Adventures in Space Racism: Going beyond the Turing Test to determine AI moral standing

by Nicholas A. Novelli

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba in partial fulfilment of the requirements of

the degree of Master of Arts

Department of Philosophy

University of Manitoba

Winnipeg

Abstract: In pop culture, artificial intelligences (AI) are frequently portrayed as worthy of moral personhood, and failing to treat these entities as such is often treated as analogous to racism. The implicit condition for attributing moral personhood to an AI is usually passing some form of the "Turing Test", wherein an entity passes if it could be mistaken for a human. I argue that this is unfounded under any moral theory that uses the capacity for desire as the criteria for moral standing. Though the action-based theory of desire ensures that passing a rigourous enough version of the Turing Test would be sufficient for moral personhood, that theory has unacceptable results when used in moral theory. If a desire-based moral theory is to be made defensible, it must use a phenomenological account of desire, which would make the Turing Test fail to track the relevant property.

Acknowledgements: Funding from the University of Manitoba over the last two years has made this research possible. I would also like to thank the following people:

The best thesis committee a philosopher could ask for in Rhonda Martens, Rob Shaver, and Sarah Hannan. It was truly a pleasure to work with them, and their help and advice was invaluable.

The other fantastic professors in the U of M philosophy department who have helped and supported me, especially Joyce Jenkins, who has contributed a great deal to my command of ethics from the very beginning of my philosophical career, and Carl Matheson, who has always been there in the nick of time when I needed something.

My fellow students, for conversations that have given me intuitions and insights which have proven tremendously useful, and that were tremendously enjoyable as well.

My family, for their support and for listening to all this in its unrefined state, and in particular my sister Tori, for my title and for the late-night *Star Trek: Voyager* marathons that inspired it.

Table of Contents

I. Introduction.....**1**

II: The Disposition-To-Action Theory.....**17**

II.i: The Theory.....17

II.ii: Propositional content and epistemology of desire.....19

II.iii: Problems as an element of a moral theory.....29

II.iv: Empirical detectability and practical usefulness.....40

III: Phenomenological Theories.....**42**

III.i: The hedonic theory of desire and its faults.....44

III.ii: Morillo's theory of desire and the amendments it requires.....48

III.iii: Problems for the theory.....55

III.iv: Empirical detectability and practical usefulness.....59

IV: Some Practical Conclusions.....**65**

Bibliography.....**72**

Chapter I: Introduction

As technology advances, it becomes increasingly crucial that we explore the question of the moral status of Artificial Intelligences (AI). We already have computers with unbelievable computing power, and before long we might have machines far more "intelligent" than any human. We need to know how we ought to act towards such entities to avoid acting immorally. How advanced must machines become before we should treat them as having moral standing – or even as moral persons? Already we have machines with conversational ability nearly equal to a human's – under certain conditions, computers have already passed the Turing Test (or so it is claimed).[1] Even if we deny that they have passed quite yet, there is little doubt that they will do so soon.

Turing's test was originally proposed as a test of cognition: if a machine is indistinguishable from a human in casual conversation, it can think.[2] John Searle's Chinese Room thought experiment[3] attempts to show that even if a computer can communicate in such a way that it could convince an observer that it is human, the internal processes behind it might not be the sort that

1 See: Aron, Jacob. "Software tricks people into thinking it is human". New Scientist Magazine (September 2011), and "Computer AI passes Turing test in 'world first'". BBC News Website, June 9 (2014). http://www.bbc.com/news/technology-27762088
2 See Turing (1950).
3 See Searle (1980).

produce actual cognition. There have been many objections to Searle's argument that attempt to show that the Turing test is in fact a reliable indicator of cognition.[4]

But whether a machine can think or not gives us no moral guidance by itself. It might be relevant, if we adopt a moral theory where ability to think grants moral standing, but that is hardly universally accepted. We might instead claim, as Jeremy Bentham did about animals, that the question is not if machines can reason, but if they can suffer.[5] Without having an idea of what it would take for an AI to be deserving of moral consideration, we have no way to even begin to determine if we are acting correctly. As more and more sophisticated robots become integrated into our daily lives, we become exposed to greater degrees of moral risk. There is a chance that we may make machines with moral standing without realizing it, and proceed to harm them impermissibly without even being aware there is anyone to harm. Making robots to do difficult, dirty, and dangerous jobs in the place of humans might turn out not to be benevolent, but in fact be tantamount to slavery. Conversely, we might make machines that seem so "real" that we instinctively begin giving them moral consideration they do not deserve, to the detriment of humans and non-human animals that do have moral worth.

4 For example, see Minsky (1980) and Churchland and Churchland (1990).
5 See Bentham (1823), XVII.6 (footnote 122).

People might prefer to assist and benefit androids designed to be pleasant over humans that might be abrasive or socially inept, and might opt to do so when a choice arises. We need a way to tell whether we are acting appropriately.

Many people seem to believe that when it comes to AI, we are justified in operating on the basis of a sort of "Moral Turing Test" – if a machine can communicate and interact in a way that seems fully human, some would say that we are thereby justified in treating it exactly the same as we would a human. This seems to be the attitude we are expected to adopt for many pop culture depictions of AI – when an artificial intelligence character is present in a fiction, the other characters generally treat them as people, the same way they treat human characters.[6] In fact, this is often treated as morally required – characters that discount the personhood and moral importance of AIs are often portrayed as being insensitive bigots, and frequently an allegory for racism is present. Many of the things one might say to deny equal treatment to robots on the grounds of their material composition are taken to be direct parallels to the things one might say to deny equal treatment to other races merely on the basis of the colour of their skin.

6 For a few examples, consider Data and the holographic doctor from *Star Trek: The Next Generation* and *Voyager* (respectively), Holly and the holographic version of Rimmer from *Red Dwarf*, and Bender from *Futurama*.

In fact, positions like this are even expressed in some of the philosophical literature on Artificial Intelligence. Rob Sparrow, in criticising this use of the Turing Test for moral decisions, does not find fault with the basic approach of relying on our perceiving a machine to be human, but only says that we are not being strict enough in applying that criteria, and suggests a "Turing Triage Test" to ensure that we are adopting a high enough standard – a machine would have to seem completely human for us to be tempted to choose its "life" over an actual human's, so that should be the true test, according to Sparrow.[7] But the basic methodology of relying on the degree to which a machine outwardly seems human in its appearance and behaviour is not questioned.

However, seeming human is not the criterion for moral personhood (or any other level of moral standing) in any commonly-held moral theory. Very little of the philosophical literature on AI explicitly states which moral theory is being relied on, but it seems important to have some moral theory in mind in order to evaluate the moral arguments about Artificial Intelligences. Only then can we know which properties a machine would have to possess to be worthy of moral consideration. A complication is that many of the properties that form the basis of moral standing in the various candidate theories are

7 See Sparrow, (2004) and (2012).

not directly detectable. We might be justified in using the Moral Turing Test as a practical test for moral standing if we have some reason to believe that it will track the actual presence of the relevant moral properties. Whether that's the case depends on which properties are significant, and how we define them.

One property that occurs in a large number of moral theories is the capacity to have desires. Many theories of welfare include desire as part of the definition of well-being[8] – what is good for a person might be to get the mental states she desires, or it might be that the states of affairs that she desires are actualized. If we adopt any consequentialist moral theory based on such a theory of welfare, then to have moral standing, an entity would have to possess the capacity to have desires.[9] There are also many deontological theories where benevolence will be one of several moral duties,[10] which could take the form of an obligation to contribute to the welfare of any being capable of having welfare. These theories might use the same criteria to determine which entities merit consideration under that duty. There are of course other types of moral theory – non-welfare-based

8 See Parfit (1984) for a thorough catalogue of the candidate theories of well-being – unless we adopt an objective standard of a good life, all other options include what someone desires as an element of well-being.

9 Note that this includes not just theories where we maximize welfare, but ones where we help the least well off, perfectionist theories where we seek to have some individuals living the best possible life, and many other ways we might calculate welfare-based moral duties.

10 W. D. Ross' theory being but one example.

theories might assign moral standing only to rational agents, where being rational might be defined as requiring actions being based on some interaction of beliefs and desires. Other moral theories require us to respect the autonomy of autonomous individuals; being autonomous might involve having desires, and respecting the autonomy of others might mean not going against their desires. Capacity to have desires would then be a requirement to have moral standing in these theories as well. However, my focus will not be on those theories, and I will generally address my arguments to welfare-based consequentialist theories and deontological theories that include a welfare-based duty of benevolence, though the arguments will frequently be applicable to the other theories as well. Under the theories being considered, it is usually possible to have degrees or hierarchies of moral standing, where entities can have less moral importance than humans while still meriting consideration. I will examine whether AIs will be capable of having any degree of moral standing at all, not just whether they might reach the level of personhood.

Theories such as these that treat the capacity for desire as a criteria for moral standing are fairly popular and widely accepted, though there are viable competitors. Moral theories that do not make reference to desire would require a separate examination to determine whether the Moral Turing

Test might be justified for them, which I do not have space for here. I will not argue for any particular moral theory here, but only for the conditional that if we were to accept one of the theories that relies on desire, the Moral Turing Test would not be a reliable guide to moral standing. I will not discuss what it is about desire that makes it an intuitive candidate as a morally relevant property, but will assume that any property picked out by a theory of desire has at least some prima facie appeal for that role. For the sake of argument, assume the truth of some desire-based moral theory (the exact details are left to the discretion of the reader).

There are many controversies about desire, and I cannot examine all of these questions in detail here. For the sake of expedience I will make certain presuppositions about desire. For one, I will take for granted that desires have propositional content. Most theorists hold that desires must be desires *for* something – directed at some proposition or state of affairs. They are thus representational in the contentive sense.[11] This is widely, though not universally, accepted.[12]

Another slighty controversial assumption I will make is that desires do in fact

11 Though not in the indicative sense: desires have propositional content, but do not attempt to indicate anything about the way anything is in the world – see Schwitzgebel (1999).

12 Thagard (2006) is one of the few dissenters.

objectively exist. Some philosophers have argued that there is no justification for believing in the existence of desires in the first place – Paul and Patricia Churchland claim that belief in desires is equivalent to mere superstition.[13] If this were true, we would not be justified in ascribing moral standing even to humans on the basis of moral theories that depend on desires. If that were the case, we would have to reject all desire-based moral theories. However, Andy Clark has argued that there must be some property we are detecting, simply for evolutionary reasons – so many of our decisions are based on the assumption of desires that if we were wrong, surely we would not be so successful.[14] We will proceed under the assumption that humans have something that can be identified as desires, and we can usually trust their reports and our intuitions of when they have them. An interesting question, however, is the status of desires in non-human animals. Our intuitions are less clear and less universal in that case, but the implications of our theory of desire for the moral standing of creatures ranging from amoebae to chimpanzees should be taken into account in our evaluation of the theories' plausibility.

But what is it to have a desire? Everyone agrees that desires tend to have certain features, and there are many familiar cases where it would be

13 Churchland (1979) and Churchland (1981)
14 Clark (1987).

agreed that a desire is present, and in these cases the same types of things generally happen. A person desires to eat a sandwich: he thinks eating a sandwich would be good, so he is motivated to make a sandwich and eat it, and he feels satisfaction when he does so. But which feature is constitutive of desire, and which only tend to accompany desires? Which feature, if it were absent from a state of affairs, would lead us to conclude no desire was present? For the most part, the literature on these moral theories tends to be concerned only with humans, so questions about the nature of desire have not been very thoroughly explored in terms of their implications for moral standing. It seems like we can reliably identify desires in humans, and it is difficult to settle on a definition of desire precisely because all the features argued to be constitutive of desiring – finding something pleasant, being motivated to get it, thinking it's good (in some sense)[15] – tend to go together in humans and other animals, likely for evolutionary reasons. The things that are good for us (under normal circumstances) tend to be things we find pleasant and things we are moved to act to obtain, since an entity that was moved to cause itself pain or that was not motivated to acquire what was good for it would likely soon go extinct. In artificial intelligences, however, these features could come apart, since they could be designed and

15 Though not necessarily "all things considered". It is possible to desire something while thinking it is not good *on balance*, but it seems plausible that you must think there is something about it that is good even if that is outweighed by other factors – for instance, you desire a cigarette because you believe it will produce pleasure, which is good, while realizing that on balance it is harmful because of the health effects.

programmed any way we choose.

If we are to use desire as a criterion for consideration in our moral decision-making, there must be some empirical criteria that give us justification for making attributions of desires. Again, this is not seen as crucial when it comes to humans, since we know that humans are the kind of things that can have desires. But though the competing theories of desire agree about obvious cases, they are not empirically equivalent – there are cases where one theory will say that a human or other animal does possess a desire while another theory denies it. What we want to say about these cases will determine which theory we want to adopt for AIs. Furthermore, it is not sufficient to simply say that desires have something to do with consciousness or volition – consciousness and volition are themselves mysterious and controversial, and cannot be measured or perceived. It may well be that our intuitions are that desire requires phenomenal consciousness,[16] and if we were somehow certain that consciousness and/or volition were absent, we would always deny that desires were present. But those are not things we can test for directly, and any potential tests for them will be extremely controversial, so we must find a practical solution to determining whether the presence of desires is likely.

16 See Worley (1997) for a list of examples showing that we should never ascribe beliefs or desires in a situation where we know consciousness to be absent.

If we accept desires as the criteria for moral standing, there may be a way to justify the Moral Turing Test, if we adopt a theory of desire that makes desires something that reliably correlate with the types of behaviours that we are sensitive to in the Moral Turing Test. This is obviously the case if we accept an interpretationist theory like that of Donald Davidson[17] or like the intentional stance theory of Daniel Dennett[18] – according to this type of theory, what it means for something to have desires is that we legitimately treat it as having desires. Desires are something we project into other entities when it's predictively useful to do so. Obviously any entity that would pass the Moral Turing Test does in fact have the same desires as a human according to this definition, since by hypothesis we could make the very same predictions about it. However, adopting this theory of desire would make it unattractive to adopt a desire-based moral theory (if we want morality to be objective). If we are to make important moral decisions on the basis of desires, there must be an objective fact of the matter about what types of entities have desires and which ones they – since it is a morally relevant property, it cannot simply be a matter of interpretation or dependent on what knowledge and alternative explanation we have available at the given time.

17 Davidson (1980).
18 Dennett (1987).

There is, however, a simple and appealing theory of desire that would give us a definition of desire as a real, objectively existing property, and licence to use the Moral Turing Test. According to the action-based theory of desire, as articulated by theorists such as Michael Smith, desires are dispositions to take certain actions in certain circumstances. We can infer what dispositions humans have based on which actions they in fact take, and they tend to take actions in most situations where we would say that they have desires, so for the most part, this theory captures our intuitions about which desires humans have. And if an AI passes a demanding enough version of the Moral Turing Test, it would seem that it has done so because it has dispositions to behave very similarly to humans in response to similar circumstances, and so it would in fact have the same desires as humans according to this theory. We would therefore have reason to believe that we are correct in treating sufficiently human-seeming robots as persons.

For a defence of the Moral Turing Test to work, our theory of desire must be consistent with the pre-theoretical assumptions that led to adopting desire as part of our moral theories. Those moral theories were found plausible on the basis of our intuitions about desire, and if it turns out that "desires" are quite different from what we thought, that would be a reason to distrust the

moral intuitions that led us to adopt desire-based moral theories in the first place, not to simply plug in our new theory of desire to our existing moral views. Not everything that has to do with desire according to our naive folk-psychological conception of it is something we care about morally, so there can be some deviation, but if we come to a plausible account of desires that gives moral rulings that are deeply inconsistent with our intuitions, we would likely want to reject our desire-based moral theory and still have no way to justify the Moral Turing Test. Therefore, if this theory of desire fails either of these tests, it will be inadequate for our purposes. It must both match our intuitions about desires to a sufficient degree, and it must avoid unintuitive moral results.

I will argue that the action-based theory gives unintuitive results about desires, claiming there are no desires in some situations where it seems desires are present, and attributing desires in circumstances where it seems there are none. Furthermore, even if it is the correct theory of desire, it gives us extremely unpalatable results when used for moral decisions. It doesn't pick out all and only the morally relevant desires, and if used in moral theory, it would grant and deny moral standing incorrectly in obvious cases. Only if a theory matches our intuitions about obvious cases can it be useful for solving the difficult questions about borderline cases such as AI,

and so the action-based theory is of no use here.

An alternative to this is a phenomenological theory of desire, where desire is a particular feeling. Of course, the only way to empirically identify feelings is through introspection. We have to rely on the testimony of others to have any information about sensations other than our own. Since we each can introspect our own desires, we can assume that other humans have similar phenomenological experiences, since they are sufficiently similar to us that there is no reason to suppose there is a difference. We also must assume that we can trust people's own reports of their experiences to be truthful and accurate at least some of the time – assumptions that are not excessively controversial, I think, since under normal circumstances people would have little reason to lie. The methods we use to detect desires in other humans may be imperfect and fallible, but regardless of how accurate or inaccurate the techniques available to us are when it comes to other humans' desires, we are not obviously justified in assuming these methods are at all reliable when it comes to AIs, since AIs are dissimilar from humans in potentially relevant ways. Therefore, we need another empirical criteria to use for entities other than humans. A potential solution is to identify neurological patterns that correlate with phenomenological sensations of desire in humans, which will give us some justification for concluding that an

AI with similar states in its brain structure will also possess desires in the phenomenological sense. If we do identify those structural features, we might be able make AIs that reliably correlate their behaviour with their desires in the correct ways, and then the Moral Turing Test would provide adequate guidance for dealing with them. The point is that in the absence of that knowledge about the makeup of those machines, the Moral Turing Test provides no independent evidence of the the moral standing of AIs.

Philosophical examinations of the neurological structure of desire have been conducted by Carolyn Morillo and Tim Schroeder, which I will attempt to apply to the case of Artificial Intelligences. This analysis will show that if we adopt the phenomenological theory of desire, the Moral Turing Test will not by itself give us reliable guidance. We could, in principle, program a machine to exhibit all the outward indicators of desire while not giving it the structure that would give rise to the phenomenal state, or vice versa, the correct phenomenal states with no outward evidence of them (indeed, we may not even be in a position to know whether we've done those things, given our incomplete knowledge of the neurological basis of desire). But the phenomenological account, as I will show, is the preferable option for desire-based moral theories. Since this account leaves a lot of room for the Moral Turing Test to fail, I will argue that reliance on the Moral Turing Test in our

ascriptions of moral standing to AIs is not justified under a desire-based moral theory.

These two theories are of course not exhaustive of potential theories of desire, nor is it impossible for there to be other empirical indicators. My project, if it is successful, will have succeeded only in showing that what I take to be the most promising way of justifying the Moral Turing Test fails. The most attractive theory of desire that makes desires detectible with some form of the Turing Test is not compelling when combined with desire-based moral theories, and the phenomenological theory provides a viable alternative that is more attractive for use in a moral theory but does not justify the use of the Moral Turing Test. Therefore, we have no immediately apparent justification for relying on the Moral Turing Test if we adopt a moral theory that has desire as a prerequisite for moral standing. It is still possible that such a justification could be found, we just do not have any justification at the moment. Alternatively, we could adopt a moral theory that does not rely on ascriptions of desires. I will not argue for or against that option, I merely wish to make clear what we must commit ourselves to in order to be consistent.

Chapter II: The Disposition-To-Action Theory

In this chapter, I will present the theory of desire that seems best suited to defending the Moral Turing Test, wherein a desire is a disposition to perform actions to bring a certain state of affairs about. I will consider Michael Smith's arguments that this view is better able to attribute propositional content to desires while accounting for the uncertain epistemology of desire, and argue that these arguments fail. I will then turn to other problems for this view that bear directly on moral theory, and show that accepting the action-based theory of desire would render desire completely unsuitable as a basis for moral claims. Finally, I will examine the implications of using this theory as an empirical criteria for attributing desires, even if not a definition of desire. I will show that it is inadequate to that purpose as well.

1. The theory

Under an action-based theory of desire, having a desire is nothing more than having a disposition to do certain things in certain circumstances (generally it is claimed that the desire is the structure in the brain that gives rise to a particular disposition). A desire might happen to correlate with a

disposition to feel certain phenomenological sensations, but this is not essential to being a desire under these views; the disposition to action is what is relevant. This account of desires is usually accompanied by a dispositional account of beliefs, where a belief is also a structure that causes a disposition to act, and the combination of desires and beliefs produce particular actions. The main difference between a belief and a desire, on these theories, is "direction of fit":[19] people tend to be disposed to abandon beliefs when confronted with evidence against them, but desires persist when the world does not conform to them.[20] This theory of desires does allow for unconscious desires (i.e. desires you do not know you have). Although desires interact with beliefs, their existence does not depend on the existence of any particular belief – the posssessor of a desire must be disposed to perform actions they believe will bring about the state that is the object of the desire, but they need not know exactly which disposition causes the actions, or know what particular state their actions are directed at producing. A number of philosophers have held such a theory of desire, such as G.E.M. Anscombe[21] and Roger Stalnaker,[22] but I will focus on Michael Smith's formulation of the theory as among the clearest, as well as one of the few specifically concerned with moral theory.

---

19 See Anscombe (1957), S.32
20 Smith (1987), p. 54.
21 In Anscombe (1957).
22 In Stalnaker (1984).

2. Propositional content and epistemology of desire

Smith argues for the disposition-to-action view mainly by attempting to defeat the rival phenomenological theory, wherein desires are introspectable. His main arguments are based on cases that he claims are intuitively explained by dispositions to action, not phenomenal states ("feelings"). However, all of the cases Smith relies on can be accounted for by a phenomenological theory of desire while respecting common-sense intuitions. One claim Smith makes is that desire cannot be a feeling, because if it were, we could always tell when we had a desire, as we can with feelings. Smith claims we cannot do this – his example is of a man who goes to a certain newsstand that has mirrors behind the counter, but would deny (with all sincerity) that he goes there for that reason. Smith claims that if this man would be disposed to stop going to this newsstand if it removed the mirrors, and would be disposed to go to another newsstand if that one put up mirrors behind the counter, it is clear that he has a desire to look at himself in a mirror while he buys his morning newspaper, even if he cannot introspect any feeling of that being a desire he has.[23]

23 Smith (1987), p. 46.

However, this argument is not persuasive. For it seems that the man's desire in Smith's example is not an intrinsic desire, but an instrumental desire – not a desire for something for its own sake, but only as a means to a state that is desired for itself. And much as we are not always cogniscient of everything our beliefs entail, we need not explicitly desire every means to the satisfaction of our desires. It seems possible that the desire the man has in this case is the desire for a specific sensation, that he happens to get from seeing himself in the mirror when he buys his paper. If you asked him why he buys his paper where he does, he might not be able to give the ultimate reason, but he would likely tell you that he can introspect a feeling associated with buying it there. He may not even be able to articulate this desire, or accurately describe it, but he is introspectively aware of it. Smith has not established that it is possible for an individual to have a desire without knowing he has any desire at all, since there is always the possibility of offering this alternative story. It might still be the case that an individual always knows when she has some desire, even if she cannot say exactly what it is for. She might have any number of instrumental desires in the service of the intrinsic desire, but the intrinsic desire is what is important.

Smith's other example of the supposed failure of a phenomenological account of desires illustrates this as well – Smith describes a situation where

you search the refrigerator not knowing what you want from it, but eventually "realize what it was you wanted all along."[24] I see no reason to take that description literally. We use idiomatic expressions like that all the time while knowing they are not literally true, and on reflection we would accept a paraphrase readily. Requiring a paraphrase in this case should not be taken to be evidence that the phenomenological theory goes against common-sense intuitions. I think people would readily assent that in that case, you desired a means to a particular sensation, and evaluated which of the items in the fridge produced would lead to satisfaction of that desire. Once you formed a belief that a particular item would cause the sensation, you formed an instrumental desire for it. In everyday parlance, we say it was "what you wanted all along" because we cannot spend all day ensuring accuracy in our descriptions of desires.

The cases mentioned so far are ones where a sensation is what is desired, which makes it more plausible that we could introspect a desire without being able to articulate its exact content (sensations are frequently things we cannot explicitly describe). Handling this type of desire would be sufficient for a moral theory that adopts the view that mental states are the only things that are intrinsically desirable and that contribute to welfare. However, we would have reason to adopt the action-based account if we

24 Smith (2011), p. 46.

want to allow for intrinsic desires that do not have sensations as their object, if it is the only theory that can handle them. But that is not the case, since a phenomenological account could still handle such desires. Under phenomenological theories, the desire is a sensation, but its object/content need not be. Let us now turn to an example of an intrinsic desire for an external state of affairs rather than a mental state.

Smith addresses an argument against the phenomenological account intended to demonstrate the possibility of a person believing he possesses a certain desire, but being mistaken. In this case, a man claims to have a desire to be a musician, but he also desires not to upset his mother, who also desires that he become a musician. When his mother dies, he loses all dispositions to attempt to pursue a career as a musician. Smith claims that we should conclude that he had no fundamental desire to be a musician, and so believed himself to have a desire when he did not. However, we need not accept that the individual was mistaken about the presence of a desire, but only that we was confused about whether the desire was intrinsic or instrumental. As Smith describes the case, the man did not believe there was a desire where there was none, but correctly introspected a desire sensation, but incorrectly attributed its object. It seems quite implausible that a person could believe she had a desire but have it turn out she had

none at all, so once again Smith's objection fails to refute the phenomenological theory.

Smith, however, would take issue with this response to his cases, and claim that even if we can introspect a phenomenological sensation of desire, if we cannot accurately introspect the propositional content of desires (as we clearly cannot) we would still need to supplement the phenomenological conception with an "independent and self-standing" account of how desires get propositional content.[25] Smith claims that this reduces the motivation to adopt the phenomenological theory of desire, since he seems to think that once we find this independent account of propositional content, it by itself would give us a workable account of desires, and the phenomenological part would add no explanatory power and could be discarded as unnecessary. However, Smith does not make clear exactly how the disposition-to-action account is supposed to have a non-independent account of propositional content in a way that is relevantly different from the phenomenological account.

One way to defend such a claim would be to say that whatever brain state disposes one to take the actions associated with a desire, it itself represents the desire's propositional content, solely in virtue of its structure and

25 Smith (1987), p. 48.

function and not in virtue of being conscious or phenomenological (much as a map of North America represents North America without having any inherent phenomenological content). Desires would then be representations, with intrinsic propositional content that cannot be separated from them, though vastly different (and completely independent) from other representational states like perceptions. This would be adopting the position described by Timothy Schroeder as being a claim that "to desire that *P* is to have a mental representation that *P* which plays a certain causal role, namely, that of disposing one to bring it about that *P*", and therefore "believing that *P* involves a mental representation that *P* playing one functional role, while desiring that *P* involves a distinct representing object (token) with the same content playing a different functional role"[26] (as opposed to saying that the exact same representations can be believed, desired, both, or neither). If this were the case, the desire would have propositional content directly under the action-based account, in a way that is unlike how it would work in the phenomenological account.

It is unclear whether this is what Smith himself holds, but anyone defending an action-based account of desire on the grounds that desires themselves represent (rather than inheriting their content from beliefs or perceptions) will encounter serious problems. For it is not enough to simply assume that

26 Schroeder (2004), p. 24.

there is some particular structure that causally underlies the observable actions associated with each desire without having any idea where in the brain that might be. And the assumption that there is some easily isolated brain state associated with each desire that could carry its propositional content has proven to be unfounded. Schroeder claims that there are no such representations identifiable in the brain of humans (and so of course we would have no way to begin identifying such representations in AI). "In the whole of the cerebral cortex, there is no plausible home for the scores of mental representations required by our scores of desires",[27] and were there one, we should have found it by now with our neuroscientific understanding of the brain being at the level it is. We have already identified which parts of the brain are associated with each of the other representational states. Schroeder says that it would be implausible to claim that there are representations associated with desires that are non-localized in a way that is very neurologically demanding and completely different to how the representations that we have in fact identified function (for example: sensory/perceptual representations being found in the primary sensory cortex; the limbic association area and hippocampus being responsible for memory representations; etc.)[28] – and yet they would have to be, since the actions involved in desire are multiply realizable. To borrow a case from

27 Schroeder (2004), p. 24.
28 See Kandel, Schwartz, and Jessell (2000), p. 351.

Zenon Pylyshyn,[29] the desire of a person attempting to phone for help after witnessing a car crash cannot be associated with a structure localized in the fine motor control area that produces the hand movements that dial 911, since it might produce the action of dialing 999 if the person is (or believes herself to be) in England, or 0 to get the operator if, say, the 9 key on the keypad is broken. If the phone is not operating at all, it might involve the coarse motor control area causing her to run to find another phone, or it might involve activating the perceptual centres to obtain more information. And the case is far worse with a desire such as wanting to be economically comfortable in one's old age. It becomes implausible to say that the conjuntion or disjunction of the operation of such a large number of parts of the brain in such a massive number of ways is what represents the content of the desire.[30]

On the phenomenological view, by contrast, "desiring presupposes the prior existence of the capacity to bear the content P in a perceptual or cognitive form—in the form of it seeming to be the case that P." The representational part is "found just where the representational capacities for perception and belief are found", and the content is set by how those interact with the

29 Pylyshyn (1986), p. xiii.

30 Note that it is not impossible that AIs might be capable of having such states, since their neural structure might be vastly different from humans'. But that would be no help, since it is

"feeling" part. Therefore, "all that needs to be added to the brain in order to have desires is... neurons connecting the representational capacities" to where the desire feelings are found, "and this is a much more modest demand on brain space than that apparently called for by the representationalist version of the [action-based] theory."[31] For the phenomenological theory, the "content" or "object" of a desire is determined by the belief or set of beliefs that will affect the "desiring" sensation in the right way. This allows us to be fallible about our desires – we might find that if we were to come to believe a certain proposition, we would cease to have the relevant desire-sensations, or they might persist but demand a different interpretation, but this may be because of background beliefs that might be false and that we do not realize are playing a role. There is a sense, then, in which the propositional content is "independent and self-standing", since it already exists in independent, self-standing beliefs and perceptions. But this does not seem to be a problem, since those propositions being the content of the desire is still inextricably linked to phenomenology.

It seems that any advocate of the action-based account should adopt the parallel model of how desires would get propositional content – connection between the representational centres (perception, memory, imagination, etc.) and the action centres. Depending on how the representations are

31 Schroeder (2004), p. 29.

connected to action, we can tell whether given propositions are believed or disbelieved, and desired or not desired. The way to tell which desires a subject actually possesses will require testing which combinations of beliefs will lead to actions and which will not, which will allow us to create a theory of exactly what the desire is. But the action-based theory would thereby lose Smith's supposed advantage over the phenomenological theory for content and epistemology of desire. The way desires get their content under the phenomenological account will not be very different from the disposition-to-action theory if this is the case – they will still inherit them from perceptions and other representational states. It will be no more "independent" in one case than the other.

Once we have this account of the propositional content of desires worked out, we can see that the phenomenological theory in fact gives us a more plausible epistemology of desire than the disposition-to-action theory. Under the phenomenological theory, we are not infallible about our own desires, but we do have privileged access to them. From introspection, we get a type of information about our desires that others cannot have. This does intuitively seem to be correct. We make indirect inferences about other people's desires, but these are frequently subject to revision based on the testimony of the possessor of the desire, who has this special access to a

different kind of information. It seems this is how we generally conduct ourselves in our daily lives. We do frequently make mistakes about other peoples' desires, and we tend to accept their claims when they inform us that we are mistaken (unless we have a clear explanation of why they are making a mistake in a particular case). Under Smith's view, the only reason we might be more reliable about our own desires than other people is that we have more opportunity to observe our own behaviour. If there were someone were observing all my actions, then it seems that if there was any disagreement between us about my desires I would have no more reason to believe I was correct than that the observer was, since we have exactly the same type of information. But except in very unusual cases, it seems that the possessor of a desire is the more reliable guide to its content, or at least has some special access to it, in a way that is not explained simply by their having observed more of their own actions.

Thus, the action-based theory does not have any advantage when it comes to combining propositional content with a plausible epistemology of desire.

3. Problems as an element of a moral theory

Disposition-to-action theories have been criticized on the grounds that they do not provide adequate explanation of actions. Agnes Gellen Callard, for example, says that defining hunger (i.e. the desire to eat) as a disposition to eat is useless, since the mere fact that someone has a tendency to eat when hungry gives us no understanding of why they eat, it only pushes the explanation back – we can still ask why they have the disposition to eat.[32] Similarly, Nagel argues that it might be the case that a dispositional desire is necessary for action without desire providing any explanation or reason for action.[33] However, this type of objection is of no concern for the purposes of my project. Desire need not have any explanatory power to be morally significant – or rather, it needs to explain moral facts, but not any empirical facts. There might be some deeper property that grounds the dispositions, but it doesn't matter what that property is for our present purposes. The explanatory role of desires might be important for other purposes, but a property need not fill that role to fill the role of "desire" as it appears in moral theories. Thus it is not "trivial" that a desire must be present for beliefs to motivate (as Nagel claims), as long as that fact has moral importance.

And indeed, it has been argued that desires in the sense of a disposition to

32 Callard (2008), p. 109.
33 Nagel (1970), p. 30.

action do play a role in morality. Aaron Simmons has argued that we should adopt an action-based account of desires in order to get the correct results when it comes to non-human animals.[34] It is generally assumed that animals such as pigs and cows do not have beliefs about life and death, in virtue of not having concepts of life and death (though it is difficult to be certain due to the present impossibility of adequate communication with them). If they can be said to consciously entertain beliefs at all, they have, at most, very simple beliefs, about eating and avoiding predators and such. Some theories of desire would therefore say that they desire food, and desire safety, but do not desire life. If we accept this, and also adopt a theory of well-being that depends on desire, we would have to say that we do not harm such animals merely by killing them. Simmons points out that such theories attempt to avoid this unintuitive result by claiming that killing animals harms them by inflicting pain, or by depriving them of satisfaction of future desires, but Simmons maintains that it is intuitively obvious that the killing itself harms them. We would therefore have reason to prefer a theory of desire that can accomodate that intuition. The disposition-to-action theory can do so, because pigs are disposed to take actions that will result in a states of affairs in which they continue to live. Therefore, it can be claimed that they desire life. This is true whether or not they are aware that such a state is "life", have any concept of "life" or "death", or are aware they have such a desire

34 Simmons, 2009.

at all. Behaviour such as fleeing from predators, locating and consuming food, etc. are actions, and as such are motivated by desire. Furthermore, this disposition is (presumably) responsive to beliefs – counterfactually, if a pig did have beliefs about life and death, they would affect its behaviours. Thus, if we adopt a dispositional theory of desire, we can say that killing animals harms them – a conclusion which it seems might appear intuitively plausible even to those who think we are permitted to kill animals (obviously, any number of factors could outweigh welfare considerations depending on which moral theory we adopt).[35]

However, there are problems for using this theory of desire in our moral deliberations. The action-based theory does not make the correct distinctions between the states that matter in a desire-based moral theory and those that do not. The property that it picks out is not the correct basis for making moral claims, since it includes habitual actions that are not

35 Note that this type of view is also compatible with non-welfarist deontological moral theories. Smith's own moral theory is that what we are morally obligated to do are things we would do if we were ideally rational, in the sense of internally coherent. Smith argues that to avoid being self-thwarting, an individual must necessarily have certain desires (in the sense of being disposed to do certain things), since these desires are a requirement to be instrumentally rational and coherent, and will desire that others have their rational capacities developed as highly as possible as well. This is defined solely in terms of dispositions to action. Therefore, it is possible for desires as conceived in the dispositional theory to be the foundation for a deontological moral theory, wherein entities that we have obligations towards are those with the capacity for belief-desire rationality in the procedural sense (Smith, 2011). Thus, this type of desire could be relevant even to deontological theories that give no moral importance to welfare – the entities that we have moral standing are still those that have the capacity for desires. I will not explicitly address this theory, however, and will focus on more widely-held moral views.

"desires" in the sense moral theories mean when they say that desires are morally important, and excludes wishes that do have moral importance and should count as genuine desires. If we did adopt the action-based theory of desire, that would only lead us to conclude that desires are not what we should base our moral theory around, and so we would still be left with no way to justify the Moral Turing Test, as will become clear.

Smith claims that we frequently act on the basis of desires without having any real feeling of desire – we cross the road while feeling completely dispassionate about it, but it would be absurd to say that therefore "I cross the road... even though I do not want to!"[36] However, feelings of desire come in degrees, and in that case it is simply that the desire is so faint as to be barely noticeable. For this reason, we could say that frustrating that desire has less moral weight than frustrating stronger desires, an intuition that the action-based theory is less equipped to deal with. Indeed, if there was genuinely no feeling at all that the possessor of the desire was aware of, and if preventing that outcome caused no frustrations of any actually felt desire (including the desire not to be mildly, momentarily irritated, say) it seems plausible to claim that there was no morally-relevant desire at play. Cases like that seem more like someone who crosses the street out of habit, because she has done so many times before. If she were actually going

36 Smith (1987), p. 49.

somewhere different this time, the phenomenological theory of desire allows us to actually say that she crossed the street even though she did not want to, which in fact seems the correct description, contrary to Smith. The action-based theory must say that she simply had an irrational desire to go the way she usually goes, when it seems like she had no such desire at all. Smith is therefore correct that the phenomenological account precludes subconscious desires, but it seems that actual cases of truly subconscious motivation to action are often not correctly described as desires. At the very least, it seems like those "desires" do not have moral importance in the same way that other desires do, and it seems like an advantage for the phenomenological theory to be able to rule out such motivations from the class of morally relevant states. Even if there was no conflicting desire to make such a "desire" irrational, it seems absurd to say that satisfaction or frustration of it would have an effect on well-being. It would not be morally impermissible to interfere with that type of action unless it prevented the satisfaction a real, felt desire, either at the moment or in the future.[37]

The phenomenological theory is better able to handle the sort of desires that

37 In addition, if one wanted to adopt a rationality-based moral theory, it seems having "desires" of that kind does not make an entity a rational being. And if we want an autonomy-based theory, those do not seem to be expressions of an individual's autonomy and it is plausible to say that interfering with them is not an infringement of autonomy in the way that interfering with the satisfaction of a felt desire is, so it seems phenomenological theories of desire fare better on that front as well. There may be ways to make those claims compatible with action-based theories of desire for that type of moral theory, but I will not explore this in greater detail here.

could never lead to action, which is a difficult case for a theory that defines desire as a disposition to action. Take, for instance, the desire that God not exist. It seems no actually possible set of coherent beliefs could dispose someone to take any action in service of that desire, but it could certainly have phenomenological effects. One could say that there is an impossible counterfactual that makes it the case that these are still dispositions to action, but there would be no way to determine if such a disposition exists, and it seems strange to say that any such disposition is what such a desire actually consists of. However, it seems perfectly possible to have beliefs one way or the other, and our feelings caused by those beliefs are what makes them the content of desires. Even for propositions where no change in our belief is possible, we might still determine the content by entertaining different beliefs and their negations, resulting in some modification to the desire feelings. No such solution is available to the action-based theory. The phenomenological account might have difficulty handling desires such as wanting a round square to exist – it is impossible to believe that without being conceptually confused, but it might be possible to desire it. If so, it might be possible to imagine a round square existing in some sense, even if not a very accurate one, and that might be enough to effect the phenomenological changes that are relevant. At the very least, that is more plausible than a non-conceptually confused individual taking any action in

service of that end.

Timothy Schroeder suggests that an adherent of a dispositional account could say that these aren't "desires" in the sense they are concerned with, but are mere "hopes" or "wishes", belonging to a separate category.[38] However, it seems they are desires in the sense we are concerned with in the realm of moral theory – most would say they have the same moral import as desires that could lead to action and should belong to the same category. It would be unintuitive to claim that the satisfaction or frustration of these "wishes" does not contribute to well-being, for example. If someone had strong "hopes" or "wishes" about God, and it turned out God did not exist, many moral theories would hold that this person's life has gone much worse than if God did exist, even if her "hopes" about God weren't the kind of thing that affected her actions. Similarly, Galen Strawson proposed the thought experiment of the "weather watchers", beings that have no capability to act and thus have not evolved structures that dispose them to act, but that have beliefs about the weather and hopes about how the weather will turn out that involve feelings.[39] It seems their welfare is increased if their "wishes" are satisfied rather than frustrated, and it seems that if anyone else was in a position to take action to satisfy those "wishes",

38 Schroeder (2004), p. 20.
39 Strawson (1994), ch. 9.

they would have a moral reason to do so (*ceteris paribus*). Thus from the point of view of morality, for those who wish to adopt desire-based moral theories, there is no distinction to be made between "desires" that lead to action and "wishes" and "hopes" that do not.

Thus, using the action-based theory of desires could lead us to grant and deny moral standing to the wrong entities. For one thing, it would grant moral standing to beings capable only of the type of habitual, instinctive actions that are irrelevant to morality (the unthinking street-crossing kind of behaviour). It is ill-suited to distinguish complex, sophisticated animals (including humans) from even extremely simple organisms, such as insects, paramecia, and protozoa, since those entities do have dispositions to act in that sense. Also, it seems that such a theory will grant moral standing to practically *any* machine. A home computer has "beliefs" and "desires" in the dispositional sense, even a robot vacuum seems to, but it is obvious they have no moral standing. The advocate of the action-based theory might say that there is a threshold of dispositions, wherein a certain level of complexity is required before actual desires are present. However, this is not persuasive. This modification correctly excludes modern robot vacuums, but there is no reason we could not make such a vacuum with more memory and many more dispositions without significantly altering the way they are

programmed. The reason we do not ascribe desires to them is not that there are not enough ways their behaviour is affected by different stimuli, but that we know how they are made. Conversely, a human would not fail to have desires in virtue of having very few dispositions. It is possible to imagine such a person who has extremely simple dispositions, perhaps of the same number and complexity as a robot vacuum, but if these desires were associated with the same phenomenological sensations as in normal humans, such a person would still have desires – and certainly would not fail to have moral standing.

An action-based theory of desire will likely require that the dispositions must be responsive to some beliefs – humans have a disposition to convert carbohydrates into lipids and then adipose tissue when they are introduced to their digestive system, but that does not mean that people "desire" to turn ice cream sandwiches into fat. This is because there are no beliefs that could affect that disposition in any way – there is no information anyone could learn that would prevent them from performing that function. This seems necessary if we are to use this theory morally, as Simmons attempted to – if dispositions to do things that keep an organism alive are by themselves sufficient for having a desire for life, then the ham in your ham sandwich is no more morally problematic than the wheat used to make the

bread – both had a desire for life, and had moral standing on that basis. It might seem that responsiveness to beliefs could be used to make the relevant distinction between the cases – wheat, protozoa and robot vacuums do not have genuine "beliefs", and as such cannot have desires. However, to make this distinction do the necessary work to rule out all the problem cases, a dispositional account of desire would need a far more robust account of belief than a mere disposition. Protozoa, robot vacuums, even wheat do process information, and treat it in a way that tends to conform to the way the world is, and are disposed to respond on that basis. It is unclear what definition of "belief" we could use to distinguish those cases from actual beliefs, but it seems difficult to do so in a way that allows us to reliably identify beliefs without the benefit of introspection. For creatures with vastly different sensory apparatus than us, it will be impossible to tell whether a certain processing of input is akin to our eyes taking in photons and processing it into images (which involves belief), or to our stomachs taking in food and processing it into energy (which does not). How are we to tell which takes place when ants "communicate" by vomiting chemicals into each others' mouths, or when computers interpret the pattern of electrical impulses caused by the input of binary data? This abandons a significant advantage of having a purely action-based theory of desire. Furthermore, even if we have this robust conception of belief, there can still be

counterexamples: Stampe provides a case where a tennis player believes that serving in a certain way will cause him to fault, and his nervousness causes him to do exactly that – he has a disposition to take whatever action he believes will cause him to fault, but he clearly does not desire to fault.[40] Furthermore, even if we have a robust conception of beliefs, the dispositional theory of desires makes it far too easy to get from having beliefs to having moral standing, since all it takes to have desires in the full sense is performing some behaviours that could be affected by beliefs. In essence, it makes beliefs the more important criteria for having moral standing, which does not seem independently plausible.

4. Empirical detectability and practical usefulness

A significant practical advantage of the action-based theory is that it would make it (relatively) easy and straightforward to tell if an entity possesses desires. Science generally proceeds on the basis of identifying dispositions of one kind or another, so if desire is nothing more than a certain kind of disposition that responds to certain other dispositions in certain ways, we can identify it in the same way we identify any other entity accepted by science. We simply look at which

40 Stampe (1986).

conditions lead to which results.[41] And it is easy enough to observe actions. Of course, it will not always be easy to determine exactly which desires a being has – in a certain situation, the dispositions might be blocked by various factors, such that (for example) a desire might be present but never lead to action, because other desires always override it, or because relevant beliefs are absent or false beliefs are present. There will be various equally empirically supported theories about which set of beliefs and desires led to the data that was observed. But it will generally be possible to tell if the being is the type of thing that has desires – this will be the case if it has certain dispositions that disappear in the face of perceptions of contradictory information about the world (beliefs), and if these dispositions affect other dispositions that persist in the face of contradictory information, and result in attempts to force the world to conform to them (which would be the desires). This test could be applied to any entity we choose, even those radically different from humans.

Indeed, even if we do not accept the action-based theory as a definition of desire, it might be useful as a test of the presence of desires for pragmatic reasons. If desire involves phenomenal states, we will have knowledge of

41 There are some general problems with dispositions, of course, but these can be put aside for present purposes, since, as we shall see, any theory of desire will likely have to make some use of dispositions of one kind or another.

our own desires, but no direct way to tell if other beings have desires. If the morally relevant desires reliably correlate with a pattern of actions where we can identify a disposition, then even if desires aren't dispositions, we can use them as the basis for making correct assignments of moral standing.

However, this is only useful if there are not a large number of cases where the dispositional theory clearly gives a result that is at odds with obvious moral intuitions. It seems obvious that dispositions that meet all the criteria of the action-based theory can exist where there are clearly no desires (and certainly where we would likely want to say there is no moral standing).

Therefore, as we have seen, the dispositional account of desire is of no help in justifying the Moral Turing Test. It is no more plausible as a theory of desire than the phenomenological theory, but even if it were, it picks out the wrong properties for a moral theory, and it would assign moral standing to the wrong entities.

Section III: Phenomenological Theories

Based on the arguments in the previous section, we have two options

available: find a new, non-desire-based moral theory, or find a new theory of desire. If we take seriously the intuition that desire matters morally and could form the basis for moral standing, then it is a constraint on a theory of desire that it give results that are consistent with moral intuitions. In this section, I will argue that there is a plausible theory of desire that is compatible with desire-based moral theories, and that it does not justify the use of the Moral Turing Test. Thus, we can continue to maintain a moral theory that uses desire as a basis for moral standing, but if we do so, it will require deeper investigation than social interaction and observation of behaviour to determine which entities deserve moral consideration.

I will begin this section by examining what I take to be the naive phenomenological theory of desire, the hedonic theory,[42] and identifying the problems with it. I will then present a more sophisticated theory proposed by Carolyn Morillo, and identify some lingering problems for it and how to address them. I will thereby arrive at the theory that I believe is the most attractive if we adopt a desire-based moral theory.

42 Though Smith accuses his critics of implicitly adopting a phenomenological theory of desire, very few have explicitly articulated and argued for any such theory. Thus, I am not sure how widely held any version of a phenomenological theory actually is. Schroeder (2004) treats the hedonic theory as one of the main contenders in his examination of desire, but does not provide examples of anyone who explicitly advocates it.

1. The hedonic theory of desire and its faults

An initially attractive statement of a phenomenological theory of desire is what is sometimes called the hedonic theory of desire – to desire a state of affairs is to take pleasure when it seems that state of affairs has obtained, and to take displeasure when it seems that state of affairs has failed to obtain. However, the hedonic theory of desire has some difficulty with theories of pleasure. According to some theories, what makes a state pleasurable is that it satisfies a desire. It is circular and uninformative at best to define desire in terms of pleasure if pleasure is to be defined in terms of desire. Thus, we can only say that to have a desire for a state of affairs is to have a feeling of pleasure when it seems that state of affairs is actualized if pleasure is a distinct sensation. Some have argued that pleasure is such a sensation, like seeing the colour blue, that sometimes forms a part of our conscious experience and it is present whenever we are in a pleasurable state. This theory of pleasure is relatively unpopular, however, since most deny that it is possible to introspect a common feeling that remains the same across the various and diverse pleasurable experiences – there is nothing similar, it is claimed, between the feeling of eating a delicious meal and of apprehending a clever mathematical proof, though both might be pleasurable. To deny that undermines the reason for

adopting many of the desire-based moral theories – the lack of such a sensation is often a motivation for saying that what is good for a person is to have the mental states she desires, rather than saying that a simple sensation of pleasure is good. Also, some people adhere to a philosophy of asceticism, and desire to not feel the simple sensation of pleasure. Others are masochistic, and desire the sensation of pain. These are paradigm reasons for including desire in our theory of welfare, and we should not define desire in a way that prevents it from handling these cases.[43] We should say only that there is a sensation involved in desiring.

A potential problem for phenomenological theories of desire such as this one is that they might not define desire in a way that is acceptable to everyone who uses desire in their moral theory. Part of our goal is to ensure that we can agree on the term, to ensure that people aren't simply talking past each other, and if our definition of desire rules out a number of desire-based moral theories and pushes us to a particular view, then it seems like we are proposing a new concept rather than offering an interpretation of a term common to a number of theories. However, despite how it may seem at first glance, this theory is compatible with most theories of welfare, and we can

43 Of course, there have been various attempts to handle such cases within the context of non-desire-based welfare hedonism, and I do not wish to argue against that theory here. I only wish to point out that many people have been convinced by these types of arguments, and the issue is contentious enough that it is desirable to be able to remain neutral about theories of pleasure.

remain neutral about how various terms such as pleasure and pain relate to this concept of desire. If we wish to claim that what is good for a person is that her desires be satisfied, and desire satisfaction is a type of pleasure, this is still not necessarily equivalent to saying that what is good for a person is pleasure, and does not commit us to welfare hedonism. It might also be the case that disposition to pleasurable feelings is what establishes a person's desires, but what is good for them is that the state of affairs obtains that would dispose them to that pleasure if they knew of it, not that the pleasure is in fact felt. We are still able to hold a view where what is good for a person is something other than a mental state. Alternatively, it might not be the case that all episodes of pleasure involve desire satisfaction, depending on how we define "pleasure". Then only a special kind of pleasure would be good for a person, the desire-satisfaction kind, which is relevantly distinct from welfare hedonism. And of course, welfare hedonism is still a viable option as well. This theory is therefore consistent with a wide range of desire-based moral theories.

The hedonic theory of desire, however, has a significant disadvantage compared to other phenomenological theories, in that it does not posit a phenomenological sensation until it seems either that the state of affairs has obtained or that it has failed to obtain, and so it would give us no way to

introspect a desire until we have reason to believe things have turned out one way or the other. This seems false – before my sports team begins to play a game, I can clearly have a strong feeling of desire that they win, but this feeling is neither pleasure that they win, nor displeasure that they lose, since I do not believe either proposition at that point. Under the hedonic theory, we would have to infer whether we desire something based on expecting that we will feel pleasure if it comes to pass, and displeasure if it fails to occur. But this is unintuitive – if I were to have some novel food described to me, I might expect (correctly) that I would derive pleasure from eating it, and yet it still seems entirely possible for me to not desire it. Furthermore, it also seems possible for me to desire to eat it and yet not derive pleasure from it when I do. It would not be a case of being wrong that I desired to eat it all along – I desired that, but ceased to do so once I found out the results weren't pleasurable. Pleasure and pain can cause our desires, and cause us to revise our desires, but taking pleasure in something is not what it takes for that thing to be desired.

Therefore, desire is a sensation that tends to produces pleasure and displeasure under certain conditions, but is not merely a disposition to feel pleasure or displeasure when it seems to be satisfied or unsatisfied. We can still easily accomodate preference-satisfaction theories of welfare under this

theory – a belief that the state of affairs obtains would remove the desiring sensation. That is what determines the content of the desire. We do not necessarily need to come to believe it for the desire to be fulfilled. We could still maintain that there is some proposition that, were it known, would alleviate the desire – and if that proposition is true, welfare has increased, whether it is discovered or not.

2. Morillo's theory of desire and the amendments it requires

Carolyn Morillo proposes a more sophisticated theory of desire that is still fundamentally phenomenological in nature. The basis of desire, according to Morillo, is a "reward event", an experience that drives all our motivations.[44] Morillo points to empirical evidence that suggests that this experience is always present when we are motivated, and concludes that all our desires are ultimately for this state. One might worry that it would have significant moral implications if Morillo's theory were true – it would rule out the possibility of claiming that some external state of affairs could be good for a person, since they could never desire such a state of affairs intrinsically. It would also mean that all that people ultimately ever desired intrinsically was their own pleasure, ruling out altruistic desires. However, I believe that this

44 Morillo (1990).

conclusion is unwarranted, and in making the necessary modifications to Morillo's theory, we will avoid committing to psychological hedonism or ethical hedonism.

Unlike in the hedonic theory, a feeling must precede the motivation for Morillo's theory to be correct. Morillo explicitly rules out "desires operating prior to, and independent of, any associated reward event"[45] like in the food case mentioned previously. Rather, the feeling is prior, and is what sustains individuals' motivation. This avoids that problem for the hedonic theory and allows that individuals can introspect the presence of their own desires.

Morillo wishes to claim that the reward-event is the only object of desire, but realizes she must answer the following question: "Even if something like [the reward-event theory] is true, why does that not merely tell us more about the mechanism of motivation, about why we have the many different objects of motivation we do have? Why should that mechanism itself count as the only ultimate object?"[46] She claims that the focus is on these outward objects for evolutionary reasons, since obtaining them is what is in our best interests, but "the reward event would still be the aspect of these more complex experiences which is what we are motivated to obtain."[47] But there

45 Morillo (1990), p. 179.
46 Morillo (1990), p. 177.
47 Morillo (1990), p. 177.

is no reason to divide up the experiences like that to isolate which one particular aspect is desired and claim the rest are not. The presence of the reward-sensation might be what causes the whole state of affairs to be desired, yes – but the other features might be what causes the sensation to be present, which should lead us to say that those aspects are desired. Desires can clearly have propositional content that involves things far more complex than pleasure or the basic things that lead directly to pleasure (as Morillo admits), and merely because these desires are likely derived from or explained by basic drives that involve pleasure does not show that pleasure is all we intrinsically desire. Morillo grants that "one might, for different purposes, wish to emphasize the differences among the physiological states (so there would be many different motives), or the differences among the external behaviours and objects (so there would be many different objects of motivation)."[48] Articulating and explaining how desires fit into moral theories seems like a purpose where we would want to do so. Thus, Morillo makes an error in how she connects desire to motivation. The problem might be an ambiguity in claiming that we are "motivated by" a reward-event. That could mean that the experience is what causes us to be motivated, or it could mean that we are motivated to get that experience. The former does not entail the latter, and it would be a mistake to conflate the two senses.

48 Morillo (1990), p. 182.

If Morillo is correct that there is a phenomenal state that is necessary for us to be motivated, it seems far more plausible to say that it *is* desire, rather than that this mental state is the only thing we desire. There is a clear distinction between saying that pleasure is all we desire, and saying that desires are the only things that can motivate. The second, Humean point is far more plausible and widely accepted, and still allows us to make the common-sense attributions of desires that most moral theories depend on. Morillo should say that what she describes as the reward-event is the desire, not that it is the object of our desires.

Morillo grants that desires can sometimes be aimed at things we have never experienced.[49] This cannot be explained by a reward and reinforcement system, and it would undermine Morillo's theory to say that motivation precedes the reward and we are motivated simply by the belief that we will get the sensation in the future. Morillo instead explains these cases with "the hypothesis that we have the ability to envisage future, or possible, or merely imaginary states of affairs, and that such envisagement, particularly when vivid, can link directly to the reward event."[50] Presumably we then assume that we will get more reward if the state of affairs we are imagining actually comes to pass. But there would be no reason to take any action to

49 Morillo (1990), p. 179.
50 Morillo (1990), p. 180.

accomplish things that we know we will never personally experience occuring no matter what we do (for example, things that will happen after our death). Many desires are like that, as Morillo admits – desires for success of one's children after one's death might even exist even in some non-human animals. Perhaps this could be explained by claiming that the more we do to accomplish those goals, and the more they seem likely to occur, the more vividly we can experience the reward-event. This model, however, is problematic in light of cases where no pleasurable "reward" sensation is present. In many cases, our desires involve only unpleasant sensations, and contemplating them is the antithesis of "reward".

Morillo says that her theory "anchors all *positive* motivation in the reward event", but admits that "of course creatures can be, and undoubtedly are, motivated to avoid some things". She presumes "that all such aversion-based learning is also internally anchored, in some aversion event (or events)."[51] Since contemplating these states of affairs will cause us to experience aversion-events instead of reward, the expected effect would be to discourage certain actions rather than promoting them. Morillo is correct that some cases have a positive phenomenal character while others have a negative one, but makes a mistake in apparently claiming that only the positive sensations (the "desires" involving the "reward-event") motivate

51 Morillo (1990), p. 176.

attempts to bring certain states of affairs about, while the negative "aversions" only aim at avoiding certain outcomes by preventing action. In fact, either one might motivate to action, and a desire for a certain state of affairs might produce no pleasant sensations, but only negative sensations when frustrated. A person might desire that some criminal be brought to justice, though the criminal being punished gives him no pleasure, it is just that her escaping punishment upsets him greatly. The difference, then, is not between desire for something to come to pass as opposed to aversion to it, nor is there a difference in motivation to action – pleasurable sensations might be found in the idleness of avoiding doing some difficult or unpleasant work, and desires associated with negative feelings could be just as strong a motivation to action as positive desire. Morillo should not endorse a distinction where something being "aversive" means it "diminished and eliminated operant behavior"[52] as opposed to a real "desire" that promotes action. Both are desires, just with different characters. In fact, it seems an advantage of a phenomenological theory that it can distinguish positive and negative desires based solely on the character of the sensation rather than its effects. But if that's the case, and desire can occur with either positive or negative feelings associated, then Morillo's reward-reinforcement model is flawed. Just contemplating certain states of affairs can produce the relevant

52 Morillo (1990), p. 177 (footnote). Morillo seems to be endorsing that definition, though it's not clear.

sensations, which are not always pleasurable and rewarding but still motivate to action. The sensation is therefore not a reward-event that is the only thing we desire.

This mistake aside, I believe that things are essentially as Morillo describes them. There is a phenomenal state that causes individuals to be motivated, though it can be present even when there is no disposition to action. It is possible for people to do things that are not motivated by this state, but all cases like that seem to be part of a separate class of behaviour that should be treated differently both theoretically and morally, grouped together with (and treated like) things such as digestion and blinking. Morillo admits that it is possible that sometimes perceptions and beliefs could lead directly to behaviour in the absence of this desire sensation, but these cases of "Kantian" motivation[53] are not actually deliberate and rational actions, but "reflex responses, such as the toad's zapping small moving objects ("bugs") with its tongue when they are detected visually."[54] These are not genuine desires, and certainly do not have moral importance – interfering with them would not be morally problematic without an independent reason why we

53 It is interesting to note that both Morillo and Michael Smith, in the papers in which they articulate their views on desire, have as their aim defending the Humean theory of motivation. Humeanism about motivation says that we only act to do what we desire to do, as opposed to a Kantian theory of motivation, wherein beliefs alone can lead people to action. Smith and Morillo each argue for Humeanism on the basis of their theories of desire, despite their theories being wildly different – about as close to opposites as theories can be.

54 Morillo (1990), p. 178.

should not do so. And desire-based moral theories would likely want to deny moral standing to beings capable only of that kind of behaviour. Thus, we arrive at the theory of desire I wish to endorse as the most plausible if desire is to be the basis of a moral theory – desire is a phenomenal sensation that precedes motivation to action, but is not a feeling of pleasure.

3. Problems for the theory

There yet remain a number of objections to the phenomenological theory of desire that must be addressed. The first, most obvious problem is the issue of standing rather than occurent desires. It seems there are many desires we are not aware of at a given moment, creating difficulty for the phenomenological account. A man might desire that his children be successful – when he considers his childrens' success, he might find himself feeling a sensation of desire for it. He might be made satisfied when he learns of increases to his childrens' success, and made dissatisfied when he learns of things that will hinder his children from becoming more successful. However, when he is occupied with tasks that require a great deal of concentration – perhaps a sporting competition, a game of chess, or attempting to solve a difficult mathematical or philosophical problem – it

seems likely that a desire for his children's welfare is completely absent from his consciousness. It seems unintuitive to say that at those moments, he does not possess a desire for his childrens' welfare. Certainly for many moral theories we would want to be able to say that he is made worse off if something negatively affects his children during those times, and contravening a decision he made in the interests of that goal would still constitute interfering with his autonomy during those times. It clearly still contributes to his rationality, as well – if he were to wager his childrens' college fund on a poker game, we would not say he wasn't acting irrationally merely because he wasn't feeling any desire about his children at that moment.

Smith claims that though the phenomenological theory cannot handle these cases, the action-based theory can accomodate them easily, since the disposition to action is present even when it is not triggered. However, Smith errs when he rephrases the claim that "desires are states that have phenomenological content essentially" to "if there is nothing that it is like to have a desire, at a time, then it is not being had at that time."[55] The former does not entail the latter, since it is possible to have a theory where a desire can be had in some sense even if it is not being felt at a particular time, and yet still have desires be defined by their phenomenological content. A desire

55 Smith (1987), p. 48.

might involve a disposition to feel certain sensations when the relevant things are brought to mind, which would still make it the case that phenomenological content is essential to desire. Surely we must assume that even at the moment, the man in question has a disposition to feel the right sorts of sensations when the prospect of his childrens' success or failure is brought to mind. If we say that a standing desire is just a disposition to have an occurent desire when thinking about the right things (or a structure that disposes one to have the occurent desire sensations when thinking about those things), desire is still essentially phenomenological. The differences between the two types of desire could justify treating them differently in various ways, allowing us to accomodate different moral theories. And we still exclude fully unconscious desires, desires that could never become conscious, thus avoiding a problem with the action-based account of desire.

Another potential difficulty for the phenomenological theory of desire is the results it would yield about the changing strength of desires due to states such as depression. Timothy Schroeder claims that a person suffering depression might feel his desires far less acutely than when he was not depressed (which certainly seems to be a standard effect of depression), and that we would not therefore want to say that he had come to desire things

to a lessened degree, as the phenomenological theorist must say.[56] A possible response would be to say that the depressed person's abnormal mental state blocks the feelings, and the person's desire levels should be set by what he would feel under "normal" conditions, i.e. what he would feel if he were not depressed. Schroeder argues that this will not work, since we cannot avoid the unintuitive results by excluding changes resulting from depression without also excluding some genuine cases of altered desires due to "abnormal" brain states – for example, an 89-year-old woman whose syphilis caused her to experience an increase in sexual desire, which she chose not to "cure", considering it to be a real desire.[57] However, this response is not necessary, since it seems perfectly plausible to say that the depressed person's desires are diminished. Schroeder claims that a depressed man who fails to have strong sensations associated with his wife receiving a promotion would say that he does not care any less than if he were not depressed. "'Of course I still want you to succeed, I'm just having a bad patch' is the sort of thing the moderately depressed husband might say to his wife, after being criticized for failing to show happiness upon learning that she has been promoted, and he is likely to be believed."[58] But the real reason the husband has an excuse is that his desire has not diminished *relative to his other desires*. His wife's success might still be one of the most

56 Schroeder (2004), p. 32.
57 Schroeder (2004), p. 32.
58 Schroeder (2004), p. 31.

important things to him, he just feels less desire for everything in his life. This might explain both why people become depressed, and why depression is bad – people subjected to sufficient misfortune sink into depression as an evolved defence mechanism to prevent severe loss of well-being, but it also prevents them from deriving significant changes to their welfare from things that should make them better off.

Thus, the modified form of the phenomenological theory is capable of handling the objections that have been leveled against such theories.

4. Empirical detectability and practical usefulness

But how do we identify the capacity for phenomenological desires? Desires must have propositional content. Under the theory I have proposed, the content of desires comes from sensitivity to beliefs. Introspectively, we can tell *that* we have desires, and *when* we have them. We are not infallible about our desires, since we can be mistaken about their content, but not about their presence. Our desires are responsive to our beliefs, and we have a multitude of beliefs at any given moment, and when we gain beliefs we usually gain many beliefs at once due to things being entailed by other facts

– conjunctions and disjunctions of beliefs, etc. It is not always a simple matter to tell which exact belief affected our feelings. But this does not show that we are not feeling some desire at those times. It is not so easy for entities other than ourselves, but still the representational power of desires can be explained by the representational power of beliefs. A good deal of neuroscientific research has been conducted and the representational states associated with beliefs have been isolated.[59] These states may not by themselves be enough for their possessor to have true "beliefs", but it is enough for our purposes to identify the representational aspect of belief, since that is what factors into desire. It might be claimed that we need a more complex theory of belief to avoid unintuitive results, but this is a separate theoretical issue. It may be that this theory will grant the capacity to have beliefs far too easily – present-day computers have representational states, and it does not even seem to be very difficult to design a computer that has representational states of the same structure as those associated with belief in human brains. It may seem counter-intuitive to say that these machines thereby have beliefs, properly speaking, but though this is a theoretical problem, it is not morally problematic, since the presence of beliefs by itself does not entail any moral facts. This issue, then, can be set

59 See Kandel, Schwartz, and Jessell (2000). We can have a believing-attitude towards the propositions represented in our perceptions or memories, though what it takes to have such an attitude is admittedly somewhat more mysterious.

aside for our present purposes.[60]

We also need to identify the neural state associated with the phenomenological component of desire. Schroeder provides an overview of various candidate structures – the activity of the anterior cingulate cortex, in particular the perigenual region, correlates to certain phenomenological sensations, and modifications to it alter these sensations and attitudes towards certain states (it is unclear the degree to which this is directly correlated with pleasure and displeasure).[61] The ventral tegmental area, and the substantia nigra pars compacta, correlate with reward and punishment, and seem to cause some sensations, though not always pleasure and displeasure.[62] The circuit connecting the nucleus accumbens, ventral pallidum, and brainstem parabrachial nucleus is another candidate, and Kent Berridge claims that it is this structure that is the seat of "liking", which is very similar to desiring.[63] Some of the reasons given to reject some of these structures as the seat of any phenomenological sensations of desire are directed specifically at the hedonic theory, and Schroeder admits they do not

60 Note that this is a problem for some moral theories that claim that knowledge is intrinsically valuable – if the value is not simply that a proposition is known by someone, but value is added for each person who learns a proposition, then it may be that if relatively simple machines can have beliefs, we morally ought to add as much memory capacity as possible to all of them, so we can make them "know" a huge number of facts irrelevant to their purpose – a clearly counterintuitive result.

61 Schroeder (2004), p. 78.

62 Schroeder (2004), p. 81.

63 Berridge (2003).

apply to anything other than "what people commonly denote by ‘pleasure’ and ‘displeasure’."[64] We have identified such structures in humans, and have identified the parallel structures in other animals relatively similar to humans. It becomes controversial with animals with nervous systems more dissimilar to human biology – there has been a great deal of debate about the degree to which fish, for example, possess the capacity to feel anything like desires. But this result puts the borderline where we would expect, where our intuitions about desire and morality are unclear, giving evidence that we are tracking the correct properties. Though more scientific research remains to be done, this is at least a viable research project.

As for synthetic artificial intelligences, there remain some questions about what it would take for them to have desires in this sense. Plausibly, neurons and synapses are just one instantiation of this structure, and the phenomenological properties could be reproduced by similar structures made from silicon and metal. But it would have to be very complex to truly approximate the structure of the brain of even the simplest creatures that seem to have the desire sensations. The structure might be instantiable as a program, but again, it would have to be an incredibly complex program, requiring an unimaginably powerful computer to run, far beyond the capacity of any currently existing machine. According to Anders Sandberg and Nick

64 Schroeder (2004), p. 82.

Bostrom, although we are close to being able to run programs that simulate the arrangement of neurons present in the brains of animals such as rats and cats, current simulations run on immensely powerful supercomputers, and even then "most are a hundredfold to a thousandfold slower than biology,"[65] and "achieving the performance needed for real-time emulation appears to be a... serious computational problem."[66] We are not in danger of assigning desires to entities that obviously fail to have them, like modern desktop computers and smartphones. But even accomplishing this emulation would only produce a structure that might potentially have the capacity to instantiate desires, not necessarily something that actually possesses desires. We can assume "that this is the appropriate level of description of the brain, and that we [will] find ways of accurately simulating the subsystems that occur on this level," but ultimately "we are still largely ignorant of the networks that make up the brains of even modestly complex organisms."[67] In fact, these types of emulations might never be enough, and some of the functions of the human brain might be impossible for any computer program to truly emulate.[68] There remain a number of questions about what it would take for us to conclude that a machine likely possesses desires.

---

65 Sandberg and Bostrom (2008), p. 72.
66 Sandberg and Bostrom (2008), p. 81.
67 Sandberg and Bostrom (2008), p. 83.
68 As argued by Lucas (1961), Dreyfus (1972), and Penrose (1994), among others.

Another possible criterion, suggested by Schroeder, is that actual physical alteration and reinforcement to neural patterns through learning is crucial to desire. If that correlates with desire experiences, it might be the case that the substance of the "brain" would have to be adaptable to possess desires. An artificial brain might pass if nanites could modify it in the right ways, for example. That might be sufficiently similar to what takes place in an organic brain that is capable of desiring. It might instead be argued that requirments like that show that conscious experiences such as desires can only truly be instantiated in organic matter, as Searle claims.

But structures like these are clearly neither necessary nor sufficient conditions for having any particular dispositions to action, and not even for passing the Turing Test. As the performance of modern chatbots suggests, machines with structures completely dissimilar to anything that might possess desires could even pass the Turing Test. And conversely, we could in principle design a machine, made out of whatever substance we like, that has the desire structures but has no dispositions to actions that we would find appropriate to the desires, and thus have desires that are not detectable by the Moral Turing Test. We could create an entity like Strawson's "weather watchers", feeling desire but never reacting or moving at all, deserving

moral standing but not inspiring sympathy from others. This means that the Moral Turing Test is of no use on its own in determining whether machines ought to be granted moral standing. It could only be contingently reliable, in cases where we already know we have created a robot where the presence of actual desires correlates with the behaviour normally associated with them. It could never be the answer to how we determine whether a machine deserves moral consideration, it could only provide a guide of to how act in particular circumstances once we have answered that question already. But we are far from being able to do so. And even if we did, it might be dangerous to become conditioned to complacently rely on assuming that all and only robots that seem like they have desires actually do, when there is always the potential for someone to make a machine with a mismatch between outward behaviour and inner feeling.

Section IV: Some Practical Conclusions

We have seen that if capacity for desires is the correct basis for assigning moral standing, then under the most plausible theory of desire compatible with moral theory, it is possible for a machine to have desires and not show it. It is also possible for a machine to pass the Turing Test and seem fully

human, yet still fail to have desires. But we are not equipped to detect the relevant properties directly, and so it is very tempting to continue to rely on the Moral Turing Test. After all, relying on our ability to recognize desires in other beings has served us well for millions of years of evolution. It will not be easy to simply ignore our impulse to believe the evidence of our senses. It seems, then, that we should avoid creating structures that might be capable of instantiating desires if we are not certain that they in fact do – and also avoid creating human-seeming robots until we are better able to reliably determine the presence of specific phenomenal states. Otherwise, we will be exposed to situations where we might make wrong decisions about whether to consider these entities in our moral deliberations. Some might say that we could simply adopt a policy of "better safe than sorry" – we should just treat all machines that seem human as though they are human, and avoid harming them or doing anything impermissible to them. If they do have desires, we avoid acting immorally, while if they don't, no harm done. However, this is not an attractive option, since it is not always costless to avoid treating machines in ways that would be impermissible if they had desires. Often we must make choices about which of several entities would receive treatment that would be harmful, or which will receive (and which will be deprived of) what would be a benefit, if they were capable of being harmed and benefitted. We cannot simply treat robots as people "just to be

safe" without risking causing senseless harm to humans if we are wrong.

Perhaps then we should just ensure that the androids we make do not have desires, but we can still make them look as human as we wish. As long as people are aware that these entities do not have genuine desires, they will know how to act appropriately, one might argue. However, I do not think this information would be effective at preventing people from acting wrongly, and I think we should avoid making robots that appear too human. Even with relatively simple robots, people develop strong emotional connections – Matthias Scheutz tells of soldiers who form strong attachments to ordinance disposal robots, worryingly, since this seems like exactly the sort of situation that could lead to dangerous consequences from incorrect moral deliberation.[69] He also tells of other similar situations with social and personal care robots. It seems that we can remind ourselves that they are unthinking automata only with difficultly. If a machine were to be made that was nearly indistinguishable from a human, I doubt we could prevent ourselves from thinking of it as human. Rob Sparrow argues that an immediate, primitive moral reaction is required before we ought to treat a machine as a person. However, contrary to Sparrow's arguments, such a response is just as inevitable towards a robot that acts completely human as it is towards an actual human, and yet that response is undeserved and

69 Scheutz (2012).

might lead to disastrous consequences if directed towards a robot that did not possess desires.

Even those who manage to resist the urge to treat such machines as persons and succeed in reminding themselves that these machines are not worthy of moral standing might be socially pressured into doing so. The science-fiction examples that make attitudes towards artificial intelligences an allegory for racism show how severe this might be – if it has been continuously reinforced that certain claims one might make about the correct treatment of machines are analogous to ones that might be made about treatment of other races, and that discrimination against robots is tantamount to racism, that would likely produce severe stigma for anyone who would deny moral personhood to completely convincing AIs. People quite rightly do not want to be racist and treat other races unequally, and merely due to perceived parallels they might treat AIs as equals just to avoid being labeled as some kind of future sci-fi space racist. Therefore, if capacity for desires is the criteria for moral standing, we should avoid making human-seeming robots until we can reliably determine whether or not a given entity has the relevant phenomenal states.

However, this course of action might not be costless either. There are great

advantages to having robots that can be interacted with socially, for example as therapeutic robots to help the elderly and individuals with developmental disorders.[70] Even in cases where social ability is not an obvious requirement of a machine's function, it is a significant advantage in terms of ease-of-use to be able to issue commands and have them confirmed in a natural, familiar way – this is why computer interactivity has moved from punch-cards and entering lines of code towards programs like Apple's Siri. The temptation might be to go further, but I believe we should take a careful look at the actual gains we would acheive from making machines that are even more convincing human analogues and social beings compared to the increased risk of making incorrect moral attributions and the consequences that would arise from that. I suspect that we will reach a point of diminishing marginal utility. Another solution might be to ensure that the robots we cannot help but treat as having moral standing are in fact deserving of that status. We will then, of course, have to determine how to avoid causing harm to them and the costs associated with that.

In fact, giving robots desires might sometimes be advantageous even when they are not required to interact with humans. It is quite plausible that desires are a large part of what makes humans and other sophisticated

70 See Robins, Dautenhahn, Boekhorst, and Billard (2005), and Kidd, Taggart, and Turkle (2006).

organisms successful and effective. Robots that are required to learn and adapt might benefit from having real desires. This raises complex questions for cases such as automated bomb detection/disposal robots whose function is inherently dangerous. Even if they have moral standing, we might be able to minimize bad consequences if we are careful not to imbue them with desires that are likely to be frustrated by injury and death the way most humans' desires are. But it is far from obvious that we are in any position to be sure of which desires we are giving them and how to avoid giving them the unwanted ones. This highlights another potential problem: we are not presently equipped to reliably determine when an AI would have desires, so as we give robots more and more complex structures to fulfill complex roles, we might be in danger of giving them desires without realizing it, and harming them without knowing. It is not desirable to avoid giving any structure that might potentially lead to the capacity for desires, since that might rule out useful and necessary features. As computer techology advances, we have an increasingly urgent need for a reliable way to identify the features that give rise to the relevant phenomenological states.

We could, of course, adopt a different moral theory, that does not make reference to desires. In that case, we need to determine what those theories will claim are the properties are necessary for moral standing, and how we

could tell when an entity possessed them.

Bibliography

Anscombe, G.E.M. (1957). *Intention*. Harvard University Press.

Bentham, Jeremy (1823). *Introduction to the Principles of Morals and Legislation*, second edition. Hafner.

Berridge, Kent (2003). "Pleasures of the brain". Brain and Cognition 52, pp. 106–128.

Callard, Agnes Gellen (2008). *An Incomparabilist Account of Akrasia*. Proquest.

Churchland, Paul (1979). *Scientific Realism and the Plasticity of Mind*. Cambridge University Press.

Churchland, Paul (1981). "Eliminative materialism and the propositional attitudes". Journal of Philosophy 78, pp. 67–90.

Churchland, Paul and Churchland, Patricia (1990). "Could a machine think?". Scientific American 262:1 (January 1990).

Clark, Andy (1987). "From Folk Psychology to Naive Psychology". Cognitive Science 11, pp. 139-154.

Davidson, Donald (1980). *Essays on Actions and Events*. Oxford University Press.

Dennett, Daniel (1993). *The Intentional Stance*. MIT Press.

Dreyfus, Hubert (1972). *What Computers Can't Do*. MIT Press.

Kidd, Cory; Taggart, Will; and Turkle, Sherry (2006). "A sociable robot to encourage social interaction among the elderly". Proceedings of the 2006 IEEE International Conference on Robotics and Automation, pp. 1050–4729.

Lucas, J.R. (1961). "Minds, machines, and Godel". Philosophy 36:112-27.

Minsky, Marvin (1980). "Decentralized Minds". Behavioral and Brain Sciences 3:3.

Morillo, Carolyn (1990). "The Reward Event and Motivation". The Journal of Philosophy, 87:4 (April 1990), pp. 169-18.

Nagel, Thomas (1970). *The Possibility of Altruism*. Oxford University Press.

Parfit, Derek (1984). *Reasons and Persons*. Oxford.

Penrose, Roger (1994). *Shadows of the Mind*. Oxford.

Pylyshyn, Zenon (1986). *Computation and Cognition*. MIT Press.

Robins, Ben; Dautenhahn, Kerstin; Boekhorst, R.T.; and Billard, Aude (2005). "Robotic assistants in therapy and education of children with autism: Can a small humanoid robot help encourage social interaction skills?" Universal Access in the Information Society 4:2, pp. 105–120.

Ross, W. D. (1930). *The Right and the Good*. Oxford University Press.

Sandberg, Anders and Bostrom, Nick (2008). *Whole Brain Emulation: A Roadmap.* Technical Report #2008-3, Future of Humanity Institute, Oxford University.

Scheutz, Matthias (2012). "The Inherent Dangers of Unidirectional Emotional Bonds Between Humans and Social Robots". In *Robot Ethics*, ed. Patrick Lin, Keith Abney, and George A. Bekey. MIT Press.

Schroeder, Timothy (2004). *Three Faces of Desire.* Oxford University Press.

Schwitzgebel, Eric (1999). "Representation and desire: A philosophical error with consequences for theory-of-mind research." Philosophical Psychology 12:2, pp. 157–180.

Searle, John (1980). "Minds, Brains and Programs". Behavioral and Brain Sciences 3:3.

Simmons, Aaron (2009). "Do Animals Have an Interest in Continued Life? A Desire-Based Approach". Environmental Ethics 31 (Winter 2009), pp. 375-392.

Smith, Michael (1987). "The Humean Theory of Motivation". Mind 96, pp. 36–61.

Smith, Michael (2011). "Deontological Moral Obligations and Non-Welfarist Agent-Relative Values". Ratio 24:4 (December 2011), pp. 351-363.

Sparrow, Rob (2004). "The Turing Triage Test". Ethics and Information Technology 6.

Sparrow, Rob (2012). "Reflections on the Turing Triage Test". In *Robot Ethics*, ed. Patrick Lin, Keith Abney, and George A. Bekey. MIT Press.

Stalnaker, Roger (1984). *Inquiry.* MIT Press.

Stampe, Dennis (1986). "Defining desire". In *The Ways of Desire*, ed. J. Marks. Precedent.

Strawson, Galen (1994). *Mental Reality*. MIT Press.

Thagard, Paul (2006). "Desires are not propositional attitudes". Dialogue 45, pp. 151–156.

Turing, Alan (1950). "Computing Machinery and Intelligence". Mind 49:236 (October 1950).

Worley, Sara (1997). "Belief and Consciousness". Philosophical Psychology 10:1, pp. 41-55.