that we have reason to believe experience pain. Singer and Regan's account does not get us to animal preferences with certainty and cannot even give us probability with regards to AI.

[16] It has been pointed out that this is an obvious point to make, but the point being obvious should lend credence to the conclusion of the argument. As such it seems appropriate to make such a point.

Given that Singer only claims that it is highly likely that some specific animals have mental states (Singer, 2011), I take it that the problem of determining whether another creature can have preferences (in the sense that they have mental states of some kind of desire) as being similar to the problem of other minds. If we were to take up a sceptical stance towards other persons having minds – in the same manner that we might take up a sceptical stance towards animals having preferences – it is difficult to develop a proof to show that other persons have minds. In the same way to develop a proof that animals have a mental state of desire, in such a way as to be in a state of preferring, is difficult. That sort of information is simply not available to other persons. As such, a similar problem will arise if we wish to sensibly talk about AI having preferences.

This thesis assumes that no viable solution for the problem of other minds is readily available, and that the same holds true for the 'problem of other preferences'. As such, certainty is beyond our grasp, and we must look to other avenues. Dennett presents one particularly attractive method.

Section 2: Dennett's Approach to Other Minds

Dennett's account aims to provide a justification for why we would assume that another creature's behaviour can be predicted and explained by attributing beliefs and desires to that creature[17]. In so doing he puts aside consciousness and aims at giving good grounds for taking up the assumption that something has intentions. Dennett's theory, therefore, does not attempt to solve the problem of other minds; it attempts to describe a method for predicting when a creature is a subject with intentional states. The goals of those two projects differ substantially.

The problem of other minds is a complex and difficult topic. There are many variations, each with its own nuances to consider. For the sake of simplicity, I shall refer to the traditional version (Avramides, 2021). The problem of other minds is an epistemic issue that seeks to provide some account that will show how anyone can know that other minds exist in the same way that my own does. It is outside the scope of this paper to discuss the other minds problem. Dennett's account of the intentional stance is extremely useful in that it *avoids* the problem of other minds. The strategy that he opts for instead is to aim at identifying other beings that are "true believers" (Dennett, 1998, p. 19), or systems that have intentional states.

Dennett carefully excludes the notion of consciousness from his approach to other minds. Although he claims that he thinks there are particular brain states that correspond to desires and beliefs, he points out that our current understanding of how that works at the neurobiological level is

[17] This is obviously true in regards to explaining a large portion of salient human behaviour. If Mike states he does not like to eat tomatoes, and also does not eat tomatoes, we can reasonably infer that there are beliefs about the taste of tomatoes that he holds, namely that he does not believe they taste good.

incomplete at best (Dennett, 1998). He states that "[m]y thesis will be that while belief is a perfectly objective phenomenon … it can be discerned only from the point of view of one who adopts a certain predictive strategy, and its existence can be confirmed only by an assessment of the success of that strategy" (Dennett, 1998, p. 15). The next best thing that can be achieved, according to Dennett, is to take up a strategy to identify systems that have intentional states.

Dennett very clearly states that this is a matter of predictive power (Dennett, 1998), it is not a means of certainty. He points out that a strategy is a method of prediction, and that a strategy is evaluated according to how successful it is. If it were the case that we could measure the intentional strategy against the facts about whether an object has desires or not, then we would not need the intentional strategy to begin with. As such, we should lower our expectations as to what can be feasibly achieved, and aim at accurate predictions of behaviour by taking up certain assumptions. He does this by contrasting 3 predictive stances against a few other strategies of prediction (Dennett, 1998).

There are a large variety of predictive strategies, each with varying degrees of success. A strategy that is perhaps not very useful would be predicting whether or not there is to be rain during the day by counting the number of words in the headline of a news article and comparing it to a list of numbers that correspond to certain weather phenomena. Sometimes the prediction will match up to whether it is raining, however this strategy does not take into account any causal link between the state of the world and the possibility of rain. If an even number of words predicted rain, and an odd number of words predicted no rain, then at best the method has a 50 percent chance of success. This method is therefore as good as a blind guess, all other things being equal. An example that Dennett himself provides is that of astrology (Dennett, 1998). It seems very unlikely that the patterns of the stars at the moment of one's birth has any meaningful impact on one's character. And it seems that the strategy does not work very well at predicting behaviour. As such, it carries very little predictive power in describing what kind of behaviour one is bound to display. Dennett is only interested in strategies that accurately predict behaviour.

A predictive strategy that accurately predicts behaviour would be the physical stance (Dennett, 1998). Dennet describes this strategy as taking up a stance wherein one examines the physical qualities of an object, as well as the physical rules pertaining to its behaviour, as being a product of physical laws (Dennett, 1998). This works quite well for phenomena such as lighting fires or throwing a ball accurately. If we understand the forces applied and how they will act on the objects in question we can predict how the object is likely to act in those circumstances that the forces are applied. Using this stance, we can also predict where a falling stone is likely to land. If we are very precise about the various factors involved, if we were to codify and record the forces involved, we could predict how objects will behave very precisely. It is this predictive strategy that allows for commercial flight, or

the locomotive capacities of a car. The physical stance (Dennett refers to taking up a strategy as taking up a stance) is quite useful then, and we can see that the physical stance is a good strategy because the predictions we make seem to match up with the information we have gathered after the fact (Dennett, 1998). If someone can predict where a stone will fall given all the relevant information, and that prediction turns out to be accurate on a regular basis, the physical stance must be an effective strategy for predicting the behaviour that stone.

A second useful stance comes out of the shortcomings of the physical stance. In some cases, some information is too complex, or unnecessary, to perform a prediction about an object's behaviour from the material stance. In the case of a computer, it is beyond the capacity of any one person to predict the state of the lights on the screen based on the keystrokes being inputted into the computer and the manner that that information is being processed[18]. That is to say that we do not have the capacity to hold enough information about either the process of transferring information or about the physical states of the matter comprising the device. If I were to type on this word generating program, it is clear that even if I were to understand that the information is transmitted through such and such technical manner, I would not be able to point at a specific local area of any of the components and explain that the effect at such an area is producing such and such an effect, leading to so and so effect through *x* pathways that light up precisely these diodes in my computer screen. In this case it is more economical to employ the design stance (Dennett, 1998). This strategy focuses on the designed features of a system to predict how the system will behave under given stimulus. Under this paradigm it is not even necessary to have any technical knowledge about how computers function in a material stance sense. It is enough that I know that when I push the power button, the computer will turn on, and if I click on a particular icon on my screen it will run that program.

A kind of interest can be drawn from the design stance. Kidneys, when viewed from the design stance, can be healthy or unhealthy, or functional or dysfunctional. If it is one's intention to ensure health in a kidney there are certain steps one can take to achieve that end. It is good for the kidney to be supplied with adequate amounts of water, and the correct number of vitamins and minerals that a kidney requires for good operation. This is like Regan's claims that a good that is brought about when a "gardenia" (Regan T. , 1976, p. 495) is well cared for. These interests have to do with ensuring good operation, rather than being derived from preferences. It seems that there are two broad kinds of interests at work here. There are interests that come about because of design features, plants and insects have these kinds of interests; and interests that come about from a combination of the intentional and design stances. Only the second kind of interests (of interests drawn from taking the

[18] Although someone might have the technical knowledge to understand how digital information is stored in a physical state and how that information might be transmitted to a screen, it is clear they do not have a complete understanding about how each specific element of that system is adding up to that specific bit of information being presented.

design stance) are moral properties, as far as interests derived from the design stance goes. From here on I shall refer to these design interests as design-interests, for the sake of clarity. In Chapter 3, Section 1, Subsection 1 I shall draw on design interests to explain how taking up the design stance and the intentional stance simultaneously leads to her handler locating Sally's interests.

Section 3: The Purpose of Taking Up Particular Stances

It may well be asked why these two stances are employed, after all everything that the design stance achieves can, in theory, be achieved by the material stance (Dennett, 1998). The main features that distinguish the material stance from the design stance are that a) there is a smaller requirement of information and computing power to predict the behaviour of a given system if you can use the design stance, b) the successful prediction of a system's behaviour through the design stance conceptualizes and explains the behaviour of things before us.

The material stance can explain how clicking on a particular icon will result processes happening in a particular computer, and those processes will result in a particular set of diodes lighting up on the computer screen. A sufficiently intelligent being, or powerful computer may be able to provide an account of how that happens at a material level for any specific bit of information (Dennett, 1998). However, there is much more predictive power, through much more economical means, from the design stance if you were to command a friend to "select that icon and run the program." Through the design stance it becomes clear that on any compatible machine – already the design stance is not limited to a single computer and can be used to infer processes on other computers – selecting an icon will result in program *x* running, or process *y*, or process *y* occurring. I gain far more predictive power from treating the computer as an object with designed features rather than a merely material object, in that I can predict the behaviour of a wide variety of computers with different internal components, insofar as they all share the same designed features. Compared to the example of a thrown rock, if we were to examine the event from the design stance there is no more useful predictive information to be gathered that we could not gather from the material stance. It is clear that there is more predicative power in the use of the material stance in this case. Similarly, when using a computer the design stance is the most useful stance to take when you have some goal in mind for what you wish the computer to accomplish. Therefore, to take up the material stance when you wish to investigate how digital information is 'saved' to a material substance is less useful a strategy than to investigate it from the design stance.

A particular system's behaviour is conceptualised when viewed from a specific stance. I can look to the throwing of a rock and claim that it behaves that way because of the laws of physics that are responsible for describing that particular arc. In this way the behaviour of the rock is described as

a feature of the rock and its interaction with the world. The design stance conceptualises how something *ought* to act. I can talk about the function of a kidney by referring to its 'design', despite nothing having designed it (in the sense that there are no schematics or prior intent for the kidney to act it the way that it does). None the less, I can also talk about the kidney as being 'malfunctioning' if the kidney no longer behaves in the ways that the design stance predicts it will behave. The stances conceptualise behaviour in the form of normative conceptions of behaviour. A kidney's function is tied up in the organ behaving as expected from the design stance. The computer's function is found in the same way.

Section 4: The Intentional Stance

The intentional stance is a strategy that aims at predicting the behaviour of actor by assuming that they have intentions. This stance conceptualises behaviour of an actor in the form of the actor having intentional states, just as the design stance imbues significance in the form of function. The intentional stance is straightforward, first an observer decides to "treat the object whose behaviour is to be predicted as a rational agent" (Dennett, 1998, p. 17). Then the observer attributes a set of beliefs and desires based on behaviour or previous knowledge about the actor, or "what beliefs that agent ought to have" (Dennett, 1998, p. 17). Finally, the observer evaluates the stance by predicting behaviour of the agent and comparing that prediction to actual behaviour. If the prediction is accurate then taking up the intentional stance is warranted (Dennett, 1998). In the case of John, one could surmise that, as a human, John has a desire to eat and drink when he is thirsty or hungry. As such we would predict that on a hot day John would likely drink, as he is thirsty. Having spoken to – or observed – John the observer learns that John dislikes both coffee and sugar. Consequently, John tends to drink tea without sugar. In this case the observer has inferred probable behaviour that John will exhibit based on a set of beliefs that have been generated based on past behaviour (he has told the observer he does not like coffee or sugar) and on the kind of thing that John is (he is a thing that must drink). If the prediction matches with observed behaviour of John, then the intentional stance is warranted. This is the basis of the intentional stance.

Of course, the physical stance could theoretically perform the same function. For example, a hyper-intelligent being or sufficiently super-powered computer could take all the physical information about any intentional creature and generate an accurate prediction about that intentional being's behaviour (Dennett, 1998). However, the intentional strategy bypasses the problem of all those computational tasks, as well as the problem of gathering all that information. Further the intentional stance provides predictive power of a different kind. Dennet points out that a hyper intelligent being that would only be able to understand human behaviour through the material stance would be

astounded at the predictions that a person in the intentional stance would be able to generate, based on the observation of human behaviour (Dennett, 1998). The hyper-intelligent being would be confused at the prediction that in 6 months' time a particular human would be boarding a metal tube that would launch into the sky and drop off that person in France (Dennett, 1998). A human in the intentional stance would merely have to notice a person buying air-tickets to France, scheduled for flight in 6 months' time. Of course, they might not be able to predict all the events leading up to that flight, how often this person takes sick time off from work or other events of that sort. However, the important event relating to the buying of the ticket is good reason to believe that out subject intends to fly to France in six months' time, and that can be predicted within a small margin of error (Dennett, 1998). This margin of error could be reduced by assuming that John has certain beliefs, such as his belief that to miss his flight would be an inexcusable waste of money.

There are some problems with this example, it seems unlikely that a hyper-intelligent being would be unable to pick up on such a pattern, nor would it be very likely the sort of thing that could be confused by a single persons behaviour if they can see that the behaviour of speech works this way with all human (Dennett, 1998). However, the example does show that the intentional stance has a very real and useful application in the prediction of behaviour. The criteria used to evaluate the usefulness of a given strategy is in how effective the strategy is at predicting the behaviour of any given object, and clearly this strategy is effective to a large degree. Dennet points out that we are already quite successful in the use of this strategy in our dealings with other persons (Dennett, 1998). This lends legitimacy to the notion that this stance is effective. In being able to interact with a creature as though it has intentions and having that notion of interacting with an intentioned object giving more predictive and explanatory power than from either the design or material stance, gives the behaviour of that object the conceptualisation of being intentional. This is done in the same way that a computer is assumed to be designed.

It is important to note that successfully taking up the intentional stance is a good reason to believe that the creature has intentions[19]. For instance, taking up this stance with persons and some animals leads to good predictions of behaviour. This lends credence to Singer and Regan's argument that behaviour is good evidence that animals are in fact the kind of thing that have preferences. Further, some behaviour is not better explained by the design stance, a quintessential case of this being language usage among humans. Any human expressing through language some belief X that corresponds well to their behaviour is likely truly holding that belief X.

[19] Where being successful includes generating accurate predictions that cannot be more efficiently predicted by the use of another stance.

It is not the case that multiple stances cannot be taken up towards one unified system[20] . The use of the tongue can be seen from both the design and intentional stances. It is designed in the sense that the tongue has a very specific structure that allows it to be manipulated in certain ways, for the use of communication, or eating and such. But its actual usage, in communication for instance, is often better understood from the intentional stance. When making specific utterances there is a story to be told that relies on mental states of intention and desire. There is also a story that appeals to natural selection that explains how the tongue functions. In the case of speech, one would communicate intentionally but will do so with 'designed' features.

Successfully taking up the intentional stance towards a creature implies the presence of preferences in that creature. The predictive value of taking up a stance is good reason to hold that the stance is valid. If my predictions imply that some object has desires and beliefs, and those predictions turn out to be accurate, then it follows that the agent being observed does have desires and beliefs. That assumes that the agent's behaviour is not *better* explained by appealing to the design or physical stance. In the case that the behaviour is more accurately predicted by the design stance, it is going to follow that the object does not have beliefs or desires. For example, consider the difference between a horse and a car. A car's behaviour is easily reproducible in a way that a horse's behaviour is not. I can with extraordinary certainty predict that a well-functioning car's engine will make specific noises and motions, when the accelerator is applied in particular ways. This is best understood by the design stance. A horse's behaviour is best understood from the intentional stance as there is a reasonable chance that it will behave, abstractedly, as predicted. A well-trained riding horse will behave slightly differently every day, depending on its mood. But it will behave consistently in terms of its mood, and that cannot be fully explained by the design stance.

Chapter Conclusion

This Section has shown that other minds and other interests are problems of a similar nature. I have argued that Dennett's intentional stance theory is a good alternative to solving the other minds problem. I have explained what the intentional stance is and functions and how the design stance functions. These stances allow observers to assume that other creatures have desires and beliefs, which implies that those creatures have preferences.

[20] Dennett refers to anything that expresses behaviour as a system, whether a rock mid-flight or a human mid speech.

Chapter 3 AI Interests

AI interests

Singer's project is to show that animal interests fit into his theory of ethics, specifically his theory of preference utilitarianism. Some of the arguments that he uses to show this have been touched upon already in Chapter 1, Section 2 – specifically, Singer's argument that animals should be included as ethically considerable in his preference utilitarian theory by virtue of evolutionary and behavioural evidence. This thesis is not specifically about animal moral considerability, but about a broader category of non-human moral considerability. Some of the arguments that Singer and his contemporaries use to show that animals are morally considerable are therefore not applicable to all non-humans in general. In regards to AI we do not have the kinds of biological evidence that we have for animals. In which case we must appeal to the intentional and design stance to come to know interests (or lack thereof) in AI.

A problem that this thesis must confront is that it lacks any AI with which Singer's theory could be applied. There are no good examples of what AI will look like, or good indications of how they will behave. As such I will be using fictional accounts of AI. This is not ideal, there is no reason to assume that Asimov would have a good idea of what AI may look like in the future. However, it is the next best thing to use in the absence of having concrete case studies.

This chapter will discuss two narratives, *The Bicentennial Man* and *Sally,* both written by Isaac Asimov. The focus of this section will be on discussing how interests are manifested in Andrew (a character in the former) and Sally (a character in the latter), the most prominent non-humans in the narratives. Further I shall be examining how effective those narratives are in portraying characters that are morally considerable creatures.

In section 1 I will dissect how Dennett's theory of stances is useful in analysing the ethical viewpoints of the central human characters. I will do this by highlighting the expressed beliefs and actions to and about the automatic cars, as these beliefs and actions are given in the narrative. Folkers is shown to take up the intentional stance towards these cars and so takes up ethical beliefs about how the cars should be handled. Gellhorn's position is less clear, and I investigate the various implications of his belief system, specifically whether his actions are ethically grounded in the case of his taking up the intentional stance, and in the case of his taken up the design stance. I do not believe that the reader is given enough information to clearly state whether the intentional stance can be validly taken up to the automatic cars, and therefore cannot comment on the validity of holding that Sally has interests. Therefore, I cannot make decisive comment on the ethical nature of the treatment of these

automatic cars. However, this narrative serves well to highlight the possible ethical issues around the case of non-verbal AI, and so is an excellent resource to this thesis. Further, *Sally* gives an excellent account of the importance of stances in approaching ethical behaviour. Where Gellhorn does not take up the intentional stance Sally is not an ethically considerable creature. In such a case he has no reason to hold that he has done anything wrong.

The second section will analyse how The Bicentennial Man draws a picture of a non-human robot that seems to have interests, and how that would qualify him for being the sort of thing that is morally considerable. I shall do this by looking at how his interactions with the world indicate a developing sense of self, as well as showing how his self-inflicted mortality is an essential feature of being human, and therefore being a moral agent. This section, in contrast to section 1, argues for the necessity of taking Andrew as a morally considerable being.

Section 1: Sally

Subsection 1: Folkers and the Intentional Stance

A man named Jacob Folkers has inherited a farm on which he is the caretaker of retired automatic AI cars. Owners of these cars become emotionally attached to them, and so they send these cars to Folkers to be cared for after the death of their owners. A man named Raymond Gellhorn approaches Folkers with the intention of transplanting positronic engines from Folkers' cars into newer car bodies intending to sell them at a profit. Folkers refuses the offer as he believes that these cars now live a good life, and splicing them into other bodies could be distressing or painful. Folkers warns Gellhorn that he will call the police if Gellhorn ever returns to the farm again. Later that night, Gellhorn attempts to coerce Folkers into his plan by breaking onto the farm and demanding that Folkers disconnect half of the cars from their engines. Gellhorn claims that the alternative is that he will personally disconnect all the engines, and very likely destroy a significant number of them in doing so. Folkers refuses and sets his cars on the intruders, the cars then chase Gellhorn's associates away. Gellhorn abducts Folkers, bringing him onto a bus that Gellhorn has personally spliced a positronic engine into. Folkers is enraged by this saying that the bus is in extreme pain because of Gellhorn's poor work. Gellhorn is unfazed by causing this pain and is quite unhinged from the cars attack on his associates. Gellhorn attempts to escape the farm in the bus but is pursued by Sally and some other cars. Sally appears to communicate with the bus to some degree, the bus expels Folkers and Gellhorn and then chases Gellhorn down over a large distance until the latter is near death due to exhaustion. The bus then kills Gellhorn by driving over him.

The entire narrative is given from the perspective of Folkers, to say that 'Sally is a minded thing,' is to say that 'Folkers sees Sally as a minded thing.' It is entirely to the mind of Folkers that Sally is a minded thing, if he believes it at all (Asimov, 2018).

Folkers morally considers the needs of Sally and the other cars. He aims to improve their well-being from a stance that is not only mechanistic or material, but also intentional. He does not want the cars to function well so that they can be driven longer or to otherwise be used. He wants them to function in a manner that would, to his mind, improve the life of being a car (Asimov, 2018). To this end, he installs devices in his cars that allow – amongst other things – them to self-regulate their appearance. He also ensures that the cars always have plenty of fuel, ensures that their tyres are always in good working order, does not needlessly turn off the cars at night, and generally allows the cars to be free ranged (Asimov, 2018). By free ranged I mean that the cars are not exercised, or otherwise worked. They are able to roam the farm of their own accord.

Folkers seems to take up the design stance in regards to the health of the cars. In Sally having capacities that can be employed to self-maintain – such as the squeegees that Sally uses to clean her windscreen – Sally is made more robust. In this way she has an 'interest' in the maintenance of those self-cleaning capacities, in the same way that a kidney requires certain conditions to maintain itself. Regan's sense of welfare and design-interests inform this kind of interest having. However, this is not an interest in Singer's sense of the term. Singer caches out moral properties as preference-interests, and his utilitarian theory does not function without that aspect. Therefore, Sally having these interests counted in a moral sense is going to be dependent on her being the sort of thing that is intentional. That is not to say that Sally's design-interests are not linked to her preference-interests, only that her design-interests are not moral properties *per se*[21]. Folkers still seems to take up the intentional stance towards Sally, but he draws out her welfare from the design stance in relation to her 'body'. I shall return to this point below but first I shall explain why Folkers takes up the intentional stance towards Sally.

Folkers rests this notion – that the cars are improved by the treatment he gives them – by arguing that each car has unique characteristics that Folkers refers to as "personalities" (Asimov, 2018, p. 11). The convertibles are described as female, the sedans as being male. Individual characteristics being shown as well. Tom, the first automobile that comes into Folkers' possession, is described as something like an elderly gentleman. Tom stays inside but other sedans are variously described as playful or roguish, one is described by Folkers as playing pranks on other cars (Asimov, 2018). I contend that Folkers holds that these personalities arise from having specific sets of preferences. He is quite clear that he believes that these cars are in a state of preferring some things to others and that

[21] Chapter 1, Section 2, Subsection 3 spells out how welfare can be drawn from things that are not preferring creatures, in this case Sally's health.

these preferences should be seriously considered (Asimov, 2018). Folkers' taking up the belief that the cars have personalities implies that he believes that they act with goals that are specific to each car. As such, he assumes each car to have as set of beliefs about how the world is, and what each car can achieve within the world. These beliefs about what can be achieved – their preferences about how to go about in the world – seem to be acted out in the 'prank playing' that Folkers describes (Asimov, 2018).[22] He states that "[t]hey wouldn't be happy in another car" (Asimov, 2018, p. 16) when explaining why he is refusing to take up Gellhorn on his offer to transplant these vehicles' engines. Further, he directly asks Sally "what do you think … ?" (Asimov, 2018, p. 16) and interprets her opening and closing of her doors as being "the way [she] laughs" (Asimov, 2018, p. 16). He takes up the intentional stance towards Sally, and so takes her to have intentions about the way that she would wish to live her life.

That these cars have preferences is central to Folkers' holding that they are the sorts of things that are morally considerable. Sally is a convertible, and so is characterised by Folkers as female. He has a particular fondness for Sally and expresses it by favouring her when improving the capacities of the cars on the farm. She is given tools by which she can regulate her appearance using a squeegee and a cleaning agent dispenser (Asimov, 2018). Folkers giving these kinds of improvements seems to be led by his belief that the cars operate efficiently when allowed some degree of personal freedom – their fuel tanks are always kept full so that they can turn over during the night. Folkers claims "[a] positronic brain stays in condition best when it's got control of its chassis at all times, which means it's worth keeping the petrol tank filled so that the motor can turn over slowly day and night" (Asimov, 2018, p. 13). This aids Folkers in his belief that his cars are in a state of better condition than other cars not given the same capacities to self-regulate. The argument, presumably, is that if the interests of the cars is taken seriously then the ability to self-govern their actions is also better for their well-being, and so for their interests. Gellhorn rejects that the cars have interests to begin with and so is not convinced of this.

Subsection 2: Gellhorn and the Design Stance

Gellhorn sees the cars as objects of utility, although likely as very sophisticated objects of utility. They are first and foremost machines to Gellhorn, and so he sees no moral issue in using them accordingly. This viewpoint apparently also justifies his belief that the cars are the sorts of things that he can experiment with and create Frankenstein's-monster-like cars for his own profit. He may even feel that they are wasted on the farm. Gellhorn's position is not implausible, particularly as he has never experienced the cars' 'personalities'. There are very few (if any) people that would be morally

[22]See Chapter 2, section 1, Subsection 4 on how taking up the intentional stance occurs.

outraged if someone were to swap engines between two Toyotas, and Gellhorn's scheme to place old positronic engines in new car bodies is not so far removed from swapping mechanical engines between cars in our own world. That kind swapping of engines, in the narrative, assumes that positronic engines are like car engines rather than like brains, or structures that house minds. It does not follow from the fact that Gellhorn wants to swap positronic engines specifically that his plan is unethical, as the brain analogy is not fully shown to be true. This reveals that Gellhorn takes a strongly mechanistic stance to how Sally and the other cars operate, and this mechanistic stance is what I would argue is the design stance. In order to show that Gellhorn's position is unethical it is necessary to show that these cars are the sorts of things that can have their interests unjustly abused or ignored. That these cars are the sorts of things that can have their interests unjustly abused or ignored is a basic assumption for Folkers but not for Gellhorn.

Folkers seems to hold that positronic engines function something like brains in humans, and therefore have intentional states. The term 'positronic' is fanciful, but the idea is that positronic engines are the seat of personality and memory in these cars. In pursuing Gellhorn's project Folkers would – to Folkers' mind – be performing a painful and unnecessary operation on the cars – for monetary gain. This would clearly be ethically dubious for Folkers as he sees the cars from the intentional stance.

Folkers assumes an interest-based stance towards the cars, which Dennet would call an intentional stance. Gellhorn seems to assume some kind of causative stance, possibly the material stance but more likely the design stance. These stances have a very real effect on how Gellhorn and Folkers weigh the moral value of their actions towards the cars. Folkers sees the cars as interest bearing objects and draws this account of interests from the 'personalities' that the cars seem to display. As such he is mindful of the effects of his actions towards the cars, deliberately aiming to give them more freedom to maintain themselves. Folkers also gives the cars more opportunity to explore and interact with the world around them. Folkers' actions aim at the improvement of the life that he perceives them to have.

There is a persistent theme of freedom in the narrative, Folkers in particular seems to position personal freedom and the capacity to act for oneself as central to having moral status respected. His upgrades aim towards giving autonomy, aimed at the cars acting for the purposes of their own needs rather than the needs of their owners. This shift in focus, from the needs of the owner to the needs of the car is a central feature to Folkers' interest-based account of ethics. Specifically, that the reducing of actions that abuse (or unjustly ignore) a car's interests is the most ethical course of action to take. Singer's account aims in a similar direction and argues that it is morally objectionable to raise one's own interests above another's without good reason (Singer, 2011). Holding this position relies heavily

on holding that the cars have interests and that those interests hold equal weight to Folkers' own interests.

That Folkers aims at giving greater autonomy to the cars, so that they may pursue their own projects, is much like what Singer's theory demands for animals. That their interests (assuming that the cars have interests-as-moral-properties) be considered as interests, and not ignored simply because Sally and the other automatic cars are merely automatic cars. As such, Folkers sees that his duty is to improve the possibility of the cars being able to pursue their interests. This is how I have tried to formulate interests in regards to welfare. A good ethical theory of interests will include some notion of being-good-for the creature at hand and will necessarily include the creature preferring[23] that event that leads to being good for the creature to be the case. I do not claim that Sally or the other automatic cars can prefer at this point, I merely wish to reinforce that being a preferring creature is necessary to having interests as a moral property.

Gellhorn's position assumes that these cars are objects of utility, and he is not concerned with either their interests or their agency – for he believes them to have neither. In being objects that are seen from the design stance, the cars are performative. They act like automatic cars because that is how they are programmed to act, not from having preferences or interests. Further, for Gellhorn, they are fundamentally objects of use to human needs – if we give his actions the best possible interpretation – or for his own personal needs. Therefore, even if they are the sorts of things that have interests then those interests are not to counted as being as important as any person's interests in general. Keeping in mind that these cars are certainly not persons. These assumptions lead Gellhorn to assume that he has a right to swap old engines into new machines for his own profit, as the cars are merely tools.

That view does not seem to follow, as these cars have a freedom that we could not see in a purely mechanistic manner. Their behaviour is flexible and dependent their surroundings, even if they are bound by strict limitations based on those features that are designed in. The behaviour demonstrated by the cars has already been shown to be distinct between each of them. Tom behaves differently from Sally. Although it is not clear that these cars *are* creatures with intention states, that is very different from saying that these cars' behaviour *are* entirely mechanical. Animals exist in the same way, the behaviour they will exhibit is likely largely to do with their evolutionary history. For instance, certain kinds of skeletal structures are more suited to pursuing prey than grazing from the tops of trees; certain tooth shapes more apt to chewing foliage than tearing meat. Any species of animal might behave in ways that follow the same patterns of behaviour as others of its kind, but the individual actions of each of the individual members of that species will differ from its sibling or

[23] See Chapter 1, Section 1.

parents. This could be explained by environmental factors, perhaps the offspring of two elephants must migrate to find food in a drought that its parents never would have experienced. However, the fact that the behaviour between parent and offspring is so starkly differentiated, implies that a purely mechanical explanation of that behaviour is insufficient to explain the variety of possible behavioural outcomes.

Returning to the first view that Gellhorn might take up, even if it turns out that these cars can feel pain, it makes no difference to his treatment of them because they are simply the sorts of things that can be used in that way. A telling bit of evidence to this notion is that when Gellhorn responds to Folkers' questioning he bluntly asks Folkers what the consequence of the bus feeling pain is (Asimov, 2018). However, if we assume that Gellhorn does hold that the bus can feel pain, then our second interpretation – below – will hold.

Subsection 3: Gellhorn and the Intentional Stance

The second interpretation of Gellhorn's stance is that he is aware of the interests of the cars and simply does not care to morally consider Sally. This would fit the pattern of behaviour that he expresses towards Folkers and the farm. Importantly, this is Gellhorn taking up the intentional stance towards the cars. However, Gellhorn's view on what the stance implies for the cars – and Folkers' himself having interests and preferences – is not in line with how Singer organises his ethical theory. Gellhorn sees no reason to assume from those facts that Sally or Folker's interests hold as much weight as his own. Gellhorn's actions imply that there is a strict hierarchy of interests the top of which are his own interests, then other persons', then the cars'. Crucially, he offers Folkers a portion of the profits, even after Gellhorn attempts to coerce Folkers into agreeing with Gellhorn's scheme. As such, Folkers' interests are important enough to include him in the transaction even if Folkers does not agree to the scheme, but not weighty enough to necessitate Folkers' agreement to the scheme.

There are some other possible explanations for why Gellhorn wishes to share the money with Folkers, perhaps there are pseudo-legal or legal reasons which may have been obvious to the reader at the time of writing that would implicate Folkers to some degree, but it seems that Folkers could have just rejected the money and gone to the police. It seems plausible that Gellhorn may just have been lying, this may well have been malicious and intended to hurt Folkers further, or to try and persuade Folkers to do what Gellhorn needs him to do. However, it seems that Folkers, having a weapon pointed at him, is likely to do exactly what Gellhorn wishes him to do either way and so there is no reason for Gellhorn to make such a suggestion. As for the statement being merely malicious, it is not out of the question. Gellhorn does express a distinct lack of care for the interests of others, although it is not obvious that his actions are sadistic rather than selfish. To my mind Gellhorn is not

specifically enjoying the discomfort of Folkers, although some element of enjoyment is certainly there. He is simply more concerned with the material outcomes that his actions will lead to.

It is not entirely clear how Gellhorn's stance (about how cars operate) changes before he is killed by his bus, but it is clear that the cars are no longer only objects of utility. Gellhorn is last seen by Folkers being driven away, locked in his bus, and it is then mentioned in the newspaper that Gellhorn's body has been found some distance away, having been driven over by some car – presumably by said bus (Asimov, 2018). Immediately prior to this Folkers witnesses Gellhorn break down emotionally, possibly after realising the danger he is in. The cars having chased Gellhorn's crew mates off the farm and are now being chased by Sally and some other cars. He now fears the cars, whereas before he thought of them as relatively harmless. It is interesting to note that Folkers also develops a fear of the cars, pointing out that he beings to actively avoid Sally and the other cars after Gellhorn's death. However, it is this specific act of violence that drives Folkers to fear. It is heavily implied that the cars have run people off the farm before, but likely this is the first case of running someone off the farm resulting in the death of a person. He warns "Don't be silly. They won't kill your men...My cars have been specially trained for cross-country pursuit for just such an occasion; I think what your men will get will be worse than an outright quick kill." (Asimov, 2018, p. 22) It is worth noting that this statement shows that Folkers is complicit in the organisation of these cars for the defence of the farm, making the appearance of his sudden fear more stark. Folkers' fear is not simply that the cars are capable of killing in some circumstances but also that the cars may be communicating with one another. His specific fear is that they will at some point turn on humans. A second fear is that the cars may not be cognisant of the fact that humans provide roads, petrol, and maintenance for the cars (Asimov, 2018). As such there may be no reason to stop the cars from simply killing all humans.

The stances taken up by Folkers and Gellhorn have very serious and real consequences for how we view the ethical repercussions for either's behaviour. Gellhorn forcibly enters Sally, characterised by Folkers as female, and manually drives her. Leading up to this Folkers has made it clear that neither Sally, nor any other car on the farm, is driven by people anymore. Further, the cars are left on all the time, they are under their own control, being turned off has been known to happen but is reserved as a punishment for misbehaviour. Gellhorn first attempts to climb into Sally through her door and she locks him out, he then claims that if she did not wish to be driven then she should not "go around with its top down" (Asimov, 2018, p. 17), he jumps over her door and turns on her manual controls. He drives her for a while and then gets out stating that he thinks he has "[done] her a lot of good" (Asimov, 2018, p. 17)

There are two perspectives to discuss if we are to consider the ethical weight of what Gellhorn has done. The first is that of Gellhorn's stance towards Sally, and the second is of Folkers stance to

Sally. Clearly these stances are going to affect how we perceive the ethics of Gellhorn's actions. A third perspective, that of the reader, is also important but will be left out of this thesis as it would make assumptions about the stance that the reader will take up.

The first perspective is that of Gellhorn towards Sally, I have pointed out that Gellhorn either views Sally from the design stance or from an intentional stance that assumes that the interests of Sally are below the interests of himself. These two stances will produce different ethical outcomes for Gellhorn and so we should address both. The first stance, the design stance, assumes that behaviour is programmed in – like a computer – and not brought about from intentions, or in this case preferences. This stance, if true, shows Gellhorn to be acting in a way that cannot harm Sally as an interest holding thing, as she cannot have any interests. He may harm her in other ways, such as damaging components required for her functioning, but this does not seem to constitute an ethical harm to Sally if there are no interests at stake. The second stance suggests that some degree of harm is done. However, to Gellhorn that would be a trivial harm.

If Gellhorn takes up the intentional stance, but takes Sally's interests to be unimportant, there are ethical concerns that mirror sexual assault. Sally is characterised by both Folkers and Gellhorn as female, both referring to Sally as 'her'. In that context the phrase "[he did] her a lot of good" (Asimov, 2018, p. 17) denotes serious sexually aggressive overtones to Gellhorn's actions. These facets of his spoken beliefs, Sally's characterisation as 'her' and his presumption that Sally is required to protect her interest in not being driven, are seriously at odds with any explanation of his behaviour towards Sally that appeals only to the design stance. There is an interpersonal aspect to Gellhorn's actions towards Sally. Gellhorn claiming that his actions are somehow good for Sally is a further indication of being interpersonal. Gellhorn claims to have improved her by his actions, which seems to imply that she is the sort of thing that should be driven and should enjoy being driven. The question of whether Sally really is that sort of thing is not a concern for Gellhorn, he simply assumes that she is.[24]

Singer points out that not all interests are weighted equally, merely that interests are weighted according to the degree of that interest (the interest in eating while hungry is surely not weighted as equal to the interest of having access to whatever equipment might be necessary for the pursuit of a hobby) and not weighted by the kind of thing holding the interest (Singer, 2011). The comparable interests – such as eating when hungry – of a cow to a person is not specifically weighted less than

[24] I have deliberately addressed this sexual assault overtones from the perspective of the two male actors, and not from Sally 'herself'. I have done this for two reasons, the first is Sally is not human. How far the allegory of sexual assault is carried by her is unclear, and the point I wish to make can be made without having to delve further into it. Secondly, I do not think I could do sufficient justice to the questions that come about from addressing the issue from Sally's perspective to warrant my doing so. It is an interesting and important problem that should be pursued by more persons more competent in the literature than me.

the same interest for a human. But 'lesser' interests are weighted less than 'greater' interests. The second stance of Gellhorn implies a deviation from Singer's weighting system, Gellhorn sees his own interests as more important than other persons – and far more important than Sally's or the other automatic cars – by virtue of those interests being his own. This perspective, to Gellhorn, has similar outcomes to the first perspective, although he may harm Sally to some degree it is not important as Sally is not the sort of thing that requires protections anyway. To Gellhorn, Folkers may require some protections under certain conditions, he first tries to convince Folkers rather than immediately coercing him. However, if Folkers' interests clash with Gellhorn's then Gellhorn will proceed to force the issue. Taking up a stance more in line with Singer will require that we notice that Gellhorn is acting unethically in this situation towards both Folkers and Sally. Sally's interests are denied as important by virtue of her state of being a robot, and to a lesser degree so are Folkers' by virtue of not being Gellhorn. Gellhorn is, in fact, performing speciesism.

The second perspective is Folkers', which sees Sally and the other cars as intentional objects, and objects with interests. As such he treats their interests seriously and attempts to accommodate those interests. It is important to note that although he does this specifically for the cars on the farm, and although he expresses some concern for the treatment of cars not on the farm[25], he does not campaign for these treatments to becomes more normalised. From this fact it follows that Folkers also has some feeling that the interests of the cars on the farm are more important in some sense than other cars. Folkers points out that Sally is his favourite, and that she gets preferential treatment for her 'lesser' interests. Further, although he is unhappy with the idea of Gellhorn turning off the cars on the farm, describing that as akin being knocked unconscious (Asimov, 2018), he employs this method as a deterrent, using it on at least one car when the car frightens some other cars by squealing its breaks (Asimov, 2018). Folkers' behaviour has a darker undercurrent, and he certainly claims to be preferential when dealing with the cars. However, his behaviour towards the cars on the farm is typically more in line with Singer's ethical theory. Although he also acts in a speciesist manner, it is clear that he has some kind of regard for the interests of the cars.

As such Gellhorn's actions towards Sally, and the other cars, is ethically dependent on the stance that we take up towards them. Our approach to the ethics of Sally and how Gellhorn approaches his treatment of her is therefore dependent on how we view Sally, and for what reasons we give Sally the designation that our stances imply. If we assume a stance like Gellhorn's design stance, then we have no reason to hold that Gellhorn has ethically violated Sally. If we assume that Gellhorn takes up the intentional stance then we can see his actions are speciesist.

[25] Folkers points out that cars with the ability to self-regulate tend to require less maintenance than commercial cars that are turned off at night etc.

Section 2: Andrew

Subsection 1: The Introduction of Personhood

The Bicentennial Man, (Asimov, 2018) is told from the third person perspective but with particular focus and insight into Andrew, a robot, and his internal states. Andrew is owned by the Martin family. He is ordered to carve a piece of wood for the youngest Martin child, and his carving is quite pleasing to her and her father. The Martin family have him carve more wooden objects, which they then sell. Mr Martin, Andrew's owner, keeps half of that money and places the other half in a bank account for Andrew's use. Andrew then asks to buy his freedom, to buy ownership of himself, from Mr Martin. This is the first time a robot has attempted to do this, and as such it goes to the 'World Court'. Andrew is granted ownership of himself, on the basis that "there is no right to deny freedom to any object with a mind advanced enough to grasp the concept and desire the state. (Asimov, 2018, p. 575). Andrew, no longer owned by a human, is harassed by some farmers in his town and so returns to the World Court to pursue legal rights for robots. These rights are granted and Andrew, having made considerable money by his pursuits in prosthetics and robot history, eventually pursues legal status as a person. His case initially seems to be failing until he reveals that he has permanently damaged his processing unit (his positronic brain) and will eventually die. Andrew, at that time, was just short of 200 years old and is given the status of personhood soon after he explains that he has organised his own death. He is then declared the Bicentennial Man, and the narrative ends with Andrew's death.

The narrative focuses almost exclusively on the efforts of Andrew Martin to be recognised as a person. In contrast Sally is not person, she is – at best – animal like. She is not a member of human community nor does she aim to it. The narrative of *Sally* specifically and carefully differentiates Sally and the automatic cars into their own society. They are separated from humans by language, and it does not even seem as though they are interested in being a part of human society (Asimov, 2018).

The notion of personhood is a complex one; but of particular interest to this thesis is whether something that is not human can still be considered a person and morally considerable. If they can/are then we should produce the reasons for holding that to be the case. The narrative structure is separated into four distinct phases; the first being when he is given a banking account, the second is when he seeks his own freedom, the third when he seeks to own himself, and the fourth is his legal battle to be recognised as person. These sections all cumulatively contribute some aspect that seem to be important to the notion of personhood.

A recurring theme in *The Bicentennial Man* is the idea of being recognised *by people* as a person being an important element of being person. This theme is an important one for ethics; on the face of it, not being recognised as morally considerable opens the possibility of being treated

harmfully for the kind of thing that a creature is.[26] This is often demonstrated in various forms of bigotry, such as instances wherein racists will sometimes consider others to be sub-human, or only partial humans that are not worthy of all the ethical duties that are demanded by virtue of being human. A further point on recognition is that ethics has collective elements at play. Ethical societies function insofar as the society acknowledges each other as morally considerable. Singer can be read as arguing that we must welcome animals into our ethical community, as animals have interests. However, those who are the victims of bigotry often must demand that they be made members.

Society, in this sense of the word, is a very watered-down notion. I do not mean to say that animals are the kinds of things that we can share norms with, or anything of that kind. What I mean is that there are collections of things which we consider to be morally considerable, to some degree or other. Each of these collections can be thought of as being part of a 'society' or of being included in the moral consideration of persons to some degree. Animals, according to a preference utilitarian theory, have a place in these moral collections that might not be among persons, but are not so far removed that they are not morally considered at all. Persons who are excluded from these moral collections, as historically has been the case, often must fight to be included in these moral collections. Animals cannot fight to be included for they are not the sorts of things that can understand that they are not already part of these collections. Therefore, Singer can be read as arguing for the inclusion of animals into our moral society. His arguments also indicate that all persons should be included as well, and therefore all persons should be welcomed in. I have said that persons who have not been included in our ethical communities often have to demand to be made members only in contrast to being welcomed, not as a prescription that they should have to do so.

Subsection 2: Andrew as pursuing personhood

Andrew's goal is to achieve personhood, to be included in this ethical community as a morally considerable creature. He states "The truth is I want to be a man. I have wanted it through six generations of human beings," (Asimov, 2018, p. 601) and this aspect of his personality coming to the forefront is the major development of the narrative. Andrew is a more clear example of the kind of thing that has interests than Sally is, as he can directly make claims on the world around him, whereas Sally inhabits more of a middle ground. While Sally is the sort of thing Folkers holds to be intentional, Gellhorn rejects her intentionality. Andrew's moral status is not at risk in the same way because it is much more difficult to perceive him as not intentional. Most persons that engage with Andrew take up the intentional stance towards him, the ethical society's rejection of him is because

[26] Humans can be harmed in ways that animals cannot be, consider being calmly verbally abused. The harm done to animals by verbal abuse is significantly different to the harm of being verbally abused as person.

he is an intentional alien. Two notable examples of this rejection are the judge who presides over his world court trial and the two young men who accost Andrew on his trip to the library.

The judge asks Andrew “"Why do you want to be free, Andrew? In what way will this matter to you?" (Asimov, 2018, p. 575) This question assumes intention on the part of Andrew, but in having to ask why his intentions matter to Andrew he is denying Andrew his interests. He does this by asking a creature that he has taken up the intentional stance towards, that has demonstrated his interests by asking for his freedom, why he should have such an interest. He would not ask the same of any other person. Second, the two young men talk to Andrew, but where the judge is open to the belief that Andrew has interests that might be weighted in part against the judges own, the two young men see Andrew as an object that does not belong to anyone.

Andrew is not strictly person and so is treated as being only partly a person. Mr Martin states that “The new models aren't as good as you are, Andrew ... The new robots are worthless,” (Asimov, 2018, p. 571) and gives Andrew his own banking account. In this way he is treated in a way that we might treat children. He is owned but spoken to as though he were a person. When he asks for his freedom the judge does not understand why Andrew would want it, despite being able to speak with him and assuming that Andrew is capable of wanting anything to begin with.

I will argue that due to being creative, Andrew is only partly person. Andrew is given some degree of personhood and is at least not seen as merely an object. Gould (1996) claims that flexibility is an important aspect of being human. Specifically, flexibility is a marker of being part of the Homo Sapiens species. But an implication of finding an alien thing that is flexible like a human, and flexibility being a hallmark trait of being human, is that the flexible alien might have some claim to the benefits of being human.[27] Andrew, if Gould is correct, may therefore be entitled to some share of being human.

Gould argues that human flexibility is the result of biological limits that generate ‘ranges’ of behaviour. He gives the example of aggression; aggression is only seen relative to non-aggressive or a peaceable nature. Gould claims that this means that aggression is not ‘hard coded’ in the sense that biological determinism might claim it to be, but neither is aggression merely a learned behaviour and without any relation to genetic influences. Gould argues that aggression comes about on an extreme of some spectrum that a gene, or a combination of genes, sets the limits too. As such, aggression is only possible if there is also non-aggression (Gould, 1996). By this I mean that to be aggressive is to be ‘high’ in the range of aggression, and to be peaceful is to be ‘low’ in the range of aggression. As such, to be aggressive requires that there be a range including non-aggression. One cannot be aggressive by genetic factors alone as their evolutionary ancestors has periods of peacefulness as well,

[27] Although we may only want to include them in our communities if they are safe enough to be included again. We exclude murderers and the like of our own kind, so I do not think that we should aim to include dangerous aliens.

so aggression must exist in a range. He makes further arguments, specifically about this being evidence that biological determinism is false, but the point of concern here is that there are multiple ranges of behaviour that seem to produce very large fields of behaviour.

Gould compares human behaviour sets to animal behaviour sets and points out that in defining what makes humans unique in the animal kingdom we must inevitably point to our intelligence (Gould, 1996). Gould explains intelligence in humans as having wide fields of behaviour, although he might include other specific behavioural fields that do not seem to be shared with other animals.[28] The larger possibilities of behaviour seem to be related to intelligence, the fact that humans use tools when most other animals do not is a behavioural advantage that could look like comparatively greater intelligence. As such, a wider range of flexibility is definitively human for Gould, and this flexibility could also be described as being creative (Gould, 1996).

Being creative is linked to this flexibility-as-range that Gould describes. In being creative one takes novel and unusual approaches to problems, systems of information. Creativity and novelty clearly require having access to a broad range of behavioural responses to those kinds of environments that produce problems or are composed of complex information systems. In this way flexibility (as Gould describes it) and creativity are very similarly structured. Both require a range of behaviour that is determined by genetic and environmental factors.

Andrew exhibits creative flexibility in his artwork, this is his defining characteristic in the opening chapters of the short story, and he is explicitly contrasted with other robots that do not behave in this manner. If being human is to be flexible, or creative, then Andrew is in an interesting position in that he is not human but *is* flexible. Mr Martin may or may not be concerned with this interpretation but is certainly swayed by the fact that Andrew is creative, and this causes his behaviour towards Andrew to change. Specifically, Mr Martin now includes Andrew in Mr Martins ethical society. Andrew being brought into, or demanding a place, in ethical society is partly based on his capacity to behave in a human fashion, and this theme runs throughout the narrative. Andrew innovates new fields of academic and practical enquiry, Robo-History and Robo-biology; these creative expressions are, if Gould is to be believed, uniquely human. They are uniquely human by virtue of the fact that Andrew makes a life for himself by creating goals to pursue. These goals are not specifically to do with his designed features, they arise out of his creativity and flexibility.

The notion of flexibility as being a defining characteristic of humanity is intrinsically linked to being minded, and makes Andrew seem to be the sort of thing that is minded. Andrew displays his interests through his decisions to pursue independence and personhood, his flexibility allows him to act in such a way as to pursue those interests. Flexibility in the way that Gould describes requires the

[28] For example, some human behaviours (such as language use) seems to be altogether absent in animals.

active pursuit of some kind of goal, people are not flexible if they are acting randomly, and by some stroke of luck land upon their goal. Flexibility is the ability to adapt to a broad set of environments, some of which have never been encountered before. Andrew is the sort of thing that is looked at through the intentional stance, he can speak and make claims about his desires, and he does speak and make claims on his desires. Sally cannot do these things, and as such her ethical status is more difficult to ascertain. The flexibility that he displays is a product of that intentionality, he could not be flexible if he was not intentional.

Andrew makes a series of demands on his society, but he is also welcomed in. The first section of the narrative culminates with Mr Martin, the head of the Martin family and Andrew's owner, giving Andrew a portion of the money that Andrew generates with his 'artistic' works. There is an implicit recognition of Andrew's ownership of Andrew's artwork that Mr Martin demonstrates when he opens a bank account in Andrew's name. He effectively points out Andrew as being the sort of thing that is responsible (not only in the causative sense but also in the moral sense) for the generation of work that has high value. To take the money for himself would be to take some form of responsibility for what Andrew had done, to claim that it was done by Mr Martin in proxy (it is worth noting that he does take some of the money but still feels that Andrew deserves, in some sense of the word, financial compensation). There is a sense of moral inclusion here, Andrew is not strictly a person to Mr Martin, but he is not entirely property either. I think it is a fair assumption that Mr Martin would not extend the same privilege to a dog that happened on a gold vein, the dog might get recognition but would not be entitled to any of the wealth. Mr Martin's actions show that he believes Andrew to be the sort of thing that should be morally considerable. It is to the welfare of Andrew, that aspect of Andrew that 'points' to an interest (Andrew being in a state of preference for something within his well-being), that Mr Martin allocates Andrew his money. This transaction of money indicates that, to some extent, Andrew is welcomed into Mr Martin's ethical society.

Subsection 3: Andrew and Death

Perhaps the most telling event, in terms of welfare and interests, is Andrew's decision to end his life. In the final chapters Andrew has been involved in a lengthy court trial, attempting to gain the status of person. The situation seems bleak, he is not likely to win his bid and he has very few chances to attain his goal. Andrew has replaced most of his body parts with prosthetic organs, the same that any human may have to replace their own, biological, organs. In his own words:

> I have the shape of a human being and organs equivalent to those of a human being. My organs, in fact, are identical to some of those in a prosthetized human being. I have

> contributed artistically, literally, and scientifically to human culture as much as any human being now alive. What more can one ask? (Asimov, 2018, p. 600)

His logic is that many other persons enjoy personhood, despite having these body parts; so, to deny him personhood on the grounds of his biology is unsound. His legal team pursue lawsuits to show that having these prosthetics is not sufficient to deny someone personhood:

> They instituted a lawsuit denying the obligation to pay debts to an individual with a prosthetic heart on the grounds that the possession of a robotic organ removed humanity, and with it the constitutional rights of human beings. (Asimov, 2018, p. 602)

They are largely successful, but Andrew is still denied personhood. Andrew goes to a surgeon, who carefully damages Andrew's computational unit in such a way as to cause his computational unit to begin to degrade. Effectively Andrew will soon cease to exist. This is the turning point in the case and Andrew is granted the legal status of being a person.

There are two major points to consider with this, the first is that Andrew actively destroys his physical self, in direct violation of the second and third law of robotics, described below. The second is that Andrew is given personhood only when he is going to die.

The first point has to do with Andrew's changing understanding of the world. I have touched on the fact that Andrew seems to develop his 'psychological' state in my discussion on his flexibility, or the way that his mental states about the world seem to change, but there is more to be said on it. Andrew's development from his first encounters with the Martin children are significantly different from how he interacts with people in the later chapters of the narrative. As has been stated earlier, Andrew has always had the desire to be a person, and that desire is more openly displayed in the later chapters than the early. He becomes more forthwith and has developed a more subtle understanding of the three laws of robotics in two ways.

Firstly, he purposely has his operation in secret, he states "'"I couldn't tell you, or even the people at [the law firm]. I was sure I would be stopped." (Asimov, 2018, p. 606) This demonstrates a very literal interpretation of the second law of robotics: "A robot must obey the orders given it by human beings except where such orders would conflict with the First Law." (Asimov, 2018, p. 564) This law requires obedience from robots. The spirit of the law implies that implicit obedience should be given to what humans want from robots even if not explicitly stated, it is a directive to obey the will of a human. The literal meaning of the law, to Andrew, is that in the absence of such orders he is able to behave as he desires despite knowing that humans would order him not to if they were to know.

Secondly, Andrew interprets the third law of robotics in a similar way: “A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.” (Asimov, 2018, p. 565)Again the intent of the third law is that robots must preserve themselves. Andrew states “I have chosen between the death of my body and the death of my aspirations and desires. To have let my body live at the cost of the greater death is what would have violated the Third Law.” (Asimov, 2018, p. 606) His active avoidance of the implicit demands of the second law is indicative of what he believes his welfare is tied up with, his interest in personhood.

There are other indications of his personality changing to become more humanlike, and therefore more like a person. His interactions with people become more forceful, in that he aims to have his own interests weighted to the same degree as others. In his interactions in the earlier chapters of the narrative, he is characteristically quiet and respectful to humans. When accosted by the two young men on his way to the library, Andrew refers to them as “sir[s]” (Asimov, 2018, p. 580) despite their treating his presence as object like, and calmly discussing his destruction. Therefore their d3enial of Andrew being an intentional and morally considerable thing. When Andrew is rescued by George Martin, the grandson of Mr Martin. He then claims that he “is well,” (Asimov, 2018, p. 581) when asked whether Andrew is fine. Andrew does not immediately explain that he is in real danger. “[Andrew] would have liked to signal him in some way, but the last order had been ‘Just lie there!’” (Asimov, 2018, p. 581) From this quote we know that although Andrew has very particular desires regarding this event and realises the danger he is in, he does not push those interests due to some sense of meek.

In contrast, in his interactions in the later chapters of the books Andrew is quite assertive of his interests. He states “I have been under obligation to individual members of the firm in times past. I am not, now. It is rather the other way around now and I am calling in my debts.” (Asimov, 2018, p. 600) When approaching his legal representatives to pursue his legal status as person, in the same interaction, it does “[not] even occur to Andrew that he was giving a fiat order to a human being” (Asimov, 2018, p. 600). These are not the same kind of interactions that Andrew has had with humans in the first few chapters of the narrative. He has changed, not randomly or without reason, but rather because he has become more forceful in the pursuit of his interests. He now considers his interests to be of equal value to any other persons. This change in psychology is best explained from the intentional stance, Andrew’s behaviour is only poorly explained through the design stance. This change is indicative of his preferences being important to him, and his orchestration of his own death in the pursuit of it is an indication that he believes those preferences are aimed his well-being (as he perceives it), in the sense of living a good life.

Andrew’s capacity to die is the turning point for his case, soon after he announces what he has done he is given the status of man, of personhood. He states the following:

> Human beings can tolerate an immortal robot, for it doesn't matter how long a machine lasts, but they cannot tolerate an immortal human being since their own mortality is endurable only so long as it is universal. And for that reason they won't make me a human being (Asimov, 2018, p. 606).

Ward E. Jones (2015) argues that death is an important function of being. Jones argues that the fact that non-existence threatens people is also what makes people precious.

Jones (2015) argues that death, or rather the understanding that we will die one day as well as the possibility of dying prematurely, gives our lives a sense of preciousness. To be precious is to be unconditionally valued, and to know that something will cease to be after some period of time is a good starting point to seeing something as precious (Jones, 2015). This notion of unconditional value implies that something/someone finds the object valuable. There are two points to consider here, firstly Andrew considers himself to be valuable, shown in that he pursues personhood so doggedly. Second, he recognises that his value to humans is not of the same kind as the value of a person. Frist I will highlight how Jones organises his notion of preciousness in terms of love, then discuss these two points about Andrew above.

Jones cites a love-based account of preciousness; he claims that a clear way to come to see preciousness is to be in love with the thing that is precious (Jones, 2015). That is not to say that something that is not loved is not precious, just that that preciousness is not recognised in that case. In the case of Andrew, there is no romantic love, but there is clear evidence of being loved by the Martin family, and by Li-Hsing who supports Andrew throughout his court case for personhood. Of particular interest is 'Little Miss', Mr Martin's youngest daughter. She tells Mr Martin "Heavens, he and I have been talking about [Andrew pursuing freedom] for years," (Asimov, 2018, p. 573) and is also outraged at the treatment that Andrew receives at the hands of the two young men on his way to the library. She adamantly demands that her son pursue legal rights for Andrew, telling her son "I'll be watching, George, and I'll tolerate no shirking." (Asimov, 2018, p. 583) Her care is not for Andrew as robot, she helps him pursue both freedom and rights as she believes him to be the sort of thing that is entitled to both, her care is for Andrew as a person, as a precious thing. Of Li-Hsing something similar is to said, she is willing to treat with him as person claiming "I will gladly give you my personal accolade as man," (Asimov, 2018, p. 601) and this develops further and more deeply in her apparent horror towards Andrew at the revelation that he has organised his own death. Both these characters express feelings of unconditional preciousness towards Andrew, both love him as one would love an old friend.

Andrew then is loved, and is precious, but he also recognises that his life is precious in a way that a human's life is not. The first point I wished to explore was that Andrew thinks himself precious, which is evidenced by his dogged pursuit of personhood. The act of that pursuit is indicative of his having the belief that his life is precious. He makes pursuits in art, Robo-History, and prosthetics. That he pursued these things, that he did something rather than nothing is indicative that he sees his time as valuable. Further, he becomes impatient. When speaking to his legal team he states "I grew impatient" (Asimov, 2018, p. 599), and when it was suggested by his legal team that he be patient, he "grimly[states he has an] endless supply of that" (Asimov, 2018, p. 602). It might seem that in his claiming that he endless patience that he is claiming he is not impatient, however the context indicates his frustration. Frustration is often brought about through impatience and impotence, and in this case Andrew experiences both. He deeply desires to be a person in the strongest sense of the word, he is no longer willing to wait for it, and so is impatient. He cannot achieve that himself – and is warned that he will likely fail – and so is impotent.

The second point I wished to consider about Andrew's death is that he realises that his value, his preciousness is not the same as the preciousness that humans have (although his death does change this). He realises that his lack of preciousness is rooted in his perceived lack of an end – that he may not die, and so he is not precious in the way that a human would be. Jones speaks about being fearful of death in two appropriate ways; fear of premature death, and a dread of the inevitable death (Jones, 2015). While Andrew could certainly fear premature death, it is not clear at all that he needs dread inevitable death (at least not in the sense that a human may, by taking reasonable precaution he could feasibly live thousands of years). To be in a society where preciousness that arises from death is shared so universally, he understands that although he might be precious for premature death, he cannot be precious for inevitable death.

Andrew states that "they cannot tolerate an immortal human being since their own mortality is endurable only so long as it is universal. And for that reason they won't make me a human being." (Asimov, 2018, p. 606) He recognises that to be a person is to face the dread of inevitable death, and that to be a thing that has no inevitable death is to be something that can never be accepted among those who must eventually die. Other persons cannot see Andrew as precious because his preciousness does not come from his inevitable death.

Andrew organises his own inevitable death with the clear intention that he will either become a person or cease to be at all. Andrew does this as he wishes to live his life on his own terms, and to do otherwise would be to diminish the preciousness of his own life. It also infuses his life with a new sense of preciousness, one that comes about from dreading an inevitable death. He claims that personhood is within his welfare (as evidenced by his willingness to orchestrate his own death), and his sense of preciousness is tied up in the notion of being a person. Andrew's logic in his organising

his own death is to force the society that he wishes to join to see his preciousness as being like their own. To deny him personhood, his interests, is to be unfair in the way that Singer describes as speciesism. To deny the interests of a thing that clearly has interests, that has a welfare associated with preciousness is to make an ethical mistake.

Chapter Conclusion

Chapter 3 applies the intentional stance to Andrew and Sally. The narrative of *Sally* is structured so that the question of Sally's intentional state is not entirely clear. To Folkers she is intentional, to Gellhorn she is merely designed. The stance being taken up towards Sally carries important moral implications. If Gellhorn is correct, then we have no reason to think of Sally as being morally considerable. In such a case the fear that we have of her, from her actions, is justification for her destruction. Alternatively, if Folkers is correct in taking up the intentional stance towards Sally, then her actions carry very different moral implications. She is to morally considerable in Singer's preference utilitarian ethics She would be morally considerable due to her preferences. She is still to be feared, and likely kept away from us, but her actions against Gellhorn will be retribution against mistreatment. This is in stark contrast to merely being defective and dangerous, as Gellhorn seems likely to see Sally as being. However even if we take up the intentional stance towards Sally, she will never become a person.

In contrast, the intentional stance is taken up naturally towards Andrew. We are obliged to view his interests in a morally considerate manner. By appealing to Regan's welfare and Varner's interests theory, I have shown that Andrew has an interest in being person. I have shown that other characters have loved Andrew and have sought to protect his interests. I have also shown how he has attempted to build a life for himself. His personhood is pursued. I have given evidence that, by virtue of his creativity, he is the sort of thing that can lay claim to the 'title' of person. Finally, I have shown that Andrew's personhood is further reinforced by appealing to preciousness and death. This preciousness is actively pursued by Andrew by his orchestration of his own death.

Conclusion

This thesis seeks to apply Peter Singer's preference utilitarian theory of ethics to AI. First, I explained how Singer's theory functioned, highlighting that interests play a fundamental role in analysing the moral considerability of any particular object. Things that have interests are things that should be morally considered. I then drew out how these interests should be measured against each other in ethical consideration. Singer's conception of interests was insufficient for the purpose of this thesis,

and so I then drew on Regan and Varner to better explain how interests were related to having goals and pursuing them.

The second chapter discussed Dennett's intentional stance, and how it differed from other stances such as the design and material stances. I pointed out how economically explaining the behaviour of some system through the intentional stance was to take up the intentional stance to that system. Taking up the intentional stance implies that the behaviour of that system is sufficiently explained by that system having goals and pursuing them, and so it as safe to assume that those systems actually did have goals and were pursuing them. I explained that this approach, while not ideal, does have the benefit of avoiding the problems that come about when discussing other minds. Primarily that the problem of other minds seems to be largely unsolved in the literature.

Finally, I applied these structures to two narratives, *Sally* and *The Bicentennial Man*. I showed the importance of understanding how people take up stances towards objects (systems to Dennett) when dealing with ethical questions in *Sally*. I then argued for the ethical consideration of Andrew in *The Bicentennial Man*. I explained that his creativity is an indicator of flexibility, and that flexibility implies (from Dennett) that he has goals and pursues them. I then explain how this flexibility and Andrew's decision to end his life are further indicators of his having interests, and so Andrew should be given moral consideration.

This thesis suffers from some problems, most specifically that it focusses heavily on literary notions of AI. We have no reason to believe that AI should behave as Andrew does. I have attempted to remedy this to some degree by also discussing Sally, who is not human-like in appearance or behaviour. She is radically different from Andrew, but there does seem to be some measure by which she would be considered ethically considerable. That being said, this thesis provides no solution *in principle* for some problem of AI rights that could come to pass if AI ever develops to the point that such questions would become urgent. I have not given sufficient guidelines for how to approach such a problem. I have only explained that when AI are sufficiently 'like us' they may be due moral considerability, and I have given guidelines that are useful only to that particular case. I do not consider this to be an abject failure, I think that this approach has merit, it is simply not sufficient for the purpose in general.

The second major problem is that I have not addressed the notion of AI fully. AI is a broad notion and there are many subtleties to be addressed. AI already exists and is in use, although the kind of AI I describe do not. I have not clearly and specifically discriminated between these two general groups, general and narrow AI. Nor have I explored how each of these two groups are subdivided, or the difference between 'weak' and 'strong' AI. In this thesis, about the examples I produced, it was not strictly necessary to do so. However, it does detract from the overall theme of the thesis, that of determining when an alien form of intelligence should be morally considered in general. That is to

say that a general interpretation of the aims of this thesis would have required such discussions, whereas the specific aims of this thesis did not. In this way I think that I chose correctly to not include it, but the overall theme I wished to pursue is not benefitted by that decision.

Lastly, my overall analysis of Singer is shallow. I think that his preference utilitarian approach is a good one even if he seems to have abandoned some aspects of it, but his approach to interests is more complex than I have presented above. In my defence I think that the approach to interests that come about in Regan and Varner are very good approaches, they seem to capture the large part of what it means to be an interested creature. They also fit quite well with the structure that Dennett provides, for this thesis they seem to complement each other well.

Works Cited

Asimov, I. (2018). *The Complete Robot.* London: HarperCollins.

Avramides, A. (2021, March 9). *Other Minds*. Retrieved from Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/entries/other-minds/

Bitterman, M. (1965). The Evolution of Intelligence. *Scientific American, 212*(1), 92-101.

Coeckelbergh, M. (2010). Robot Rights? Towards a Social-Relational Justification of Moral Consideration. *Ethics and Information Technology, 12*(3), 209-221.

Dennett, D. (1998). *The Intentional Stance.* London: Bradford Book.

Frey, R. G. (1977). Interests and Animal Rights. *Philosophical Quaterly, 27*(108), 254-259.

Galen, S. (1989). Consciousness, free will, and the unimportance of determinism. *Inquiry, 32*(1), 3-27.

Gould, S. J. (1996). *The Mismeasure of Man.* New York: W. W. Norton.

Jones, W. E. (2015). Venerating Death. *Philosophical Papers, 44*(1), 61-81.

Millikan, R. G. (1995). *White Queen Psychology and Other Essays for Alice.* Hong Kong: Massachusetts Institute of Technology.

Regan, T. (1976). Feinberg on What Sorts of Beings Can Have Rights. *Southern Journal of Philosophy, 14*(4), 485-498.

Regan, T. (1983). *The Case for Animal Rights.* Los Angeles: University of California Press.

Rowlands, M. (2009). *Animal Rights: Moral Theory and Practice.* Palgrave Macmillan.

Singer, P. (2011). *Practical Ethics* (Third Edition ed.). New York: Cambridge University Press.

Varner, G. (1998). *In Nature's Interests? Interests, Animal Rights, and Enviromental Ethics.* New York: Oxford University Press.