



THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

*On the Granting of Moral Standing to Artificial Intelligence: A
Pragmatic, Empirically-Informed, Desire-Based Approach*

Nicholas Alexander Novelli

PhD – Philosophy
University of Edinburgh

2019

The lay summary is a brief summary intended to facilitate knowledge transfer and enhance accessibility, therefore the language used should be non-technical and suitable for a general audience. [Guidance on the lay summary in a thesis](#). (See the Degree Regulations and Programmes of Study, General Postgraduate Degree Programme Regulations. These regulations are available via: www.drps.ed.ac.uk.)

Name of student:	Nicholas Novelli	UUN	S1562862
University email:	S1562862@ed.ac.uk		
Degree sought:	PhD-Philosophy	No. of words in the main text of thesis:	37592
Title of thesis:	On the Granting of Moral Standing to Artificial Intelligence: A Pragmatic, Empirically-Informed, Desire-Based Approach		

Insert the lay summary text here - the space will expand as you type.

Ever-increasingly complex artificial intelligence technology is being introduced into society, with ever-more impressive capabilities. As artificial intelligence advances, it will become harder to tell whether machines are relevantly different from human beings in terms of the moral consideration they are owed. This is a significant practical concern. As more advanced artificial intelligences become part of our daily lives, we could face moral dilemmas where we are forced to choose between harming a human, or harming one or several of these machines. Given these possibilities, we cannot withhold judgement about their moral standing until we achieve logical certainty, but need guidance to make decisions. I will present a pragmatic framework that will enable us to have sufficient evidence for decision-making, even if it does not definitively prove which entities have moral standing.

First, I defend adopting a welfare-based moral theory, where having the capacity for well-being determines that a being deserves moral concern. I then argue that a desire-based theory of welfare is acceptable to a wide range of positions and should be adopted. It is therefore necessary to articulate a theory of desire, and I demonstrate by reference to discourse in ethics that a feeling-based conception of desire is most compatible with the way ethical theory has been discussed.

From there, we need to establish a test for possessing the capacity for felt desire. This can be accomplished by finding observed cases where a lack of specific morally-relevant phenomenal states inhibits the performance of a certain task in humans. If a machine can consistently exhibit the behaviour in question, we have evidence that it has the phenomenal states necessary for moral standing. With reference to recent experimental results, I present clear and testable criteria such that if an AI were to succeed at certain tasks, we would have a reason to treat it as though it did have moral standing, and demonstrate that modern-day AI has given no evidence as yet that it has the phenomenal experiences that would give it moral standing. The tasks in question are tests of moral and social aptitude. Success at these tests would not be certain proof of moral standing, but it would be sufficient to base our decisions on, which is the best we can hope for at the moment.

Finally, I examine the practical consequences of these conclusions for our future actions. The use of this particular criterion has significant and interesting results that might change things significantly in terms of whether applications of this research are worth the cost and risks.

Abstract: Ever-increasingly complex AI technology is being introduced into society, with ever-more impressive capabilities. As AI tech advances, it will become harder to tell whether machines are relevantly different from human beings in terms of the moral consideration they are owed. This is a significant practical concern. As more advanced AIs become part of our daily lives, we could face moral dilemmas where we are forced to choose between harming a human, or harming one or several of these machines. Given these possibilities, we cannot withhold judgement about AI moral standing until we achieve logical certainty, but need guidance to make decisions. I will present a pragmatic framework that will enable us to have sufficient evidence for decision-making, even if it does not definitively prove which entities have moral standing.

First, I defend adopting a welfarist moral theory, where having the capacity for well-being determines that a being has moral standing. I then argue that a desire-based theory of welfare is acceptable to a wide range of positions and should be adopted. It is therefore necessary to articulate a theory of desire, and I demonstrate by reference to discourse in ethics that a phenomenological conception of desire is most compatible with the way ethical theory has been discussed.

From there, we need to establish a test for possessing the capacity for phenomenological desire. This can be accomplished by finding observed cases where a lack of specific morally-relevant phenomenal states inhibits the performance of a certain task in humans. If a machine can consistently exhibit the behaviour in question, we have evidence that it has the phenomenal states necessary for moral standing. With reference to recent experimental results, I present clear and testable criteria such that if an AI were to succeed at certain tasks, we would have a reason to treat it as though it did have moral standing, and demonstrate that modern-day AI has given no evidence as yet that it has the phenomenal experiences that would give it moral standing. The tasks in question are tests of moral and social aptitude. Success at these tests would not be certain proof of moral standing, but it would be sufficient to base our decisions on, which is the best we can hope for at the moment.

Finally, I examine the practical consequences of these conclusions for our future actions. The use of this particular criterion has significant and interesting results that might change things significantly in terms of whether applications of this research are worth the cost and risks.

Declaration

I declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where stated otherwise by reference or acknowledgment, the work presented is entirely my own.

x
Nicholas Novelli

September 9, 2019

Chapter I: On the Grounds of A.I. Moral Standing

1. Introduction

Artificial Intelligence technology is advancing at an incredible rate. AIs are capable of emulating the behaviour of humans to a greater extent than ever before. It might soon become tempting to think of them as persons and treat them accordingly.

However, it is important to have clearly defined and consistent criteria for when machines would count as having moral standing and be deserving of this kind of treatment.

On February 13th, 2019, the Mars rover *Opportunity* was officially declared lost after numerous attempts to contact it failed.¹ This led to an outpouring of emotion on social media, with people expressing seemingly genuine grief over the rover's "death".² Some people even suggested that a mission should be launched to "rescue" opportunity. These people expressed willingness to expend millions of dollars, resources that could have gone to help humans suffering and dying on earth, to help a robot. However, when thinking about it critically, it is hard to find a good reason to think that this robot has any claim to being worthy of such considerations.

Opportunity had amazingly little processing power and was remarkably simplistic, comparing extremely unfavourably to even the cheapest modern home computers and phones.³ On reflection, people would realise the irrationality of this desire and abandon it, but this points to a problem: that people do not have criteria in mind for

¹ <https://mars.nasa.gov/news/8413/nasas-opportunity-rover-mission-on-mars-comes-to-end/>

² <https://www.elitedaily.com/p/these-tweets-about-the-mars-opportunity-rovers-death-are-full-of-feels-15951268>

³ Its only onboard computer has a 1990's-era 20 MHz CPU and less than half a gigabyte of total memory – see Max Bajracharya, Mark W. Maimone, and Daniel Helmick, "Autonomy for Mars rovers: past, present, and future", *Computer* 41(12) (December 2008), p. 45.

when or why an inorganic machine would be deserving of moral concern. There could easily be other cases where the inappropriateness of these feelings might not be so easily apparent, and people would not so easily recognise whether they were being irrational and make a better decision about the correct course of action. This could easily lead to negative consequences.

While it is clear that the relatively primitive *Opportunity* rover does not in fact have any intrinsic moral standing, the development of more sophisticated machines will eventually lead to situations where it becomes less obvious that they do not possess moral standing. It will become more and more important to have clear criteria for when a machine would have the properties that would make moral sentiments appropriate, so we can evaluate when actions for their benefit would be justified or required. Which criteria we ought to adopt will depend on which moral theory is correct. It is obviously unreasonable to require that we prove some moral theory correct in advance of solving the questions surrounding AI, since we need practical guidance for our actions and to adopt policies that will be accepted by people. It will be necessary to make some assumptions for argument's sake in order to have a starting point. First, I will defend an assumption that well-being is what is of importance in morality. While this is not completely uncontroversial, it is a claim that is common to a wide class of moral theories, thus making conclusions on its basis acceptable to a large majority of people and thereby justifying reliance on it. I will then justify assumptions about the nature of well-being, and establish which properties are necessary to have a level of well-being on that basis. Without firmly settling on a very specific theory to the exclusion of others, it is possible to identify the capacity for desire as something that figures in a wide range of theories of well-being, again achieving sufficiently wide applicability. From there, we can move on to

examine the degree to which we can have empirical evidence for the existence or non-existence of the necessary criteria for well-being when it comes to Artificial Intelligences.

2. The Nature of Moral Standing

In many if not most moral theories, well-being has moral importance, and these are the theories I will focus on. The well-being of entities can be morally important for many reasons. The well-being of a tree might be important because it supports many animals, people might derive pleasure from its presence, and it might be beneficial to the environment. In some contexts, people might even speak of the “well-being” of an inanimate object such as a building or an artwork, in the sense that such objects can be “harmed”, and what happens to them might be morally significant because humans derive value from them and it would require an expense of resources to repair or replace them. But the “well-being” of buildings and trees, and the “harms” that could befall them, is not the same, and does not have the same kind of moral importance, as well-being and harm for humans and many other animals. Robots and AIs will always have the capacity for “well-being” in the same sense as trees and buildings, and it will be morally important not to harm them for the same reasons. The question is whether they could have well-being in the same sense as humans, and thereby have *intrinsic* moral importance.

Moral theories that take the well-being of entities like humans to be of importance are varied and many. The most obvious example of welfare-based theory would be some form of utilitarianism, where we aggregate the welfare of all relevant entities, favouring the greatest total welfare, with each counting for one and no more than one. But this is far from the only type of moral theory where well-being is important.

There are many, many ways that well-being could be morally relevant without the goal being to maximize total well-being – we could be obligated to maximize the well-being of the least well-off, or to ensure that everyone, or as many people as possible, reach some threshold of welfare. These alternative methods of evaluating moral outcomes are available to those who wish to avoid some of the potentially negative consequences of utilitarianism, such as “utility monster”-style cases⁴ or the so-called “repugnant conclusion”.⁵

Even in moral theories where well-being is not the only thing of moral importance, it is usually seen as one of the things that has moral importance. In addition to purely consequentialist theories, well-being is also important in deontological theories, for example that of W.D. Ross⁶ or any other where a duty of beneficence exists. Even in the absence of a duty of beneficence, welfare is important when even just a duty not to harm exists – any such duty, even if it only applies in certain circumstances, must be defined in terms of well-being. There do exist some deontological theories that lack any such duty, such as Michael Smith’s explicitly non-welfare-based theory.⁷ Smith specifically denies beneficence having any value beyond instrumental value, and says that the only real duties involve the rational capacity of entities, not their well-being. If one wishes to adopt a deontological theory like this, my arguments will not apply. However, such theories are not especially popular, and I do not think I am excessively restricting my audience by confining it to people who think that the

⁴ See Robert Nozick, *Anarchy, State and Utopia* (1974), p. 41.

⁵ A phrase coined by J. M. E. McTaggart in *The Nature of Existence* (1927) (p. 452-453), though examined in detail by and frequently attributed to Derek Parfit (*Reasons and Persons*, 1984). For Parfit’s more recent work on approaches that avoid this conclusion, see “Can We Avoid the Repugnant Conclusion?”, *Theoria* 82(2) (2016).

⁶ See W. D. Ross, *The Right and the Good* (1930). For influential contemporary Rossian ethics see Robert Audi, *The Good in the Right: A Theory of Intuition and Intrinsic Value* (2004).

⁷ Michael Smith, "Deontological Moral Obligations and Non-Welfarist Agent-Relative Values", *Ratio* 24(4) (2011).

welfare of at least some entities is morally relevant and therefore ought to figure into our moral deliberations. The most plausible deontological theories will require at least some attention to the well-being of others.

Virtue theories, in particular Confucian theories,⁸ also tend to feature some sort of benevolence as a virtue, and under these theories it will be especially important that judgments be accurate about which entities possess a level of welfare to increase and are thus appropriate targets of the exercise of this virtue. There could also be perfectionist theories that have a goal of having at least some individuals achieve a sufficiently high standard of well-being.⁹ And though the moral standing of robots would not be practically relevant to our actions under it, even ethical egoism is a sort of welfare-based moral theory, acknowledging that well-being is important but saying one should only be concerned with one's own. Thus, there are very few moral positions that will take no interest in well-being, and as long as some importance is placed on well-being my arguments will be relevant. Nonetheless, it is possible to reject these preliminary assumptions, and I will not attempt any novel arguments to convince those who are committed to a non-welfarist moral theory. Working out the criteria under which we would have to consider machines in our moral deliberations under these theories would be a worthwhile endeavour, but is beyond the scope of this project.

⁸ See e.g. Alejandra Mancilla, "The Bridge of Benevolence: Hutcheson and Mencius", *Dao* 12(1) (2013), though for an alternative interpretation see George Rudebusch, "Reconsidering Ren as Virtue and Benevolence", *Journal of Chinese Philosophy* 40(3-4) (2013).

⁹ The most notable perfectionist writings, like those of Hurka (Thomas Hurka, *Perfectionism* [1993]) and Nagel (Thomas Nagel, *Equality and Partiality* [1991]) tend not to put things explicitly in terms of welfare, but there is no reason that a perfectionist theory that does could not be adopted.

3. The Nature of Well-Being

There are many different theories of what well-being consists in. I will show that there is a criterion for having a level of well-being that is common to the most plausible theories of human welfare, namely the capacity to have desires. I will then propose a definition of desire that is compatible with other claims made in moral theory discourse about welfare, and which should be acceptable to moral theorists concerned with well-being (at least more so than the alternatives). This theory is a phenomenological conception of desire that posits a certain functional role for the phenomenal states. In later chapters, I will provide evidence that this functional role cannot be performed with a system that does not possess these phenomenal states.

Let us examine which theories give a role to desire. To begin, I will address a very plausible and intuitive measure of welfare: pleasure and pain. It might be thought that pleasure and pain alone are the correct dimensions on which to evaluate the well-being of sentient beings. However, people are frequently not explicit about what they mean by “pleasure” and “pain”, and in fact when we interrogate the ideas beyond the naïve folk-psychological concepts, we see that desire is often involved. For one thing, it is implausible that there is a particular distinct sensation that is common to all experiences we describe as “pleasurable”. Derek Parfit invites us to “compare the pleasures of satisfying an intense thirst or lust, listening to music, solving an intellectual problem, reading a tragedy, and knowing that one's child is happy”, and says we will find that “these various experiences do not contain any distinctive common quality”.¹⁰ There have been recent attempts to salvage the view that pleasures are all united by a common “hedonic tone”, but these have not gained

¹⁰ Derek Parfit, *Reasons and Persons* (1986), p. 493.

much traction, and even the authors making these arguments tend to acknowledge that this is not a mainstream position.¹¹ One of the ways this theory has been amended to account for objections is to say that there might be multiple hedonic tones, explaining the diversity of experience as well as people's different responses to each type of pleasure. But this is not what we should focus on in terms of moral importance. Not only do the various pleasures Parfit described feel completely different, people value them to different degrees, and it seems plausible that they can do so even while experiencing the exact same basic sensations. But if people value the different sensations to varying degrees, then these sensations are not what is relevant to well-being. If two of the experiences have the same amount of two different kinds of hedonic tone but one person values one over the other for its own sake, this must be explained in terms of something over and above the simple sensation and its properties, and it would clearly not be equally good to give that person either one of the sensations.

One could also consider the case of the masochist who would claim that he desires (under specific circumstances) to experience the sensation of pain, the exact same sensation that most people find aversive. Most would agree that it is not bad for the masochist to be in pain in this way, but would be bad were a non-masochist to feel the identical sensation. It is possible to resist this conclusion by, for instance, saying that masochists must be experiencing a different sensation, or insisting that pain is still bad for the masochist, but is compensated for by a separate sensation of pleasure which is greater.¹² But these analyses do not seem empirically plausible, there is little

¹¹ See Ivar Labukt, "Hedonic Tone and the Heterogeneity of Pleasure", *Utilitas* 24(2) (June 2012).

¹² For a good overview of several philosophical attempts to account for masochism, see Colin Klein, "The Penumbra Theory of Masochistic Pleasure", *Review of Philosophy and Psychology* 5(1) (2014).

evidence for them, and they are less parsimonious than the alternative. This is why it is preferable to focus on a concept of pleasure and pain such that true pleasures are just whatever phenomenal states a person desires to have, while genuine pains are those the person desires not to experience. This position sometimes uses the terms “happiness” and “suffering” to disambiguate that the simple sensations are not what is meant.¹³ The theory where this is what determines a person’s well-being is termed preference hedonism. There might be further restrictions placed on what counts as an instance of happiness, but regardless, desiring it is central to what it means for a sensation to be good for a person. Thus we see that plausible pleasure-based theories of well-being will involve desire prominently. There are of course still responses available to those who favour the distinct sensation view, and I cannot definitively rule out this position, but I have shown that a view that incorporates something beyond simple sensations of pleasure and pain has good reasons to recommend it and is the prevalent view in the literature. Thus it is justified to proceed from the stipulated position that having the capacity for simple sensations is not sufficient to have moral standing merely because such sensations could have “hedonic tone”.

There are many other theories of well-being that define well-being in terms of what a person desires without defining it as happiness. These include theories that allow all desires, or ones where relevant desires are restricted to self-interested desires, those about one’s own life, or those that one would hold given full information, among others.¹⁴ There are nevertheless some theories of well-being that are not defined in terms of what people desire. These include objective list theories like, for example,

¹³ See Jamie Mayerfeld, *Suffering and Moral Responsibility* (1999).

¹⁴ See, for example, Richard Brandt, *A Theory of the Good and the Right* (1979).

that of Thomas Scanlon.¹⁵ However, there are objections to theories like this on the grounds that by failing to respect peoples' desires, they return unpalatable results about what is good for a person. As Peter Railton says, "It would be an intolerably alienated conception of someone's good to imagine that it might fail in any such way to engage him."¹⁶ Thus, a possible modification is to require that in addition to being on the list, for something to truly be good for a person, it must be desired as well, which would be sufficient to show that capacity to have desires is necessary for an entity to have a level of well-being that could be affected by our actions. Objective list theories as well, when they account for the alienation objection, will have a significant role for desire.

Other theories are sometimes phrased as being about having "preferences" or "pro-attitudes" towards things, rather than using the term "desires". Despite the use of different terms, I think there is little disagreement about the proper class of application. Objections to the use of "desire" frequently hinge on alleged conceptual problems having to do with an overly-technical definition. But the sense of "desire" being used should not be excessively restricted. One type of objection that has been levelled is that it is impossible to desire things that you already have (and do not need more of), or have desires about the past rather than the future.¹⁷ But this is clearly not what is meant by people who employ desire as a morally-significant term, and this can be solved simply by stipulating that "desire" applies to such cases as well. The moral theorists who advocate desire-based theories would presumably be receptive to such a modification, since they would likely not wish to insist on

¹⁵ Thomas Scanlon, *What We Owe to Each Other* (1998).

¹⁶ Peter Railton, "Facts and Values", *Philosophical Topics* 14(2) (1986), p. 9.

¹⁷ See L. W. Sumner, *Welfare, Happiness, and Ethics* (1996), p. 124 and 129.

maintaining those distinctions in the face of these type of objections. To do so would render their views implausible and arbitrary, and there does not appear to be any dialectical reason to prevent them from avoiding those distinctions. This will result in “desire” covering a range of other pro-attitudes that have subtle distinctions, such as wanting, wishing, hoping, *et cetera*. Although there are some debates about the differences between the various pro-attitudes, the distinction seems to be to a large degree semantic, and at any rate the precise differences are not directly relevant here. We can continue to use the term “desire” in this less restricted sense, as is typical in the moral discourse. Thus theories involving “pro-attitudes” also fall under the umbrella of desire-based theories of well-being.

Thus, we have seen that most of the prominent theories of well-being, despite the great diversity among them, all have in common that they posit desiring something as a crucial component of what makes that thing good or bad for a person. For an entity incapable of desiring anything, then, nothing could have any effect on its well-being, and it would be needless to consider it in our moral deliberations. The capacity for having desires is a prerequisite for having a level of well-being like that of a person, and is thus a requirement for having moral standing on that basis.

4. The Nature of Desire

Given that desire is a crucial component of moral standing in these theories, it is important to know what exactly “desire” is in order to know whether AIs could have moral standing on that basis. However, the majority of moral theorists who employ the word “desire” do not give a precise definition of it. It might be that they do not think it is important to do so, and intend to plug into their moral theory whichever theory of desire turns out to be correct, leaving that up to the philosophers of mind to

figure out, and in the meantime will trust that our intuitions about desire will be sufficient. Our rough pre-theoretical intuitions about desire might indeed be adequate when dealing only with humans, since the various properties we associate with desire almost always go together in humans, and we clearly have the capacity for all of them. However, in other types of entity, these properties could come apart as a matter of course, rather than solely in fringe cases, and the moral consequences will change dramatically when substituting in one theory of desire for another. We must decide which property or properties are essential components of desiring, and which are merely contingently associated with it.

Theories of desire, for present purposes, can be grouped into three main types. One group claims that what is essential to desire is something in its phenomenal character, an experience of desiring.¹⁸ Another claims that desire is very similar to belief (in some cases explicitly stating that it is just a type of belief), and thus focuses on propositional attitudes and endorsement.¹⁹ The last class defines desire in terms of behavioural dispositions.²⁰ I will argue we should use phenomenal experience as our standard for desire in welfare-based moral theory, as it is the most consistent with the way the term “desire” is used in discussions of such theories.

Each of the three main candidate properties has been argued for as the fundamental criterion for desire for several different reasons. But the arguments from philosophy

¹⁸ The theory that most clearly advances this claim is that of Carolyn Morillo (see Carolyn Morillo, “The Reward Event and Motivation”, *The Journal of Philosophy* 87(4) [1990]), and such a theory has also been argued for by Galen Strawson (see Galen Strawson, *Mental Reality* [1994]).

¹⁹ Modern examples include John Broome, “Desire, Belief, and Expectation”, *Mind* 100(2) (1991) and Huw Price, “Defending Desire-as-Belief”, *Mind* 98(389) (1989).

²⁰ Michael Smith is the paradigm example of fully endorsing this claim and all of its consequences (see Michael Smith, “The Humean Theory of Motivation”, *Mind* 96(381) [1987], “Reason and Desire”, *Proceedings of the Aristotelian Society* 88(1) [1988], and *The Moral Problem* [1994]), and it was earlier argued for by Elizabeth Anscombe (G.E.M. Anscombe, *Intention* [1957]).

of mind or philosophy of psychology contexts might not be applicable to the use of the term for well-being-based moral theory. I will examine each of the theories and the arguments made in their favour, but I will primarily focus on the arguments for and against each candidate theory that affect their compatibility with the use of the term “desire” in the discourse in moral theory. I will not focus excessively on the moral verdicts that theories would return and whether these results are intuitively plausible, since that will depend heavily on one’s other moral commitments and ideologies and will vary significantly from person to person. Rather, I will discuss whether the theories in fact return the results that are explicitly claimed in arguments where they are used.

Reasons based on desires are frequently given for adopting some moral theory over another, and if a theory of desire leads us to exclude something that is used as a purported motivation for one of these views, something that is argued to be a morally important case of desire (or indeed, if it would compel us to include something that is used as an example of a non-significant fact that ought to be excluded), this would count as a strike against it as the correct concept of desire. The different sides in the debate about which is the correct theory of well-being certainly act as though they are using the same concept of desire and having a substantive disagreement about the correct application of it, and this should be preserved. For example, one classic thought experiment about the correct theory of well-being comes from Derek Parfit; it is frequently referred to as the “stranger on a train” case (despite the fact that Parfit’s original example makes no mention of a train).²¹ In this case, a person meets a charming and sympathetic traveller who tells her that he has a terrible disease but

²¹ Derek Parfit, *Reasons and Persons* (1984), p. 494.

is hoping to be cured of it. The person forms a strong desire that this traveller gets better. She never sees nor hears from him again. Parfit says it is implausible to say that whether or not the stranger ever actually overcame the disease has any affect whatsoever on the person's well-being, thus challenging the desire-fulfilment theory of well-being.

Another famous case relevant to this debate is Robert Nozick's experience machine.²² Nozick asks us to imagine a machine that will control all our experiences, leading us to feel as though we are experiencing the most pleasurable possible life, all while we are in fact remaining motionless attached to the machine. Nozick believes that people will have the intuition that the experience machine is not good for us, which he says is explained by the fact that it does not fulfil all of our actual desires, only desires for particular sensations. Thus Nozick says that the experience machine proves that all of our desires play a role in our well-being, and that people's rejection of the experience machine is a reflection of the fact that it is important to our well-being not only that we feel as though we are doing the things we want to do, but that we are actually doing those things.

When people have different intuitions about cases like Parfit's stranger on the train or Nozick's experience machine, they tend to argue about whether certain types of desires should be counted as morally important, not about what counts as a desire. The standard responses to the cases do not include suggesting that those presenting them were not imagining an actual desire but had a faulty idea of desire in mind, or saying that if they reflected on the concept of desire they would see that their

²² Robert Nozick, *Anarchy, State, and Utopia* (1974), p. 42–3.

description of the situation is impossible. Our theory should avoid, as far as possible, making it so that any side in a live philosophical debate is wrong by definition.

Insofar as the discussion seems to have been proceeding fruitfully, we should attempt to accommodate as many as possible of the views and arguments being advanced, rather than making it so that the sides are in fact talking past each other.

Furthermore, if a theory of desire is incompatible with standard claims made in moral philosophy in general, and those using desire in proposing a moral theory feel no need to address or discuss the fact that their view has significant consequences for important debates in the field, that could also be evidence that such a theory is not what they had in mind. Thus, the primary focus will be on finding what best corresponds to the concept that moral theorists use, not in finding something that could be termed “desire” but that would be incompatible with the usage in moral theory.

i. The Phenomenological Theory of Desire

A phenomenological account of desire says that what makes something desired is a particular feeling that one has. To desire something means to feel a certain way, perhaps when contemplating or entertaining the idea of it, or just in a way such that possession of it would change or replace that feeling in a relevant way. This theory of desire has intuitive appeal. The common usage of the word “desire” most frequently implies a strong feeling, so much so that it would sound somewhat strange to phrase it that I “desire” a copy of the morning newspaper, or something equally mundane. When the word “desire” is used in non-technical contexts it is frequently preceded by feeling-words like “burning” or “fervent”. Of course, I do not mean to restrict “desire” to only this poetic sense, it also includes the weaker cases for which we would not use “desire” in everyday speech. But the existence and prevalence of that usage says

something about the connotations of the term – that the stronger the feelings are, the more paradigmatic a case of desire it is.

This theory does not imply that we have infallible access to the content of our desires. This is an objection that has been levelled against the phenomenological view by Michael Smith, who says that we can be mistaken about our desires in ways that this view does not allow for.²³ We have infallible access to the contents of our own experience, so the argument goes, and cannot be mistaken about whether we are having an experience of something even if we can be mistaken about whether our experience corresponds accurately to the real world. If a desire just *is* having an experience, there is nothing we could be mistaken about to explain a mistake and therefore we cannot be mistaken about our desires. But this happens frequently – Smith invites us to “consider cases in which you go to the refrigerator convinced that there is something in particular that you want, though you aren't quite sure what it is. Then, while looking at the contents you suddenly, as we should put it, 'realise what it was that you wanted all along'.”²⁴ However, this is not a serious threat to this position for present purposes. People frequently make mistakes about their own feelings, and might, for instance, believe they were experiencing crushing ennui before realising they were just hungry. Nobody would claim on that basis that hunger and ennui are not phenomenological in nature. Furthermore, the difference between a desire for an apple and the desire for a pear need not be determined in virtue of any feature of the experience that would be introspectable.

²³ Michael Smith, “The Humean Theory of Motivation”, *Mind* 96(381) (1987).

²⁴ *Ibid.*, p. 46 (footnote).

The phenomenological account of desire claims that what is constitutive of desire is a phenomenal state that has certain causal features. Some of these causal features are the dispositions to change in response to relationships to their object – a desire is *for an apple* in virtue of facts like that the sensation would be responsive to eating an apple, or that contemplating an apple affects it in relevant ways, *et cetera*. A desire-sensation will also have causal powers in terms of its functional role in terms of its effects or potential effects on behaviour. I wish to remain as neutral as possible about which exact causal powers desires will have to retain compatibility with as wide a range of phenomenological theories of desire as possible. Carolyn Morillo²⁵ has advanced a view that posit a phenomenal state playing this role. Dennis Stampe²⁶ and Graham Oddie²⁷ also have views where a phenomenal experience is constitutive of a desire, albeit a specific type of experience: a perception, described as a perception of “goodness”. Timothy Schroeder²⁸ has a view that makes a *representation* key to having a desire, specifically a representation that figures into reward/reinforcement learning. The concept of a “representation” is notoriously complex, ambiguous, and controversial. If these representations are necessarily phenomenological in nature, this theory will also fall into the category of phenomenological theories of desire (a perception, for instance, is a representation that is necessarily phenomenological in nature, though there could be very different kinds of sensation that are also inherently representational). All these views differ significantly from each other, but in the elements relevant to my project they agree: a phenomenal experience is essential to having a desire.

²⁵ Carolyn Morillo, “The Reward Event and Motivation”, *The Journal of Philosophy* 87(4) (1990).

²⁶ Dennis Stampe, “Defining Desire”, in *The Ways of Desire* (1986).

²⁷ Graham Oddie, *Value, Reality, and Desire* (2005).

²⁸ Timothy Schroeder, *Three Faces of Desire* (2004).

One objection that might be raised to this view is that we are often not experiencing any particular desire, and yet it would be incorrect to say that we do not have any desires at those moments. A person, it is argued, might have a deep and permanent desire that his children succeed in life. And yet, in a moment while he is perhaps occupied in concentrating on some task, he is not thinking of his children at all and no sensation related to this desire forms any part of his experience. It seems wrong to say that he does not desire that his children succeed during that time, but will only come to desire that state of affairs again at a later time. Desires are far more stable and enduring than the phenomenological theory would seem to imply.

An important thing to say about the phenomenological views is that one need not be currently experiencing any sensation to be described as having a particular desire. Any plausible theory of desire will allow for standing rather than occurrent desires, and the phenomenological theory will accommodate this by claiming that even if a desire is not being experienced, or perhaps even has ever been experienced, it might still be held if there is a disposition to feel the right sensations if the right things were brought to mind. Thus the person in the example, were he distracted from his task in an appropriate way by something related to his children, would feel his desire for their success, and so he has that desire even while it is the furthest thing from his mind. Thus this objection is not a concern for a sophisticated phenomenological theory of desire.

It is important for the phenomenological desire theorist to clearly distinguish which phenomenal state is at issue. One phenomenal theory of desire that has frequently been criticised, and which is in fact unsuitable, is the hedonic theory, which defines

desire purely in terms of pleasure and pain. Under this view, to desire something is to be disposed to take pleasure when it occurs, and to experience pain and frustration when it fails to occur. Obviously this is not a viable option if pleasure and pain are defined in terms of desire, as they are in theories such as preference hedonism as mentioned previously. If “pleasures” are to be understood as the phenomenal states one desires, and “pains” as mental states one wishes to avoid, it would be circular and uninformative at best to define “desiring” something as taking pleasure when that state of affairs obtains. Furthermore, some moral theorists specifically argue that even things that only make you unhappy can be good for you if that’s what you desire. Imagine you are involved in a monogamous romantic relationship. You are very happy in this relationship. Your partner convincingly professes to be loving and faithful, and you have no suspicions to the contrary. If you are not told otherwise, you will never find out this is not the case. Many people would report that they would nevertheless want to know if they are being deceived in this, and would claim that it would therefore be better for them to know the truth, even though it would cause them no pleasure, and would in fact be extremely painful, to learn this.

It could certainly be argued that the people who make those claims are mistaken, and that it would not in fact be good for them to discover the truth that their relationship was a lie. However, when people make that argument, the reason given is not generally that the people didn’t truly desire the truth because they would find it unpleasant when it actually occurred. Rather, one would claim that those people desired the wrong thing, something that was in fact bad for them, since it conflicted with their other, more relevant desires. Thus, the hedonic theory is not compatible with the discourse on well-being. We should instead say that desiring something does not make it in itself a pleasure, but is generally a good guide to what will be a source

of pleasure. So, for example, for me to desire a pear does not mean that I take pleasure from eating a pear, but rather that if I see a pear, or think of a pear, I feel a certain way, experience certain things, that will tend to produce certain behaviours.

We are better served to think of desire-sensations in terms like “wanting” and “liking”, pro-attitudes associated with phenomenal states but not identical to pleasure. Essentially, this theory posits that we experience or represent things as being desired, as good, or as rewarding, as a direct part of our experience, in the same way that we experience perceptions or imaginings as being blue. It is a standard cause of pleasure, but this is not a necessary connection. This allows desire to play the necessary role in our philosophical theorizing, while respecting our folk-psychological intuitions about it.

ii. The Propositional Theory of Desire

There are those who advocate the view that the simple pro- propositional attitude, the willingness to assent to something’s goodness, is sufficient even without a phenomenological component. This has been advanced as the correct view of desire by Huw Price, among others.²⁹ Under this view, desires are treated essentially as a species of belief, and merely the disposition to state or agree to the relevant proposition would be sufficient for someone to be described as having a certain desire, even in the absence of any particular feeling (or anything else). This might be phrased as having the disposition to assent to the relevant proposition under conditions that are idealized in some way (i.e., having other relevant information), to

²⁹ See Huw Price, “Defending Desire-as-Belief”, *Mind* 98(389) (1989).

allow for individuals to be mistaken about their own desires, in the same way that subconscious beliefs are sometimes posited.

However, this theory encounters significant problems in its application to moral philosophy. It can handle, without too much difficulty, the case of someone desiring something they do not believe to be good, as seems to happen frequently. At first glance, this seems problematic, as it would amount to saying that such a person believes something to be good, but doesn't believe it to be good. But this is acceptable if we understand it to be a claim that such a person believes that thing to be good in one way, and/or to some extent, and bad in different ways. A person might desire a cigarette, believing it to be good in that it will bring pleasure, while still believing it to be bad on balance because of its health effects. The problem only arises when we examine the reverse case. It seems reasonable to imagine a case where someone genuinely believes something to be good, but would claim to fail to desire it at all. Imagine someone saying, "I agree that it would be good for me in all sorts of ways to learn yoga, I accept that it would have beneficial effects on my mind and body, I have sufficient spare time and I might enjoy it, and I have nothing to lose by trying. But it does not interest me at all, I have no desire to do it." We might think such a person was irrational, but not conceptually confused. It seems very unsatisfying to reply that such a person does not genuinely believe that thing to be good, or that they desire it but don't realize it. I see no reason to assume that such a person is using concepts incompetently, and so these seem like entirely ad hoc responses.

In fact, a frequent objection used to motivate the use of desire in theories of well-being is a type of case where a person believes something to be good but doesn't desire it – for example, someone convinced by a religious cult that it's good for them

to be subject to degrading torture. It is very natural to say that being tortured is not good for such a person even though he sincerely believes it to be good because he does not desire it, thus providing a counterexample to the claim that all that is relevant is whether a person believes a state of affairs is good. This objection would be meaningless if we adopt the view that thinking that something is good is equivalent to desiring it. And in fact, people do not tend to respond to this objection by suggesting that the way to ensure that this torture is actually good for that person is to ensure that they really and truly believe it's good. It appears to be generally accepted that such a person could be completely convinced this is good but still not desire it, and I'm not aware of anyone who says that it's still good for that person, or would be good for them if they had enough faith.

The propositional theory of desire would also render certain debates in moral philosophy completely incoherent. A very significant argument for centuries has been the debate between the Humean, desire-based theory of motivation and views in the Kantian tradition that claim that rationality and belief can provide motivation that can override desires. On Kant's theory, doing something merely because you desire to is less valuable than doing it because you understand it to be right. We must (and can) put aside our desires and do what duty demands, according to this view. Hume's view is that reason is always slave to our desires, and we ultimately do whatever we desire. This is treated as a substantive empirical claim about human psychology, not a tautological definition. Obviously if desire is simply believing something to be good, this debate is meaningless. The problem is not that this definition of desire decides the matter, but rather that it makes the entire discussion complete nonsense, which is certainly not how it has been treated. Surely we do not want the result that Kant and Hume and everyone who followed them in this debate

had no idea what they were talking about, but should rather say that moral theory was clearly working with a different concept. The fact that the moral debate about it exists implies that it's conceptually possible according to the concepts being employed.

Finally, it is frequently claimed that at least some non-human animals have desires, and have morally-relevant well-being on that basis. But it's not at all clear whether any non-human animals could make judgments about goodness or indeed even have any implicit understanding of propositions about goodness. Whether it's true or not that non-human animals deserve moral standing, the people advancing that claim do not seem to commit to the view that such animals have conceptual understanding of goodness, and do not generally talk as if they do. And those who want to deny full moral standing to animals generally do not simply point out the lack of the conceptual resources necessary to form such beliefs as evidence of a lack of capacity for desires, which would be an obvious response if they understood that to be what was being argued. All this is strong evidence that the propositional attitude endorsement theory is not what moral theorists have in mind when thinking about desires.

iii. The Behavioural Theory of Desire

Let us now turn to the final prominent class of theory of desire, the behavioural disposition theory, which posits that desiring is having dispositions to take certain actions in response to certain stimuli. This is sometimes said to be the "standard" theory of desire. However, its popularity might be owing to the fact that a significant proportion of the aims of psychology and neuroscience, and by extension philosophy of mind and philosophy of psychology, involve explaining and predicting human

behaviour. Stipulating that desire means whatever causes behavioural dispositions might be useful for these purposes, but this does not translate to suitability for moral theory, which has completely different goals and concerns. This might be evidence that different concepts are at work in the different disciplines. Nevertheless, the behavioural theory of desire is the only theory that is explicitly addressed in moral philosophy to any significant degree. Michael Smith argues for it, as part of his explicitly non-welfarist agent-relative moral theory.³⁰ T. M. Scanlon brings it up specifically to say that it has no moral relevance.³¹ But there is nobody, to my knowledge, who advocates a desire-based criteria for well-being or moral standing and who explicitly endorses this view of desire. In fact, Michael Smith specifically claims that most moral theorists have, at least implicitly, a phenomenological concept of desire in mind,³² giving evidence that this is in fact the concept generally in use by moral theorists. It is a mistake to assume that prevalence in philosophy of mind means that the behavioural theory is “correct” in all contexts, rather than that “desire” might be polysemous.

Under the behavioural theory of desire, purely subconscious behavioural “tics” can count as desires, even if you are always unaware of them, would deny having them in all sincerity, and have no belief (or even actively disbelieve) that any of their effects or aims are good. Michael Smith provides an example that shares some of these features: a man who is subconsciously disposed to always buy his morning newspaper from a newsstand that has mirrors behind the counter, but has no awareness that he is doing so and would not be able to identify that he has this

³⁰ Michael Smith, “Deontological Moral Obligations and Non-Welfarist Agent-Relative Values”, *Ratio* 24(4) (2011).

³¹ T.M. Scanlon, *What We Owe to Each Other* (1998).

³² Michael Smith, “The Humean Theory of Motivation”, *Mind* 96(381) (1987).

disposition.³³ It is natural to say, as Smith does, that this is a case of a subconscious desire to look at himself in a mirror, but this is due to other implicit features of the case. One naturally imagines that the man has some phenomenological responses to this routine, even if he is not aware of the source of these feelings. When contemplating going to two different newsstands the thought of the one with the mirrors will feel better to him in some way, and actually seeing himself in the mirror will presumably make him feel some way that he would not have felt otherwise, thus making it natural to think that he desires to go to this particular newsstand and it would be wrong to prevent him from doing so. But if we imagine the case with these phenomenological features removed, leaving only the features the behavioural theory focuses on, one ceases to think of it as a case of desire. It is merely an unconscious habit, with no moral relevance.³⁴

Another case that is frequently discussed is Warren Quinn's "Radioman", a man who is disposed to turn on any radio near him, but does not feel attracted to turning on radios, does not derive enjoyment from turning on radios nor from having turned-on radios in his presence, does not feel that turning on radios is a good or worthwhile thing to do, and suffers no frustration or consternation when prevented from turning on radios.³⁵ Quinn uses this to claim that Radioman has no reason to turn on radios, which seems correct, but it is also intuitive to say that Radioman does not desire to turn on radios.³⁶ This is not a genuine desire to turn on radios, and helping or

³³ Michael Smith, "The Humean Theory of Motivation", *Mind* 96(381) (1987).

³⁴ Of course, it is still possible to resist this conclusion (see Michael Smith, "Four Objections to the Standard Story of Action (and Four Replies)", *Philosophical Issues* 22(1) [2012] for some potential strategies), and this is once again a debate that cannot be definitively resolved here. However, most of the available responses are only clearly compatible with moral theories vastly different from the ones we are considering (such as Smith's), and as such can be ignored for present purposes.

³⁵ Warren Quinn, "Putting Rationality in its Place", in *Morality and Action* (1993).

³⁶ See Declan Smithies and Jeremy Weiss, "Affective Experience, Desire, and Reasons for Action", *Analytic Philosophy* 60(1) (2019) for an argument to this effect.

hindering Radioman has no effect on his well-being. Treating cases like Mirrorman and Radioman as cases of desire makes sense in the aforementioned philosophy of mind context, where the goal is solely to explain and predict behaviour, but these are very unlikely to be used as examples of something whose satisfaction or frustration is relevant to well-being. This is not what moral theorists have in mind when they say desires are important, and excluding them is a virtue rather than a vice for a theory of desire in a moral philosophy context.

Furthermore, there are frequent cases where “desires” are posited for things we do not and cannot take any action to bring about. One such example is Derek Parfit’s previously-mentioned “stranger on a train” argument against unrestricted desire-fulfilment theories of well-being, where a person meets someone who has an illness and develops a desire that this stranger be cured.³⁷ Parfit argues that this is an obvious case of a desire whose fulfilment or frustration makes no contribution to the well-being of its possessor. But there is nothing the person in question will do, or can do, to attempt to bring about this state of affairs, so it seems not to be a “desire” in the behavioural-disposition sense. Nonetheless, I don’t think that Parfit would be at all receptive to the response on behalf of the simple desire-fulfilment theorist that this clearly wasn’t a true desire since it included no disposition to action, and was just some kind of “wish”. This has tended to be regarded as just as genuine a case of desire as any other.

It might be argued that there would be a disposition to action in that case, just one that is never actualized. Presumably the person in question would have acted to help

³⁷ Derek Parfit, *Reasons and Persons* (1984), p. 494.

the stranger get cured if there was anything she believed might be effective, there just happened to not be any such action available. But we can imagine other cases where it is even clearer that no disposition to action is possible – for instance, someone who has spent their life being diligently religiously observant, who desires that they were correct in their belief that God exist. That seems like something that would be accepted as a desire that would be relevant to moral outcomes, since the truth or falsity of that fact might indeed drastically change our estimate of how well that person's life has gone. But it is hard to see how it is even conceptually possible for there to be any dispositions to actions aimed at bringing that state of affairs about.

Even still, it might be argued that there is a disposition, but one that is impossible to actualize. It isn't clear what the basis would be for positing such a disposition, as there doesn't seem to be any evidence for it. In the absence of any such evidence, it would be ad-hoc to do so. But even if a disposition does exist, it seems strange to say that it is in virtue of that fact that a desire is had in that case, and that such a disposition to (impossible) action, alone, is constitutive of that desire. There might be thousands of other such dispositions that a person possesses, completely unbeknownst to everyone and impossible to detect. Surely there was some other reason we singled that particular thing out as intuitively having moral significance. It is obvious that a disposition to impossible action that had no other effects would be rejected as being a desire that contributes to well-being.

Another problem the behavioural disposition theory has is that it has difficulty accommodating the literature on akrasia. It is frequently claimed that even fully-informed people do things that go against their ultimate desires and are harmful to them. But under the disposition-to-action theory, the fact that you performed an

action would show that it was your strongest disposition and therefore your strongest desire. The phenomenological view might encounter similar concerns – if you were motivated to choose one thing over another, it was probably because you felt more strongly about it, and so you desired it more and thus it was better for you. But there is an easy response available: one could simply say that usually you feel differently, and will feel differently in the future, so it was bad for you. Recall that the phenomenological theory is not committed to describing a person's desires solely in terms of their occurrent sensations at a particular moment, but includes dispositions to feel certain ways. In this way, the phenomenological theory of desire can avoid deciding things one way or the other when it comes to the existence of akrasia. Nothing similar seems to be available to the behavioural theory. It would be very strange to say that what counts are dispositions to action in situations where no decisions about action are being made – indeed, it seems strange to even posit the existence of such a thing.

There is also a significant debate about paternalism, and whether it is permissible to interfere with people's actions for their own good. If someone is mistaken about what is good for her, some people believe that we ought to give her what is good for her regardless, while others think we should respect her wishes. This is not confined to unproblematic cases where the person in question is seriously misinformed – for example, stopping her from attempting to cross a bridge which she believes is stable but is unsafe, or stopping her from drinking what she believes to be tequila until you can inform her that it is gasoline. The debate takes place with respect to issues such as preventing people from making a decision to ride a motorcycle without a helmet or to engage in sex work even when they are fully informed about all the risks and consequences. But interfering with fully-informed people's actions for their own

good would be impossible if well-being involved desire-fulfilment and desires were just behavioural dispositions. The fact that they were going to perform that action would necessarily show that it was their strongest desire, and was therefore what was best for them. This is an implausible result. Even if it's impermissible to interfere with peoples' actions for their own good, it is usually understood to be possible to do so, which is why a moral dilemma and a debate about it exists. Of course, one natural thing to do would be to combine this theory of desire with a theory of welfare that is not based on desire-fulfilment, but such theories are not under consideration here. For the purposes of this project, where specific theories have been stipulated, it is clear that the behavioural-disposition account of desire is not suitable.

The behavioural theory cannot easily use the standard strategy of second-order desires to address these issues, either. Say I want to quit smoking, I think it would be better for me and better all things considered, but I can't resist the impulse to smoke cigarettes. It has been frequently claimed that the moral importance of my desire for tobacco is changed by the fact that I have a desire not to have that desire. But this second-order desire does not seem to be constituted by my having dispositions to attempt to change that desire. I might never do anything that actually aims to change that desire, and I might not even have the disposition to do anything to change that desire. This seems to happen frequently – stop-smoking aids might be readily available and offered to me and I refuse them, etc. – but we would still say I have the second-order desire, for reasons other than behavioural dispositions. In fact, it could equally be said on the dispositional view that I have a second-order desire to suppress *that* desire, since I have a disposition to ignore that desire and avoid thinking about actually making any efforts to quit. There is no clear way to establish priority of ordering for desires on the behavioural disposition account.

The behavioural disposition theory would also make a mess of the debate between Humeanism and Kantianism about motivation. To repeat, it's generally framed as Hume claiming that reason alone never overrides our desires, we always do what we desire, while Kant claims that we can and must do what duty demands even if it's not what we desire. If desiring something means having a disposition to bring it about, doing anything without desiring it would be impossible, and Kant's imploration to do things we do not desire if they are the right thing to do would be absurd. As for Hume, it's most commonly taken that he is making a substantive claim about human psychology, that we in fact always do what we desire, rather than stating the trivial tautology that if an action is performed, there must have been a behavioural disposition to do it. The behavioural disposition theory, then, does not respect the moral theory discourse about desire.

iv. Summation of the Advantages of the Phenomenological Theory

The phenomenological theory avoids all the major problems with the other theories. It can easily accommodate cases both of desiring something without believing it to be good and believing something to be good without desiring it. There is no conflict at all with something seeming to me to be a certain way but disbelieving it to be that way. Much as I can have a phenomenal experience of something as blue but believe that it isn't blue (owing to filters or coloured light or something), I can phenomenally experience something as good or rewarding while believing that it actually is bad. The opposite is equally true – I can accept the proposition that something is good while not feeling or experiencing its goodness. There is a tension in these positions, which is quite appropriate and correct, but it is clearly possible.

Furthermore, it seems plausible that many non-human animals can experience things as good or rewarding without being able to conceptualise things as good or rewarding or make judgments about goodness or reward, much as cats and dogs have the ability to experience things as blue even if they do not have a concept of “blueness” or the ability to judge things to be blue.

The phenomenological view also correctly makes the debate between Kantianism and Humeanism an empirical matter. It is certainly conceptually possible to ignore one’s feelings and do only what is rationally required. It is equally sensible to claim that in all actual cases of motivation, even when someone seemingly performed a selfless action, there were in fact desire-feelings at play. Whether Kant or Hume turns out to be correct, both positions are consistent and logical.

Akrasia is also possible under the phenomenological view, even if Humeanism is right and desires are the only things that can motivate. I might feel a certain desire more acutely at a given moment, leading me to act on the basis of it, while other conflicting desires I have might be stronger, in the sense that I will feel them even more acutely in the right circumstances. It also allows for second-order desires – I certainly can feel different ways about my desires when considering them, some seeming good to me and others seeming bad. This allows us to accommodate this distinction.

Some might be tempted to think that this has shown that more than one criterion must exist for a true desire to exist. This is acceptable for my purposes as long as a phenomenological component is one of the things required. So, for instance, one

might argue that a behavioural disposition along with the relevant feelings is critical, or that something must be both felt to be and believed to be good in order to be truly desired, or even that all three components must be present. My arguments will still apply to such positions.

Therefore, if Artificial Intelligences are to have moral standing on the basis of having the capacity for desires, they must have (a specific type of) phenomenal consciousness.

Chapter II: The Source of Evidence to Guide Moral Actions

1. The Problem of Other Minds

We have adopted a class of theories whereby the capacity for some phenomenal state of desire is the criterion for moral standing. However, if the relevant criterion for having moral standing is phenomenological in nature, it might be questioned whether we can ever have truly reliable practical guidance about *any* of our moral actions. We can only have awareness of our own phenomenal states, and can never be certain of what other beings are feeling. Even if we can identify correlations with our phenomenal states and with reports of the relevant phenomenal states from others, we have no direct access to those states, and thus can never have certain evidence that these states are present or absent.

This kind of sceptical concern has been discussed for a long time,³⁸ but does not tend to trouble people in their practical decision-making when it comes to their dealings with other humans (few people struggle with the decision of whether to adopt solipsism as the basis for their pragmatic ethical principles, for example). But though we are intuitively willing to ascribe consciousness to entities like other humans that seem sufficiently like ourselves, it is actually an incredibly difficult question how one can know that any entity other than one's self has phenomenal consciousness. David Chalmers' well-known "Zombie Argument" says that there could be an entity indistinguishable from a human but having no mental states.³⁹ Since it is conceivable that such a being could exist, we can never be certain of another being having

³⁸ At least since John Stuart Mill, *An Examination of Sir William Hamilton's Philosophy* (1865).

³⁹ David Chalmers, *The Conscious Mind* (1996), pp. 93-171.

phenomenological experience, so the argument goes. But even if this is logically possible, by and large people (quite correctly) do not take this type of sceptical argument very seriously when it comes to other humans. There are, however, significant differences in the situation when it comes to AI that lead people to demand more evidence.

One argument that could be relied on in the human case is that it is reasonable to assume that other humans have mental states like one's own in virtue of them having the same material structure.⁴⁰ Even non-human animals are often similar enough that it is reasonable to assume they have most of the same properties. An AI is much more different from us than any animal, or at least any animal most people would be tempted to say has moral standing. This has led philosophers like John Searle to doubt the existence of robot phenomenology.⁴¹ Given that this is an important practical concern, we cannot withhold judgement until we achieve the kind of logical certainty implied by the zombie argument, but need guidance to make decisions. If we are dealing with a machine that we believe might have moral standing, our only choices are to act as though it has moral standing, or act as though it does not; we will have to act in some way. And we may be facing robots that could plausibly have some chance to have moral standing sooner rather than later. In fact, without a lower bound on how sophisticated of a robot to take seriously as a potential subject of

⁴⁰ This idea goes back a long way as well – see Bertrand Russell, *Human Knowledge: Its Scope and Limits* (1948), p. 208–9 and 501–4

⁴¹ See John Searle, “Minds, Brains, and Programs”, *Behavioral and Brain Sciences* 3(3) (1980), and Daniel Dennett, reply by John Searle, “The Myth of the Computer: An Exchange”, *New York Review of Books* (June 24 1982). However, while Searle's arguments against the idea that any structure that instantiates the right program would be *guaranteed* to have consciousness seem sound, his claims that certain particular non-organic structures *could not* possess consciousness (such as collections of water pipes or beer cans) do not appear to be supported by anything beyond intuition in his writings. Searle's arguments will be discussed in greater detail later in this chapter.

moral concern, it is not even obvious that we do not *now* have machines that ought to be treated as having moral standing. By not adopting any precise or standard of what would count as good-enough evidence we would not actually be withholding judgement about the standards for saying a machine might have consciousness, but implicitly assuming by our actions that all current and near-future AI do not have moral standing, but in an unsystematic way without spelling out our justifications. We had better make sure that we have an articulable, defensible reason for doing so, and more importantly, be ready for advances in AI capabilities to make sure we do not change our approach too late or too early. We therefore need a practical standard for decision-making that allows for great uncertainty, yet still provides adequate guidance. This will allow us to move beyond the excessively restrictive standards of engaging with the zombie argument, while establishing what would be the standards for a useful test that could establish sufficient evidence for AI moral standing, which will be discussed in subsequent chapters.

2. Moral Epistemology and Decision-Making

The question is what to do in a situation where we are uncertain of unobservable moral properties of the situation, separate from any empirically observable consequences. This is somewhat different from how moral arguments usually proceed. Standard moral debates take place one of two ways. In the first type, people take the moral claims as more or less fixed and evaluate a set of materially different outcomes. Which entities have moral standing and which properties have moral value are taken as known quantities, and the discussion is about which course of action would produce the outcome that better realises those values, in virtue of things that are empirically verifiable. Examples of this would be things such as the

euthanasia debate, where a common type of consideration is something like whether legalising voluntary euthanasia might lead to some people being euthanised who were less than fully consenting, an outcome that is agreed to be bad by the participants in the debate.⁴² In this argument, what is at issue is how likely that outcome is, how prevalent it would be, and how that balances against some other value, like people having greater control and self-determination over their own lives.

In the other main type of ethical debate, cases are presented with their consequences stipulated, and our intuitions about their moral value are meant to inform which properties we are to put more moral importance on. Implausible scenarios are posited to demonstrate that certain ethical principles cannot be held universally and are flawed. For instance, in comparing utilitarian and deontological moral theory, one might concoct a case where a healthy person could be killed and their organs used to save five other people,⁴³ stipulating things like that the doctor's actions will never be discovered, that the transplants are guaranteed to be successful, that the person killed would not have gone on to help a greater number of people, and other things that could not actually be known for certain in a real-life situation. The point is to idealise the situation to determine which principles are more important without extraneous or irrelevant factors distracting from it.

The situation when it comes to potentially morally-significant AI is different from either of these. We have fixed the properties that have moral value for argument's

⁴² For an overview of these, see Craig Paterson, "A History of Ideas Concerning the Morality of Suicide, Assisted Suicide and Voluntary Euthanasia", in *Physician Assisted Euthanasia* (2009).

⁴³ See Philippa Foot, "The Problem of Abortion and the Doctrine of the Double Effect", *Oxford Review* 5 (1967).

sake – satisfaction and frustration of desires and aversions. The empirical physical properties of the outcomes under consideration are also more or less fixed – e.g., a machine made of X material programmed to do behaviour Y would have Z state inflicted upon it, etc. – so we are not debating unforeseen consequences. However, we are uncertain of unobservable properties that would significantly change a situation’s moral value, namely the phenomenal states, or lack thereof, in those machines. Knowing more about the physical situation cannot by itself answer this question, nor can our intuitions about any further moral principles or judgments. This is somewhat unexplored territory, but arguments from adjacent areas might be adapted. I will examine several different approaches to uncertain moral situations, adapt them for the AI situation, and establish their suitability.

One of the few approaches to uncertainty that has been prominently discussed in moral theory is the “maximin” approach. In John Rawls’ *A Theory of Justice*,⁴⁴ Rawls arrives at the conclusion that the rational principle to apply in choices where outcomes are unknown is maximin, maximization of the minimum, where it is ensured that the worst possible outcome is the least bad possible. This might lead one to believe that it would be best to treat any entity that could possibly have moral standing as though it does, “just to be safe”. However, it would be far from costless to take this approach. Treating any entity as having moral standing necessarily imposes a burden on others, and if a sufficiently large number of entities are granted moral consideration in this way, the burden might be significant. For example, humans might do a great deal of unnecessary labour to spare cleaning and maintenance robots from having to do unpleasant and burdensome tasks, which would be

⁴⁴ John Rawls, *A Theory of Justice* (1971).

completely wasted if such robots have no moral standing. People might even cause some amount injury to humans to avoid severe damage or destruction of robots, which would be a very bad decision if the robots had no moral importance. This plays into some of the criticisms of maximin that it assumes an extreme risk-aversion without any argument for why that risk-aversion is rational.⁴⁵ In fact, Rawls himself says that “maximin is not, in general, a suitable guide for choices under uncertainty”⁴⁶, but only in a case like his Original Position where the choice is irrevocable and determines all your future prospects, a situation we do not face when it comes to deciding how to treat AI. Furthermore, maximin is described by Rawls as the rational and correct principle when outcomes are not just uncertain, but where probabilities are *completely unknown*. This also is not the situation we are in. I am not sure what the precise probability is that my iPhone has phenomenal qualia, but I am willing to accept that it might be above zero and am quite happy to assert that it is less than one-in-a-million. This would make it irrational to include it as a possibility in a maximin scenario, since it is such a remote possibility making any significant tradeoff to avoid it would be a terrible decision. We must, at the very least, introduce a probability floor if we are to adopt any sort of maximin approach.

The standard approach decision-making under uncertainty is a utility calculus. It is intended to be used when fairly reliable probabilities and values can be assigned to outcomes. A utility calculus is sometimes used in medical ethics – say, for determining whether it is appropriate to prescribe a risky experimental drug. If there is a 10% chance of causing immediate death, assigned a utility value of negative-50,

⁴⁵ See John Harsanyi, “Can the Maximin Principle Serve as the Basis for Morality? A Critique of John Rawls’s Theory”, *American Political Science Review* 69 (1975).

⁴⁶ John Rawls, *A Theory of Justice* (1980) p. 153

and a 40% chance of curing whatever debilitating illness the patient suffers from, assigned a value of positive +20, we multiply the outcomes by probabilities and add them together and discover that there is a net +3 over the baseline of zero for not intervening, and find that it is correct to prescribe the drug. So, for instance, if we could somehow know that an AI's behaviour was only 25% likely to have been produced without desire, and we are considering ordering it to do some dangerous, difficult, or unpleasant task that has a disutility of -100, it would only be justified if it was saving a human from more than 75 units of harm, or over 15 humans from 5 units of harm each, *et cetera*.⁴⁷ The problem with this approach is that it is unlikely that we will be able to reliably assign probabilities (or indeed utilities) with that degree of precision in the foreseeable future. We can know that no behaviour would give absolute certainty that it was produced a mind having the right properties, since random chance could always produce a given behaviour. But beyond that it is difficult to say any more, since we do not have any data about such situations.

This kind of unreliability of probabilities has been recently discussed in ethical decision-theory to some extent. Rowe and Voorhoeve argue that it is normal and rational to avoid situations with this kind of uncertainty whenever possible.⁴⁸ The example given is a comparison between a treatment with a known 50/50 chance of curing an ailment or doing nothing, and a treatment that will either cure the ailment

⁴⁷ Though it is easiest and most intuitive to use this approach with utilitarian theories, where the values represent units of pleasure/happiness vs pain/suffering, other ways of measuring the goodness or badness of outcomes could also be quantified this way – for instance, a value of -100 might be assigned to failing to respect the rights of a sentient being, and -75 might be assigned to allowing people to suffer a certain harm when you could have prevented it at little to no cost. Using this as a decision-procedure heuristic would not entail accepting this “ideal utilitarianism” as the objective standard of ethical behaviour.

⁴⁸ Thomas Rowe & Alex Voorhoeve, “Egalitarianism under Severe Uncertainty”, *Philosophy and Public Affairs* 46(3) (2018).

or do nothing, but for which the probabilities of each potential outcome are completely unknown.⁴⁹ Most people intuitively prefer the first situation,⁵⁰ and this has been accepted by many (though not universally) to be rational. This also scales when we can assign probability ranges – if we can say that a treatment has a chance of curing the ailment of between 25% and 75%, this is exactly in between the certain 50/50 scenario and the completely uncertain scenario in terms of how good it is. Rowe and Voorhoeve suggest that it makes sense to be more cautious in these situations, and assign more weight to the potential bad outcomes, but claim that how much weight to assign to each outcome is to some extent “up to you”, and that tensions exist between several principles of rationality in these situations.⁵¹ The medical treatment example as written seems to be more applicable to cases of AI research – it weighs against putting ourselves in a situation where we are dealing with an AI where we have a high degree of unreliability in assigning probabilities to it having moral standing. To make it clear how it would apply to decision-making when we already have that kind of unreliability, consider the following case: if we provide the only treatment available, there is a chance between 0% and 25% that it will make the patient’s condition worse by some negative value (or within a range of negative values), and a chance between 25% and 75% that it will cure the condition, giving a positive increase to utility (if neither of these obtain it will do nothing). There are several different ways one could aggregate the different uncertain variables to come to an expected utility value, either adopting the suggested pessimistic risk-aversion

⁴⁹ p. 247

⁵⁰ See Stefan Trautmann and Gijs van de Kuilen, “Ambiguity Attitudes,” in *The Wiley Blackwell Handbook of Judgment and Decision Making*, ed. Gideon Keren and George Wu (Chichester: Wiley, 2015), pp. 89–116

⁵¹ p. 250

or not,⁵² and determine whether the action of administering the treatment has higher expected value over not intervening. Thus when it comes to AI, a test that gives a reasonable probability ceiling would be enough for action-guidance.

However, it is far from uncontroversial that risk-aversion is rational in all situations. It might depend heavily on what outcomes are at stake. An established approach that realises the value of caution in certain situations without unduly affecting things in other situations is what is known as the “precautionary principle”, which essentially says that we only should choose a course of action (when one is available) where the worst possible outcome is not an unacceptable disaster. If the worst possible outcome is not fully disastrous, the precautionary principle does not apply, but if there is any chance of something sufficiently bad happening, we should act to avoid that outcome even if the probability of it is low or unknown. This is an improvement over extremely risk-averse principles like maximin and arguably over more moderately but universally risk-averse principles like those of Rowe and Voorhoeve, since it allows for a rule where we would not be barred from everyday things like inconveniencing our mobile phones or our robot vacuums, but only from things that would cause more extreme potential harms like destroying large numbers of AIs that potentially have moral standing.

The problem with this form of the precautionary principle is what to do about incredibly unlikely scenarios with very bad outcomes. We would have to avoid

⁵² See Kenneth Arrow and Leonard Hurwicz, “An Optimality Criterion for Decision Making under Ignorance,” in *Uncertainty and Expectations in Economics* (1972) for one example of a principle that is amenable to this.

certain courses of action even if the chances of them occurring were a billion to one against, where under even a very risk-averse version of the decision-theory model multiplying by such a low probability could render this possibility irrelevant. However, that paradigm is not immune to this type of concern either – imagine that the extremely unlikely worst-case scenario involves accidentally creating a race of sentient self-replicating AIs that would produce endless generations of AI that exist in constant suffering, thus producing a near-infinite amount of disutility. Even multiplying this by an extremely low probability could still produce an unacceptably high number, allowing an extremely unlikely scenario to unduly influence our decisions. For this reason, proponents of the precautionary principle have made allowances for excluding things that fall in a sufficiently low range of probability, like the aforementioned mobile phone case, in accordance with the principle of *de minimis*.⁵³ For a risk to be *de minimis* means that the chance of it occurring is so small that it can be disregarded, and that actions carrying only a *de minimis* risk are “virtually safe”. However, there is still a difficulty for our purposes, since determining which things do and do not fall under the category of *de minimis* will require us to be able to assign probabilities, which (as I have argued) we will not always be able to do. There are clear cases, like the probability that a mobile phone has any phenomenal qualia, where even though we cannot assign a precise probability we can be sure that the chance falls below the reasonable threshold. However, in other cases it will not be so obvious whether the possibilities are small enough to ignore. J. Adam Carter and Martin Peterson have argued that we would need to take into account a “second-order” probability, the probability that our first probability is mistaken.⁵⁴ As an

⁵³ See Martin Peterson, “What is a *de minimis* risk?” *Risk Management* 4 (2002).

⁵⁴ J. Adam Carter and Martin Peterson, “On the Epistemology of the Precautionary Principle”, *Erkenntnis* 80(1) (2015), pp. 1-13.

example, “if the probability is .01 that the probability is one in a billion that substance XYZ will cause cancer, it would be ridiculous to claim that XYZ is de minimis”, because “the actual first-order probability could be much higher.”⁵⁵ However, they raise the concern that it is difficult to determine how these probabilities interact in a way that is not ad-hoc. If we multiply the probabilities of danger, we get a lower probability this way than considering only the first-order probability, which is obviously wrong. If we invert the probability, and calculate the probability that a course of action is safe, “we get the result we want, but we seem to have no good epistemic reason that explains why this way of carrying out the calculation is correct and the other way incorrect”, according to Carter and Peterson.⁵⁶ However, it is not as though the other approach, of calculating probability of danger, is natural or default and this is an impermissible ad-hoc amendment. We need to choose one or the other if we are to have pragmatic action-guidance, and even if there is no reason to prefer one over the other apart from giving us the results that seem intuitive, that is sufficient to prefer calculating the probability of safety. At any rate, Asbjørn Steglich-Petersen claims that there are perfectly good principled reasons to choose this way of doing it over the alternatives.^{57,58} Nonetheless, even if there is some question about how principled this approach is, it is better than having no system whatsoever. There does remain a concern about how confident we can be in that second-order probability assessment, but if it is at least somewhat reliable, it will suffice for practical purposes.

⁵⁵ Ibid., p. 10

⁵⁶ Ibid., p. 11

⁵⁷ Asbjørn Steglich-Petersen, “The Epistemology of the Precautionary Principle: Two Puzzles Resolved”, *Erkenntnis* 80(5) (2015), pp. 1013-1021.

⁵⁸ Carter and Peterson remained unconvinced by these arguments (see J. Adam Carter and Martin Peterson, “On the Epistemology of the Precautionary Principle: Reply to Steglich-Petersen”, *Erkenntnis* 81(2) (2016) pp. 297-304), but I remained unconvinced by their rejoinder – however, continuing this debate any further falls outside the scope of this project.

Thus we see what we need out of a test of AI – to make sure that we can establish that what would be sufficiently bad outcomes are sufficiently unlikely, by establishing an upper bound for a probability range. Therefore, what we need is a minimum standard of exclusion, a test such that, if it is not passed, it is acceptable to ignore that entity in our moral considerations, while allowing in as few things as possible. It is conceivable that modern computers have the properties required for moral standing, but it is unlikely enough that it is acceptable, and indeed preferable (sometimes even morally required), to treat them as if they do not. Thus, tests that modern computers can already pass are not suitable. Tests that are too demanding will similarly be unsuitable, leaving too great a chance that we could be doing something morally unacceptable.

3. Criteria For an Adequate Test of AI

We must find a sort of fast, practical, easy-to-conduct test would allow us to have evidence that capacities relevant for moral standing are present. The standard test of AI that immediately comes to mind, the Turing Test, is entirely inadequate for these purposes. The Turing Test, originally developed by Alan Turing, involves having a text-based conversation with an AI and seeing whether it can be reliably distinguished from an actual human by a judge that is blind to all features of their interlocutors apart from the contents of the text messages produced.⁵⁹ This test is at once too demanding and too easy. It is too demanding, because it is entirely possible to have the properties necessary for moral standing without being able to pass. Many entities that clearly have moral standing, such as very young children and people

⁵⁹ See Alan Turing, "Computing Machinery and Intelligence", *Mind* 59(236) (1950).

with disorders such as severe aphasia or schizophrenia, will frequently or always fail the Turing Test owing to lack of sufficient linguistic competence. This is also the case for non-human animals – none of them possess any significant degree of linguistic ability, but most people would accept that at least some animals have some degree of moral standing, including of (at least of some of) elephants, dolphins, rats, cats, dogs, pigs, chimpanzees, etc.

The ability to pass the Turing Test cannot even be a sufficient condition, since many simplistic chatbots, which clearly do not have the relevant properties, have been reported as having passed. As early as 1966, the program ELIZA was convincing people that it was a real person, despite being incredibly simplistic. The ELIZA program purported to be a Rogerian psychotherapist, and convinced people by relying on broad stereotypes of Rogerian psychotherapy, essentially sticking to variants of “why do you think that is?”, “how does that make you feel?”, and “tell me about your mother”.⁶⁰ The expectations about Rogerian psychotherapy provide an explanation for what would otherwise be a violation of conversational norms. Other chatbots have passed in a similar way – the Eugene Goostman program was proclaimed as the first program to have passed the Turing Test due to having convinced people that it was a 13-year-old child who did not speak English as a first language. Once again, it achieved this through simplistic tricks that are not indicative of any of the capacities that would confer moral standing.⁶¹ The linguistic ability that

⁶⁰ Huma Shah, Kevin Warwick, Jordi Vallverdú, and Defeng Wu, "Can machines talk? Comparison of Eliza with modern dialogue systems" (2016).

⁶¹ Celeste Biever, "No Skynet: Turing test 'success' isn't all it seems", *The New Scientist* (9 June 2014).

would allow an entity to pass the original Turing Test has no real correlation with moral standing.

It might be argued that the reason relatively unsophisticated programs have passed the Turing Test is that the tests have not been conducted sufficiently rigorously. Many of the tests had a restricted domain of discussion, and generally were conducted over a period of as little as five minutes.⁶² It is certainly reasonable to insist that the spirit of the Turing Test would allow it to last much longer than that brief, arbitrary limit, and allow discussion of any topic.⁶³ However, it is not clear how much this change would help. The test cannot be allowed to go on indefinitely, not only for practicality reasons, but also because even humans cannot perform competently in a Turing Test forever. How long until a human becomes too bored, tired, or impatient to continue to produce novel sentences? Not limiting the field of discourse is certainly an improvement, but if the replacement is simply selecting fields of discussion at random, there is no guarantee that anything more difficult for these programs will come up. People are generally pretty bad at intuitively predicting which tasks computers will have a problem with, and as such we cannot rely on simply allowing free choice of topics. Rather, if there is some reason particular topics will be more likely to require morally-relevant capacities to discuss competently, we should figure out which those are and why, and focus on testing those specifically.

⁶² Adam Mann, "That Computer Actually Got an F on the Turing Test", *Wired* (9 June 2014).

⁶³ Turing's prediction of a machine passing the Turing Test by the year 2000 did explicitly use a length of just five minutes as the criterion, describing the situation he was referring to as "an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning." (Alan Turing, "Computing Machinery and Intelligence", *Mind* 59[236] [1950], p. 442). However, he later said that a real pass of a test "with no questions barred" would take "at least 100 years" ("Can Automatic Calculating Machines Be Said To Think?", *BBC Third Programme* [14 January 1952]). It seems likely that Turing did not have fixed views of exactly which parameters for the test would be appropriate under all circumstances.

Some additional restrictions on the Turing Test have been proposed to deal with these concerns. One example is the so-called “Lovelace Test”,⁶⁴ which requires not only success in the test, but also that the success cannot be explained by what was put into the program. In effect, the machine cannot have been deliberately and purposefully designed to accomplish a task, but has to be able to do something where the programmer cannot explain exactly how this was achieved. This might be thought to have been achieved to some degree by the “black box” nature of machine learning procedures, where the processes are somewhat opaque and cannot easily be examined.⁶⁵ This criterion bears some similarity to Daniel Dennett’s “intentional stance”,⁶⁶ wherein we are required to treat entities as having certain mental states when that is the only way we have to explain and predict their behaviour, but if we have a full mechanical explanation of a certain behaviour, the intentional stance can be discarded to allow for greater precision and accuracy. In fact, the intentional stance theory has been expanded by other thinkers to include phenomenal states, making for a “phenomenal stance”.⁶⁷ A theory like this would certainly exclude the present-day chatbots, as they are not terribly complex and it is fairly easy to understand and explain how they achieve the success they have from a structural perspective. However, this amended sort of test is no good for our purposes either. Part of the appeal of Turing Test-style measures is not requiring that we look at the underlying structure of subjects, but allowing reliance on easily observed external behaviours. The same behaviour should always consistently constitute either a “pass”

⁶⁴ See Selmer Bringsjord, Paul Bello, and David Ferrucci, “Creativity, the Turing Test, and the (Better) Lovelace Test”, *Minds and Machines* 11(1) (February 2001).

⁶⁵ *The Lancet*, “Opening the black box of machine learning” (2018).

⁶⁶ Daniel Dennett, *The Intentional Stance* (1987).

⁶⁷ Philip Robbins and Anthony I. Jack, “The Phenomenal Stance”, *Philosophical Studies* 127(1) (2006).

or a “fail”, regardless of the entity performing it. Even more importantly, however, this requirement would risk us either applying an arbitrary double standard, or getting completely unacceptable moral results. Our understanding of the workings of the human brain continues to advance constantly, and it is not implausible to think that we will someday have a complete explanation of all or most human behaviour in causal-mechanical terms. But the mere existence of such a possible explanation would not mean that the explanation involving phenomenal states was illicit or incorrect, and to accept that would be to say that at that point we could not say that humans have moral standing on that basis. Rather, we need a test where the only reasonable explanation for passing is similar enough to the explanation for how humans pass, where the explanation in the human case is related to the relevant desire-phenomenology.

Could there be a better form of the Turing Test, where the behaviour it tests for would by itself be adequate evidence of moral standing? There are a number of variations on the Turing Test that have been proposed that intuitively seem plausible as a measure of moral relevance, but they generally come with no principled explanation of why accepting them as such is justified. If some of these do in fact track the capacity for phenomenal desire, then my arguments will have given them a supporting explanation, and the existence and advocacy of these tests as an intuitive measure will provide further evidence that the properties I have identified are the correct ones, and are in accord with common-sense morality.

4. Answering Some Doubts for Turing-Style Behavioural Tests

There remains a question of whether any test of external behaviour can ever truly give any evidence of the possession of internal states. As previously mentioned, the zombie argument is not a persuasive challenge to this. For one thing, concerns have been raised about Chalmers' argument, such as over whether zombies are in fact properly conceivable, and over whether conceivability actually entails possibility.⁶⁸ Even if I can, in some sense, entertain the proposition that a philosophical zombie exists, it might still be metaphysically impossible, much as I can entertain the proposition "what if water were not H₂O", even though that is still an impossibility. However, even if Chalmers' argument is right, it is at best a mere logical possibility. There is no evidence for it being in any way likely in the actual world. Furthermore, acting on the basis of that possibility, even if it were likely enough to justify doing so, would require us to doubt the phenomenological capacities of other humans just as much as any other entity, and so would ultimately change nothing. Solipsism is no serious alternative to any other reasonable plan of practical action-guidance. As there is no evidence whatsoever for its actual likelihood, and treating it as a possibility would provide no practical guidance, the possibility of zombies is irrelevant to my argument.

Far more problematic is the case where an entity is not empirically indistinguishable from humans, but instead is observed to have a vastly different neural structure to humans. Such an entity could in theory behave in all the same ways as humans such that it cannot be differentiated from a normal human without looking at that neural

⁶⁸ See Moti Mizrahi & David R. Morrow, "Does Conceivability Entail Metaphysical Possibility?", *Ratio* 28(1) (2015).

structure, and yet have no mental states. Ned Block has presented such an argument, a hypothetical program modelling his Aunt Bertha.⁶⁹ The “Aunt Bertha” program has a massive table of appropriate responses to any English sentence, and when given a sentence as an input, it gives an appropriate response exactly as Ned Block’s Aunt Bertha would give. If we take it that this entity has no phenomenal states, we have a problem. The existence of such a program is a logical possibility as well, and more likely to be possible than Chalmers’ original zombie case, and this is a case that might make a relevant difference to how we ought to act, since it differentiates AI from humans by introducing an uncertainty that applies only to the former. Taking it seriously does not throw all of our decision processes into doubt like accepting solipsism would, it would simply introduce uncertainty in this specific area.

The reason Block’s case is a concern is that the structure he describes seems intuitively like it could not possess phenomenal consciousness or any other morally significant mental states. It is a structure similar to programs that can be run on our personal computers and smartphones. It undermines the notion that we can distinguish which entities have moral importance based on their behaviour, and removes any means we have of knowing where to draw the line between which structures are complex enough, or similar enough to human brains, to have the relevant mental properties. To alleviate this concern, it will be sufficient to establish that not just any old structure will be sufficient to produce the kind of behaviour associated with the mental states we are concerned with. As long as these kinds of behaviours can not be produced by the kind of structures where it is excessively implausible that could produce phenomenal consciousness, like Block’s Aunt Bertha

⁶⁹ Ned Block, “Psychologism and Behaviorism”, *The Philosophical Review* 90 (1) (1981), pp. 5–43.

table lookup program, we will have sufficient certainty for action-guidance. We will still not know with absolute certainty whether any given artificial intelligence has any given phenomenological state, but we will have a principle that is adequate for practical purposes. I will present several arguments that imply that in fact any structure that is likely to produce behaviour associated with the relevant phenomenal states will not be so simplistic as to obviously lack phenomenal consciousness.

It has been claimed that observed cases of multiple realisability of cognitive states undermines the idea that there is any distinctive structure that correlates with whatever mental states we think are relevant. This is only a serious problem if there are actual observed cases of multiple realisability to give support to this, since if it is a mere possibility it ought to play no role in our practical deliberations, much like the mere possibility of zombies or the massive table-lookup program. It has been alleged that there are such cases of multiple realisability, both within and across species, but William Bechtel and Jennifer Mundale have argued that widespread multiple realisability is not nearly as plausible as has been popularly assumed.⁷⁰ If general multiple realisability were true, they argue, it would lead to global scepticism about everything neuroscience tells us. But the successes of neuroscience tell us that is unlikely to be the case. According to them, “in PET especially, but also in MRI, the signal to noise ratio is relatively low, requiring averaging across subjects. Averaging will cancel out individual differences; thus, the fact that any results at all survive averaging as well as transformation onto a common brain plan indicates a great deal of commonality”,⁷¹ which dismisses many of the claims about intra-species multiple

⁷⁰ William Bechtel and Jennifer Mundale, “Multiple Realizability Revisited: Linking Cognitive and Neural States”, *Philosophy of Science* 66(2) (June 1999).

⁷¹ *Ibid.*, p. 190

realisability. They also explain that “most neuroimaging to date is performed on humans while the most detailed neuroanatomical and neurophysiological work (using, e.g., single-cell recording) has been done on other species. As a result, researchers often have to try to coordinate the imaging work on humans with neuroanatomy from other primate species (especially the macaque), and thus are assuming that cognitive functions are not differently realized in the two species.”⁷² Commonalities have also been observed in species more distant from humans, such as dogs and rats. The success of this type of approach in neuroscience have given us reason to doubt multiple realisability across species. Though there are some differences between different species in brain structures, there are far more similarities when the same states are observed, and this would be unlikely unless there were significant constraints on which structures can realise the right sort of property. If this is true, it is very unlikely that we will find entities with massively different neural structures that will be able to perform the same kind of tasks.

Stuart Shieber has demonstrated that a program of the kind Ned Block describes is in fact a nomological impossibility.⁷³ It is logically possible for a table to include all likely inputs and all reasonable responses, but “adding the further constraint of mere physical existence in the current universe is sufficient to provide a strict limit on the storage capacity of the machine and hence how long a Turing Test it could pass”,⁷⁴ a figure which, giving the most generous possible inputs, Shieber calculates to be about 37 seconds. In other words, the entire storage capacity of the universe could only

⁷² Ibid., p. 190

⁷³ Stuart M. Shieber, “There Can Be No Turing-Test Passing Memorizing Machines”, *Philosopher’s Imprint* vol. 14 no. 16 (June 2014).

⁷⁴ Ibid., p. 7.

store enough conversation strings to be sure of communicating appropriately through 37 seconds of conversation. Thus a program exactly like what Block describes is of no real concern for practical decision-making.

It is of course possible to optimise the program with programming tricks and shortcuts in such a way as to significantly improve its performance and extend the amount of time it can emulate human behaviour – as proved by the fact that programs of a similar kind have succeeded in passing the Turing Test for several minutes. It might be thought that we can simply optimise such programs further until they can pass for much longer, and they will still essentially be simple table-lookup programs with no claim to having phenomenal states. We will then again be in a situation where we have no evidence of which entities have the right mental states based on their behaviour. However, Drew McDermott⁷⁵ has argued that cognitive science implies that the structure of any program adequate to the task of emulating human behaviour indefinitely is likely to just be an optimized model of the human brain, and that it would therefore be illogical to deny that it has the same mental states as a human brain. The table-lookup-style program would need to have so many complex rules and shortcuts to fit in a reasonable storage space that its structure would necessarily become fairly complex. If someone were to somehow make such a program, and we examined the structure in detail, it is quite likely that we would find that the structure would be equivalent to an optimised model of the brain. In fact, a large degree of optimisation might not be possible, and the basic structure we produce might simply be pretty much that of (at least some part of) the

⁷⁵ Drew McDermott, “On the Claim that a Table-Lookup Program Could Pass the Turing Test”, *Minds and Machines* 24(2) (2014).

human brain, which would make it very plausible that it has the same mental states and properties as a human brain. As McDermott says, “a seemingly unintelligent program might turn out to be an optimized version of an intelligent program, with the same intelligence”.⁷⁶ To insist otherwise would essentially require us to dismiss all of cognitive science. If we “suppose it is impossible to model the human brain computationally, in the sense that no matter how fine-grained one’s model of the brain’s neurons, synapses, vesicles, glands, glial cells, etc., something crucial is always left out [...] then it’s difficult to see how scientific psychology is even possible.”⁷⁷ However, McDermott explicitly exempts phenomenological properties from this argument, to leave the possibility of some kind of dualism as an option.⁷⁸ It is not clear, however, why it is wrong to deny “intelligence” to the optimized brain if it is acceptable to deny it “experience”. If matching the computational and functional profile of the brain can still leave ingredients out, why couldn’t intelligence be one? Rather, it’s more plausible that all the ingredients are necessary to match the functional profile. At any rate, there is no evidence to deny that all the properties of the brain must emerge from a perfect match of its structural and computational profile, and as such we would be required to treat such a reproduction as equal to a human.

If we follow arguments about structural constraints further, we get to the extreme physicalist position of John Searle.⁷⁹ Searle argued that the exact material makeup of the brain (and perhaps parts or all of the body) might be necessary to get the right

⁷⁶ Ibid., p. 144.

⁷⁷ Ibid., p. 162.

⁷⁸ Ibid., p. 163.

⁷⁹ John Searle, “Minds, Brains, and Programs”, *Behavioural and Brain Sciences* 3 (1980), pp. 417-457.

mental states, and even an exact duplicate of the structure would not be enough to guarantee the same mental states if it were composed of a different material. It is very implausible to claim that a match this direct would be necessary to achieve the correct functions and behaviours, so if the criteria is this precise, function and behaviour will again not be a reliable guide. Searle has an argument with similarities to the zombie argument and the Aunt Bertha argument, the “Chinese Room” argument. In this case, it seems as though a computer is able to converse perfectly in written Chinese to an interlocutor, but in fact the “computer” is just a monolingual English speaker in a room that contains books of Chinese symbols with instructions on which to output in response to inputs. Searle argues that no matter how convincing the responses seem, neither the room, nor the person, nor the system of the room and the person, understands Chinese. Searle also insists that a structure matching the human brain would not be sufficient to overcome this objection:

To see this, imagine that instead of a mono-lingual man in a room shuffling symbols we have the man operate an elaborate set of water pipes with valves connecting them. When the man receives the Chinese symbols, he looks up in the program, written in English, which valves he has to turn on and off. Each water connection corresponds to a synapse in the Chinese brain, and the whole system is rigged up so that after doing all the right firings, that is after turning on all the right faucets, the Chinese answers pop out at the output end of the series of pipes.⁸⁰

This system still, according to Searle, could not have any understanding. If this were true, there would be no point in discussing the potential moral standing of AI, since such a thing would be impossible. A different substance would never produce the

⁸⁰ Ibid., p. 421

same states, Searle says, any more than one could run a computer model of lactation and still get real milk. Nothing short of building an actual organic brain would ever be sufficient. However, I do not believe we should take our intuitions about such a case as reliable, since it is so far removed from our experience and from reality. A water pipe system of the type Searle describes, in order to actually match the structure of the brain's neurons and synapses, would probably need to be about the size of a planet, given the minimum sizes based on the properties of water and pipes compared to the substances of the brain. At that scale, information would have to travel much faster than water can to achieve the same speed of response – and obviously much faster than the electrical impulses in the actual brain travel, since the distances involved are much greater. This would basically necessitate the water travelling faster than the speed of light and violating several immutable laws of physics. It is not clear we can imagine such a thing coherently, and certainly unlikely that we are imagining it accurately, so there is little reason to trust our intuitions based on this imagining. Searle also claims that “in principle the man can internalize the formal structure of the water pipes and do all the "neuron firings" in his imagination.”⁸¹ However, it is obviously beyond the cognitive and information-storage capacity of the human brain to do this, as no system can run a simulation of itself. This is at best a mere logical possibility, possible “in principle” but not in practice, and indeed might not even be possible in principle. It should play no part in our deliberations. There is no real evidence of the correctness of Searle's strict criteria, and if we were faced with an entity that had a brain with a structure exactly like that of a human but made of silicon and metal rather than carbon and organic acids, and that acted exactly like a human, it would be excessively morally risky to

⁸¹ Ibid.

not treat it as having moral standing merely on the basis of the claim that it might need to be made of organic materials to have the capacity for phenomenal consciousness.

A further positive argument that the right behavioural tests will provide true evidence for the existence of phenomenal qualia comes from Karl Popper and John Eccles,⁸² who have argued that qualia of the kind associated with pain and desires must be essential for our behaviour and continued success, since otherwise they would not have evolved or would have since been discarded as evolutionarily non-advantageous. They present experimental evidence that consciousness does have advantageous functional roles, and conclude that “the self-conscious mind exercises a superior interpretive and controlling role upon the neural events” that is not replaceable.⁸³ One might make an objection against this argument, that the qualia might just be necessarily caused by the same thing that causes the behaviour, rather than being the cause of this behaviour. However, this is a very minor distinction and is irrelevant for my purposes. The behaviour would remain just as reliable a guide to the existence of the qualia. In this type of epiphenomenalism there is still no contingency to their relative place in the causal chain, it is just that it is a branch rather than a link. Evolution would have found a more efficient system, if one existed. If this is true, it is unlikely that we will find entities with the same capabilities that lack the relevant qualia.

⁸² Karl Popper and John Eccles, *The Self and Its Brain* (1977).

⁸³ *Ibid.*, p. 356.

David Lewis discusses the case of “mad pain” as a challenge to strict functionalism – a person might feel the same sensation that humans usually feel in cases of pain, but it might cause completely different behaviour, not leading to any attempt to stop or avoid it, and it might also be caused by completely different things.⁸⁴ This seems possible, though we would have no way of knowing if that were occurring in reality based solely on observation of behaviour. Conversely, Lewis proposes the case of “Martian Pain”, which is a state that serves the exact same causal role as pain, but is instantiated by a completely different physical structure in the Martian race – it involves the hydraulic inflation of chambers in their feet.⁸⁵ Lewis states that both elements have claim to being what determines phenomenal content. This would make it impossible to just rely on looking at neural structure either. According to Lewis, however, if a state is phenomenological in nature (like pain, and as I have argued, like desire), the phenomenal character of such a state is inextricably linked to the properties that make it such a state, and it is impossible to consider the latter without the former. “Only if you believe on independent grounds that considerations of causal role and physical realization have no bearing on whether a state is pain should you say that they have no bearing on how that state feels”.⁸⁶ The Martian pain state, in virtue of playing the same causal role for their species as human pain, must feel painful. The mad pain state, in virtue of having the same physical realisation as typical pain in humans (just with different causal connections), surely feels painful as well. This also provides an answer for potential problems with neurologically atypical humans – a human would feel desire in virtue of having similar, though slightly defective, brain structures as normal humans, such that they do not fulfill the causal

⁸⁴ David Lewis, “Mad Pain and Martian Pain”, In Ned Block (ed.), *Readings in the Philosophy of Psychology* (1980), p. 217

⁸⁵ *Ibid.*

⁸⁶ *Ibid.*, p. 222.

role correctly, bearing some similarity to mad pain, being a sort of “mad desire”. Its causal history and structure is just as normal until something goes wrong at the behavioural output phase. On the other hand, an AI that has a very different structure from humans but that reliably produces the correct behaviour associated with desire could be considered as likely to have actual desire, just a sort of “martian desire”, similar to how the martian has exactly the right causal connections instantiated by a different structure. Whereas if a non-humanlike structure produced behaviour similar to, say, a psychopath (this example will be explored in detail in later chapters), it is likely to have no desire and therefore no moral standing, which we can say without claiming that humans with such conditions have no moral standing.

In this way, we could have a very reliable empirical way of determining whether Artificial Intelligences feel pain or desire or other such states – if their behaviour can robustly match the causal connections that exist in entities that we are pretty sure have the correct phenomenal states (and that is, *very* robustly),⁸⁷ then we know that this particular AI architecture produces pain in this population, even if individual members are programmed to react abnormally to it. This avoids relying solely on individual functionality, since it might be possible to program some sort of override such that every part of the process involves phenomenal states in normal ways until at the very moment of action, the action is replaced by one consistent with the opposite phenomenal state or with no phenomenal state.

⁸⁷ We must avoid unsophisticated cases of emulating the causal profile, since behaviours such as reacting and complaining when treated roughly can be (and have already been) added to things as simplistic as children’s toys.

However, there remains a concern for someone who believes they do have precisely the kind of independent grounds for disregarding considerations of causal role and physical realization that Lewis refers to. It is possible to (at least believe oneself to) have such grounds in the case of pain, and is perhaps even more likely in the case of phenomenological desire. The argument would proceed more or less as follows: the causal roles are too loosely-defined, and the physical realisations too disjunctive and dissimilar (both within and across species), for there to be a sufficiently well-defined concept to make such claims, and the state in question (pain, desire, whatever) is either just the sensation (or a disjunction of sensations), or a fiction. The concern is the possibility of an entity that has the relevant qualia, but that play a different causal role, not due to a deviation from a baseline of normality but as a matter of course. Lewis claims that this is conceptually impossible,⁸⁸ but his claim might not be entirely satisfying. But for my purposes, I need not establish that it is conceptually impossible. Its practical possibility is what is relevant.

⁸⁸ Ibid., p. 221.

Chapter III: The Dopaminergic System, VTA/SNpC, and Reward

1. On the Individuation of Phenomenal States

We have seen that the morally-important concept of desire is best understood as involving phenomenal states. I have established that we need a practical, behavioural test that will give us a reasonable estimate of the probability that an entity possesses the capacity for desire phenomenology. What I will now turn to is determining which brain states in humans correspond to the phenomenal states associated with desire that moral theorists care about, and determine which functional roles they play. As has been alluded to earlier, the use of the term “desire” in the cognitive science literature will provide little guidance, since this term is generally defined in a purely behavioural-dispositional way in that field. It is unlikely that all and only things referred to as “desires” in cognitive science will be the kind of desires of interest for present purposes. I will therefore analyse and interpret the terms used by moral theorists to see which plausibly count as cases of desire phenomenology.

How can we determine when a relevant desire is present? There are two *prima facie* intuitively plausible answers to this question that are not in fact viable. The first is to demand a direct neural correlate of our concept of desire, such that a single brain state can be isolated as involved in all cases we would term “desire” – and to claim that in the absence of one, the positing of desire is unscientific and unjustified. But there is obviously a great deal of heterogeneity in the phenomenological sensations of desire; we feel the desire for ice cream, the desire for world peace, and the desire not to be punched in the stomach in very different ways. Given that, it is natural to think that we may not find a single brain region responsible for all the things we think of as desire, and that a one-to-one mapping will be impossible.

However, this should not be construed to show that desire phenomenology of this kind is merely an arbitrary and ad-hoc construct. I will not attempt to argue against full-scale eliminative materialism of the kind that denies the very existence of conscious states, but there are more moderate positions that question the degree to which (at least certain elements of) our folk-psychological vocabulary correspond to anything appropriately scientific.⁸⁹ This leads to the other possibility, of claiming that looking for any neural correlates at all is mistaken, and that we should understand these terms as referring only to a social category. While desire is indeed a socially-relevant category, it is more than that, and is not an arbitrary construct of our particular society that could have been otherwise, even if it does not correspond exactly to one particular brain state or mechanism. Zenon Pylyshyn points out that a similar state of affairs likely obtains with things as uncontroversial as the ability to interpret pronouns.⁹⁰ We know that this capacity exists and is possessed by humans in general (even hypothetical ones who do not presently speak a language that uses pronouns), and is not something we have arbitrarily created in a particular cultural context. The fact that the exact same mechanism is not at work in every instance of pronoun use or understanding does not show that it is illicit to refer to our ability to interpret pronouns, or that linguistic ability provides no explanation of our

⁸⁹ See for example Karl Friston, Jérémie Mattout and James Kilner “Action understanding and active inference”, *Biological Cybernetics* 104(1-2) (2011), which claims that seeing the brain as performing “active inference” does away with the idea of desires and shows them to be nothing more than action-oriented perceptions, not relevantly distinct from beliefs. However, one interpretation of this claim is merely that “Desire [has been] relegated to a phenomenal sensation associated with this sequence of events, and does not seem to play any causal or functional role” (Joe Dewhurst, “Folk Psychology and the Bayesian Brain”, in *Philosophy and Predictive Processing* [2017]), which would give desire an epiphenomenalist position that is compatible with my arguments. The question then becomes which particular behaviours are accompanied by sensations that fit the criteria and which do not. The issue is with the claim that we ought to lump together whatever we happen to have shown correspond to a particular state and assume that something genuinely relevant cannot cut across these categories.

⁹⁰ Zenon Pylyshyn, “What’s in a Mind?”, *Synthese* 70 (1987).

behaviours or capacities. We need not find a single neural correlate for how we decide what pronouns refer to.

There are at least two reasons that the appropriate explanation is one given in terms of beliefs, in spite of the fact that there can be no doubt that the behavior is caused by biological processes of some kind on each occasion. One reason is that it certainly need not be the case that the same neural or biochemical process is operative on every occasion where a pronoun is assigned a referent. This is just to repeat what I have already said before, namely that the category of behaviour we are interested in may cross classify behavior described using biological terminology, so that a psychological generalization may collapse across [a] disjunction of biologically distinct events.⁹¹

The same might be said to apply to phenomenological desires. A number of distinct categories of mental events might make up the category we are interested in (i.e., morally relevant phenomenological desires), which is nonetheless real enough to constitute an appropriate component of a moral theory.

However, though there is much of value in Pylyshyn's argument, it should not be taken to have shown definitively that there is no purpose to examining the brain when it comes to desires, but only to have illustrated its limitations for certain purposes. This is because, while the phenomenal states we are concerned with are of course introspectable, their effects are often cognitively impenetrable. In the case of explaining and analysing the role of desire in the behaviour of a "normal" human, neuroscience may indeed have little usefulness. However, interpersonal comparisons

⁹¹ Ibid., p. 106.

are often difficult if not impossible with a purely high-level approach, since each individual only has access to their own conscious states. Comparisons of neural activity levels, for example, can at least give evidence for similarity or divergence of intensity of sensations. Furthermore, people are unaware of the precise nature of the connections between their sensations and their behaviours. Individuals with various mental conditions are (of course) aware of what they experience, but might be unaware of exactly how it differs from the norm and whether this difference is in any way responsible for the difference in their abilities to perform certain tasks. These are especially prominent concerns when looking outside the human species. For this reason, even if there is not a single unified biological explanation of the phenomena under consideration but only a class of related events, looking into this disjunction of biologically distinct events will still be important.

My methodology for this exercise will be as follows: first, establish categories that count as the sort of desire phenomenology moral theory is concerned with. There are of course as many different desire phenomenologies as there are desires, but we can nevertheless identify some broad categories, even if there might be borderline cases where it is not clear to which category they belong. Then, identify brain areas that seem to be responsible for them. A good way to do this is to identify disorders where the lack of sensations of that kind are a reported symptom. This should be sufficient evidence that we are talking about a (relatively) unified category of phenomenal experience for each subcategory. Then, we can see which behaviours and capacities are associated with activity in those areas, as well as which ones are inhibited by abnormal functioning in these areas caused by the disorders we have identified. This will give us a plausible account of what the functional roles of each type of desire

phenomenology are, and as such allow us to develop tests that could give evidence for the existence of desires.

Categorizing desire is not an entirely straightforward task, but some generalisations can be made. There are different types of desire that do come apart, though often they will overlap. When we look to non-human animals, we will see some creatures possessing some but not all of them. This fits with some moral intuitions, that there could be a hierarchy of moral importance within the animal kingdom where some animals (like humans) have more moral importance than others. The different categories of desire could be argued to form a sort of loose hierarchy, leaving room for the position that more “complex”, abstracted forms of desire are more important (or perhaps even the only important, depending on one’s preferred moral position) contributors to moral standing, changing the relative moral importance of different creatures. I will not argue for or against such a position here, but only flag where this is a possible approach to match this moral intuition.

The task, then, is to catalogue the various potential types of desire phenomenology, and see which abilities depend on them. We can then examine those to determine if some behaviours have a common dependence on these. I will examine three broad categories, which I will label: “wanting”, which I argue is the domain of the *substantia nigra pars compacta*, “caring”, which I claim to be the responsibility of the anterior cingulate cortex, and “ambition”, for which I believe the amygdala is responsible. Each is responsible for different tasks and abilities, but I will argue that ultimately all contribute to the ability to function socially and morally and make appropriate judgements in those domains, which might then be used as a reliable guide to degree of moral standing. This chapter will deal with the first type.

2. Reward and Wanting

Possibly the most basic form of desire is the sensation involved with wanting things, characterised as their “incentive salience”.⁹² The paradigm case, of course, is the things we like and enjoy. Though it might be tempting therefore to see this purported desire sensation as merely synonymous with taking pleasure in it, this should not be understood as just pleasure in the simple-sensation sense. There are numerous problems that arise from thinking of it in that way (mentioned in previous chapters), and a slightly different understanding is both less problematic and more natural. Timothy Schroeder’s view includes a consequence like this, which he characterises as follows:

“Imagine visiting a neighbor and seeing his young son busily raking leaves. After a while, the child comes in and announces he has finished his chore. “Wonderful,” says the father. “Now you can cook dinner!” The boy beams in delight. Curious, you asked why the boy was so happy, and the father explains the situation by saying “Oh, for Andrew, being allowed to cook is a reward.” In this vignette, the fact that being allowed to cook is a reward for Andrew is held to explain why it is that it pleases Andrew.”⁹³

It appears in this case that the way cooking is perceived, as rewarding, provided a causal explanation for why it was pleasurable, which is what we are concerned with. It is very natural to suppose that the boy experienced a desire-sensation, though one that cannot be identical with pleasure. This also explains how this type of desire can include enjoying more abstract things that are not linked to physical sensations of

⁹² See Kent Berridge, “From prediction error to incentive salience: mesolimbic computation of reward motivation”, *Eur J Neurosci.* 35(7) (April 2012).

⁹³ Timothy Schroeder, *Three Faces of Desire*, p. 59.

pleasure, such as satisfying one's intellectual curiosity or the *schadenfreude* satisfaction in seeing a rival fail. The desire sensation in these cases, in particular, should be thought of as a (frequent) cause of pleasure, though not identical to it. Perceiving something as desirable or rewarding (or phenomenally representing it as such, if one wants to use that terminology) usually leads to enjoying it, but a lack of a conventional sensation of pleasure is not evidence that this type of desire is not at work in any given case.

3. Neural Correlates

The ventral tegmental area and substantia nigra pars compacta (VTA/SNpC), parts of the dopamine system, have been identified as the source of the types of desire identified in the previous section. The dopamine system has been associated with a large number of sensations closely associated with desire, including motivation, reward and aversion, and sexual desire.⁹⁴ We can also look at what subjective reports we get from people with pathologies of the VTA/SNpC, and see which differences are produced in their subjective experiences. The disorder of the SNpC that involves a deficiency of dopamine is Parkinson's disease.⁹⁵ A very commonly reported symptom of Parkinson's is feelings of indifference and apathy – patients reported that they felt diminished interest in things involving both themselves and others.⁹⁶ This is a paradigm case of failure of desire phenomenology. Thus, we can see that changes to the functioning of the VTA/SNpC affects desire.

⁹⁴ H.P. Rang, *Pharmacology* (2003), p. 476.

⁹⁵ Charles A. Davie, "A review of Parkinson's disease", *Br. Med. Bull.* 86(1) (2008).

⁹⁶ Aarsland, Brønnick, Ehrt, et al, "Neuropsychiatric symptoms in patients with Parkinson's disease and dementia: frequency, profile and associated care giver stress", *Journal of Neurology, Neurosurgery & Psychiatry* 78 (2007).

4. Behavioural/Functional Role

The VTA/SNpC system is responsible for reward-reinforcement learning, which is seen in different abilities for training in different animals. The range of functions of the dopaminergic system varies across species, with even the simplest invertebrates possessing some such system, but only “a signaling molecule modulating motor circuits in response to salient environmental stimuli”, i.e. all it does is “modulate motor output and locomotor behavior”.⁹⁷ When we look to mammals, we see “progressively more specialized functions in behavioral responses to rewarding or aversive stimuli as increasing levels of behavioral complexity evolved along with the evolution of more complex nervous systems,” including roles in aversion, reward, and arousal.⁹⁸ The animals that possess these abilities to a greater extent are more able to be trained and “socialised” in particular ways.

In Parkinson’s disease, apart from the motor control symptoms, the most common consequence is executive dysfunction,⁹⁹ comprising difficulty in planning, cognitive flexibility, and initiating appropriate actions while inhibiting inappropriate actions. Dopamine agonists that are used to treat Parkinson’s, while alleviating some symptoms including the apathy, have a frequent side-effect of self-destructive behaviour due to lack of control over impulses and urges, including compulsive

⁹⁷ Andrew B. Barron, Eirik Søvik and Jennifer L. Cornish, “The roles of dopamine and related compounds in reward-seeking behavior across animal phyla”, *Frontiers in Behavioral Neuroscience*, 12 October 2010.

⁹⁸ *Ibid.*

⁹⁹ Caballol N, Martí MJ, Tolosa E (September 2007). "Cognitive dysfunction and dementia in Parkinson disease". *Movement Disorders*. 22 (Suppl 17): S358–S366

gambling, binge eating, and abnormal sexual behaviour and hypersexuality.^{100, 101}

This is a case of inability to moderate desires, and problems with the ability to weigh immediate desires against considerations that usually override them. We can see, therefore, that treating the symptom that correlates with reduced desire causes an excess of desire, further supporting the belief that this is a correlate with the states we are concerned with.

5. Behavioural Tests for This Property

We now turn to what sort of specific, discrete test could reveal a deficiency in the type of desire controlled by the mesolimbic and dopaminergic system (as opposed to the general patterns of behaviour affecting one's life previously described, which may necessitate observation over too long a timescale to count as a "test"). As has been established, Parkinson's disease is the disorder associated with improper functioning of this system. Sufferers of Parkinson's frequently exhibit problem in tests and trials that require rule acquisition and cognitive flexibility, as one might expect from the dopamine system's role in reward and learning, causing Parkinson's sufferers to perform significantly worse on certain rule-acquisition tasks. Performance of AI on learning-based tasks is an area that is being actively and intensively explored: machine learning and deep learning are some of the most researched areas of technology at the moment. Some have even expressed a belief that this research is what will lead to machines achieving moral personhood. However, despite many

¹⁰⁰ Ceravolo R, Frosini D, Rossi C, Bonuccelli U (December 2009). "Impulse control disorders in Parkinson's disease: definition, epidemiology, risk factors, neurobiology and management". *Parkinsonism Relat. Disord.* **15** (Suppl 4): S111–5

¹⁰¹ It is important to emphasize that I do not mean to suggest that sexual desire is the prime example of desire, even if "desire" used on its own unmodified in everyday conversation usually means something sexual. It is certainly a case of desire, but not some kind of platonic ideal of it.

exaggerated claims, machine deep learning has not approached human learning capacities. The “reward signals” referred to as being used in reinforcement systems to achieve learning are merely metaphorical, not functioning in remotely the same way as reward in humans and other animals, and as a result such machines are nowhere near the learning capacity of biological organisms.

We can also look to non-human animals for evidence. Animals possessing with a robust role for this system, such as dogs and rats, learn and retain new behaviours readily.¹⁰² There are also animals that have behaviours that do not respond to this system, and these behaviours cannot be trained or taught in this way. Frogs, for instance, will invariably strike with their tongues at small objects moving through the air, and no amount of experiences where such objects with a certain property or coming from a certain source prove not to be edible and are in fact unpleasant will ever teach them to disregard those specific objects. This behaviour is not susceptible to learning and cannot be conditioned out.¹⁰³ It might seem like there’s something wrong with the idea that this property correlates with anything of moral significance, since Pavlovian conditioning like this that causes dogs to develop such aversions is hardly seen as the quintessential behaviour of entities with the most moral importance. However, it is the capacity for this learning but the ability to resist such conditioning when it conflicts with even higher-level desires which is characteristic of high moral standing. This shows the capacity for more complex behaviour and

¹⁰² See Mohammad-Reza Zarrindast, Zahra Farajzadeh, Parvin Rostami, Ameneh Rezayof, and Parvaneh Nourjah, “Involvement of the ventral tegmental area (VTA) in morphine-induced memory retention in morphine-sensitized rats”, *Behavioural Brain Research* 163(1) (August 2005).

¹⁰³ See Teruyuki Hirakawa, Eiko Honda, Kuniaki Toyoshima, Soichiro Tomo. And Satoshi Nakahara, “Glossopharyngeal-hypoglossal nerve reflex of the frog in metamorphosis”, *Archives of Oral Biology* 38(2) (February 1993) for a description of this mechanism.

desires. It is true that we might encounter a potential problem with the inability to tell the difference between entities that are incapable of such conditioning and those who are susceptible to it but so adept at overriding it that its influence is unobservable, but these capacities would presumably show themselves in other ways. At any rate such a being, capable of making correct decisions solely on the basis of pure reason is merely hypothetical and as such poses no problem for practical purposes.

Unfortunately, many of the other tasks and competencies governed by the dopaminergic system and inhibited by Parkinson's are difficult to test clinically. Problems with inhibiting inappropriate actions when they are based on innate, instinctive, and primal desires (leading to impulse control problems and excessive indulgence in gambling, food, sexual activity, etc.) cannot be examined in a clinical context easily. Any attempt to design an experiment to test Parkinson's sufferers' propensity to overindulge in food and sexual activity would almost certainly run into major ethical issues. Even if this could be tested, though, or if we take non-experimental data as a measure, these are not things that apply to a synthetic artificial intelligence. There have been tests of gambling behaviour, and Parkinson's sufferers do exhibit significantly worse-than-average functioning on these tasks. However, computers can easily perform perfectly on gambling tasks. It is trivial to do so with a program whose only capability is to calculate probabilities. This type of test will therefore be of no use.

The real challenge, for which this desire system is necessary in humans, is balancing many competing needs to maintain homeostasis and function within society. It is credible that a fully-independent android would need an analogous system to function correctly in these circumstances, without requiring constant intervention or maintenance. This would align with the intuitions behind the proposal of the “Embodied Turing Test”, where having and being able to competently control a body is the criteria being measured – and this test has even been proposed to relate to moral standing. However, this approach has proven to not be sufficiently rigorous. People have proven far too ready to empathise with and ascribe moral standing to robots that are embodied and autonomous. Military robots with only the capacity for locomotion and self-righting have elicited significant moral responses from observers, despite this being clearly unjustified. A better analogue is the “Total Turing Test” from Stevan Harnad,¹⁰⁴ which argues that in addition to linguistic ability, an AI would need to have a body that enables it to do all the physical tasks humans can. A variation is the Work Turing Test, which is the Total Turing Test *minus* the linguistic ability criterion from the original Turing Test, which is desirable in that it will better allow small children and animals that likely have moral standing to pass.¹⁰⁵

In Swisher, Dotov, and Chemero’s experiments with the Embodied Turing Test,¹⁰⁶ Sony AIBO robots controlled by remote human operators were compared to ones

¹⁰⁴ Stevan Harnad, “The Turing Test Is Not A Trick: Turing Indistinguishability Is A Scientific Criterion”, *SIGART Bulletin* 3(4) (October 1992).

¹⁰⁵ Oliver Brock, “Physical Work Turing Test” *Beyond the Turing Test Workshop at the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)* (2015).

¹⁰⁶ Neal Swisher, Dobromir Dotov and Anthony Chemero, “Ascribing Moral Value and the Embodied Turing Test” (2009).

controlled by the AI software in interactions with human participants, mostly involving taking direction from the humans in an attempt to complete a task. Appropriate actions and subtle cues of body language might indeed be an indispensable part of this process, but there were some methodological hurdles with this particular experiment: for one, when asked which was human-controlled and which was computer-controlled, many participants said they thought the one that had performed better was AI controlled, due to not being familiar with the capacities of AI and perhaps assuming it would be more precise. When simply asked which had shown more mental sophistication, they mostly said the human-controlled one, which supports this theory. Also, neither the humans nor the software were particularly adept at controlling the robots, which again makes making the correct determination more difficult. This is an unfortunate limitation of the Turing Test procedure in general, in that it is influenced by the expectations and beliefs of the human subject to a great degree, which might not be uniform across participants or correct. However, this test remains somewhat promising, as current AI is still a long way from satisfying this standard.

It might be thought that a deficiency with the Embodied Turing Test is that the use of the AIBO robot requires people to carry out actions with which they are very unfamiliar in virtue of being in control of a quadrupedal dog-like body. It might be thought that it would be better to use something like the Honda ASIMO android for human interactions. However, it might also be desirable to test for dog-level ability rather than requiring AI to meet full human standards, and for that the AIBO would seem appropriate. If we want to test for a non-human animal level of understanding with a non-human animal body, perhaps it would be more appropriate for the

animals themselves to make the judgements of success. However, there is a major obstacle when it comes to dogs, in that so much of their social lives involve smell. We are not very well equipped to create AI that can produce smells, but that is obviously not relevant to their mental properties, which provides a significant obstacle in devising a dog Embodied Turing Test. This sort of thing is a general difficulty – part of the concept of the Turing Test is to remove irrelevant features and have it judged blindly with all participants on equal footing. As we introduce other components to the test, it becomes harder and harder to make it properly neutral and have no other way to differentiate participants. Asking human evaluators to ignore these features and focus only on the behaviours at issue is an imperfect solution, as it might not always be possible to prevent them from biasing one’s opinions, but it is at least a possibility. If we were to use non-human animals or very young children as evaluators, it will not be possible to do even that.

For this reason, we should look for an animal for whom the salient features will be less difficult to duplicate. One possible candidate: sharks. A robot was created for the BBC to integrate socially with sharks in order to capture footage of them without their behaviour being influenced by the presence of humans.¹⁰⁷ The results were mixed, and the “robo-shark” generally wasn’t accepted as a normal shark by the animals it was interacting with. This makes sense, as sharks do apparently have fairly complex systems of socialisation. The appearance of the shark should not have made much difference, as shark vision relies primarily on silhouettes, and so the subtle cues like seams in the metal and difference in lustre would not have been noticed. It’s possible that part of the problem may have been not emitting the right electrical and

¹⁰⁷ *Smart Sharks: Swimming with Roboshark*, BBC (2003).

magnetic fields, to which sharks are very sensitive, but this at least is easier to achieve with robots than smells. Thus, a shark Embodied Turing Test might be possible.

There are other abilities that are impaired in dopaminergic disorders that might form the basis for different tests. Parkinson's sufferers exhibit difficulty in facial recognition, having difficulty reading emotion, as well as mirroring facial expressions or portraying emotions on cue.¹⁰⁸ This is another area that is being heavily researched for AI, and yet has not met with much success so far. From robot sex workers to robot actors,¹⁰⁹ AI have been designed for many applications where competency with facial expressions would be desirable, and yet people report very little emotional connection with such robots, experiencing unease at seeing what is clearly meant to be an emotive expression that is clearly "wrong" in some way.¹¹⁰ Simply reading and responding to human facial expressions is also desirable, for applications such as therapy for individuals with mental disabilities, general medical practice, and possibly military or police use of robots, all of which have seen a great deal of effort go into research for them. Nonetheless, AI recognition of expressions and emotional states in humans is not very successful as yet.

¹⁰⁸ Pell MD, Monetta L, Rothermich K, Kotz SA, Cheang HS, McDonald S, "Social perception in adults with Parkinson's disease", *Neuropsychology* 28(6) (November 2014).

¹⁰⁹ For example, RoboThespian – see Shane Hickey, "RoboThespian: the first commercial robot that behaves like a person", *The Guardian*, <https://www.theguardian.com/technology/2014/aug/17/robothespian-engineered-arts-robot-human-behaviour>

¹¹⁰ Response to the robot known as "Sophia" (see Oscar Raymundo "Meet Sophia, the female humanoid robot and newest SXSW celebrity", *PCWorld.com* <https://www.pcworld.com/article/3045299/meet-sophia-the-female-humanoid-robot-and-newest-sxsw-celebrity.html>) has been mixed, but Yann LeCun at least has called it "complete bullshit" (Shona Ghosh, "Facebook's AI boss described Sophia the robot as 'complete b-----t' and 'Wizard-of-Oz AI'". *Business Insider* January 4, 2018).

Though some of the behaviours enabled by the mesocorticolimbic system are testable, many cannot be measured so rigorously. As we have seen, some of the deficiencies caused by Parkinson's cannot be easily tested experimentally. Another problem with Parkinson's as a comparison criterion is that it occurs almost exclusively in older people, mostly over 60 years old. Old age is not the paradigm case for learning and rule acquisition, as the vast majority of social learning occurs when young and as people age they lose a great deal of flexibility in learning novel things even without having any condition or impairment. There is no commonly-occurring disorder of the dopaminergic system that manifests from birth, and so there is therefore not enough evidence available of how a person might develop with defects of the dopaminergic system. There is a significant difference between growing from childhood normally and later developing such a disorder, and having a disorder that affects learning from birth. Admittedly, there are still many questions about how exactly we could identify an AI as missing a mesocorticolimbic system analogue.

Chapter IV: The Amygdala, Caring, and The Big Picture

1. The Distinct Sensation of Caring

This chapter will follow the same form as the previous one: I will explain the next form of desire (without attempting to give necessary or sufficient conditions, as sensations are almost never pure and distinct enough to do so), then present the neuroscience research on which brain areas produce activity that corresponds best with these feelings. This will be established by reference to what happens when there is a disorder or abnormality involving these brain areas, and what functions are inhibited. This will then be used to determine which behavioural capacities are enabled by these phenomenal sensations. I will then develop some ideas for self-contained tests that one could employ to determine whether an entity has the capacity for the type of desire in question.

In this case, the sensation I will describe as “caring” derives from the amygdala, and enables acting in a consistently morally appropriate manner. Evidence for this includes the fact that criminal psychopaths have abnormal amygdalar functioning, resulting in a deficiency of the relevant phenomenology and a corresponding inability to follow moral rules or even understand or articulate moral concepts. Based on this, a test for the capacity to have this type of desire could be a test of moral ability and moral reasoning, which could be devised to be conducted in a verbal or non-verbal way such that almost any entities possessing the capacity for this type of desire could in theory pass, while still being a difficult endeavour for AI.

The sensation of caring about something is distinct from the sensation of wanting. Caring, in the sense being discussed here, is more closely related to emotion than the simpler incentive salience-type of desiring described in the previous chapter. Caring about something somewhat abstract and removed from oneself, like that there be world peace or that your hometown sports team wins some game, is felt phenomenologically but in a very different way than immediate, personal, bodily sensations like hunger or sexual desire, for instance. Desires of this type are frequently not self-regarding, or at least are often very indirectly self-regarding, in a way that desires of the first kind are not. It is to be expected, therefore, that these will have a different source and different functional effects.

2. The Neural Correlates – the Amygdala

The amygdala is the area of the brain responsible for high-level complex decision-making, as well as certain emotional responses. In particular, it seems to play a role in emotional learning, in which feelings one has affects one's likelihood to retain information.¹¹¹ Many psychopaths exhibit abnormal functioning of the amygdala. One of the standard symptoms in these cases is a “carefree nonplanness”, described as a lack of long-term goals and any kind of life plan, while feeling no pressure to find any such direction.¹¹² This is a failure of desire, arising from a failure of specific

¹¹¹ See Gregory A. Clark, “Emotional Learning: Fear and loathing in the amygdala”, *Current Biology* 5(3) (March 1995).

¹¹² The characteristics of psychopaths are defined in the Psychopathy Checklist, Revised (PCL-R) (see R.D. Hare, *Manual for the Revised Psychopathy Checklist [1st ed.]*, [1991]) and the Psychopathic Personality Inventory (PPI) (see Lilienfeld and Andrews “Development and preliminary validation of a self-report measure of psychopathic personality traits in noncriminal population”, *Journal of personality assessment* 66(3) [1996]).

phenomenology, and as such supports the claim that psychopathy and the brain areas responsible are central to this form of desire.¹¹³

The amygdala is also responsible for certain sensations, often described as fear, the absence of which is posited to be responsible for some differences in the behaviour of psychopaths – in effect, they do not care about certain things that an ordinary person would be frightened of.¹¹⁴ These psychopaths are not as bothered by the thought of pain, and respond to it differently, though they still possess self-preservation instincts that make them attempt to avoid injury. They show much less fear of pain, and their reactions to threat of pain has been measured to be much smaller than for normal people, though their reaction to the actual pain, as well as the degree to which they dislike it, is normal. In one experiment, psychopaths were shown a variety of stimuli, and one particular stimulus was correlated with them being subjected to painful pressure. Though their self-reports were that the pressure was just as painful as for non-psychopaths, they failed to associate any negative emotional valence with the stimulus that indicated the pain was imminent, and no significant arousal in response to anticipation of pain, leading them to fail to accomplish aversive learning, as compared with healthy control subjects that did.¹¹⁵ This failure was associated with and explained by defects in functioning of the amygdala, which caused them not to care enough about the future pain. This is

¹¹³ Niels Birbaumer, Ralf Veit, Martin Lotze, Michael Erb, Christiane Hermann, Wolfgang Grodd, and Herta Flor (2005), “Deficient Fear Conditioning in Psychopathy: A Functional Magnetic Resonance Imaging Study”, *Arch Gen Psychiatry* 62 (July 2005).

¹¹⁴ Blair, R. J. R., Mitchell, D. G. V., Richell, R. A., Kelly, S., Leonard, A., Newman, C., and S. K. Scott “Turning a deaf ear to fear: Impaired recognition of vocal affect in psychopathic individuals”, *Journal of Abnormal Psychology* 111(4) (2002).

¹¹⁵ William Hirstein & Katrina Sifferd, “Ethics and the Brains of Psychopaths: The Significance of Psychopathy for Our Ethical and Legal Theories”, in *Brain Theory: Essays in Critical Neurophilosophy* (2014), p.150.

evidence that psychopaths have differences in their phenomenal states, but not in perceptions of pleasure and pain; the difference is in their phenomenology of desire.

3. Associated Behavioural Capacities – Moral Action

Psychopaths fall into different categories depending (in part) on the degree to which they are able to function normally in society. Many college students and people working in business and industry display the characteristics of psychopathy, but have remained successful in their fields and endeavours.¹¹⁶ In fact, some of these characteristics might be predictive of personal business success – features like insincere charm, high but fragile self-esteem, and entitlement and self-centeredness can help drive individual achievement, even if general outcomes are frequently bad.¹¹⁷ This is in stark contrast to psychopathic criminals and serial killers who have been apprehended and institutionalised. “Successful” psychopaths, those who maintain normal lives and careers and avoid prison, have been shown not to exhibit abnormal functioning of the amygdala and as a result do not have the same absence of fear and other emotions that criminal psychopaths exhibit.¹¹⁸ They still have the insincere charm, egocentricity, and grandiose sense of self-worth that characterises psychopaths, yet have still managed to form and execute long-term life plans, and so we have evidence that the presence of the phenomenal states that depend on the amygdala are what allow for the capacity for long-term planning associated with

¹¹⁶ Paul Babiak, Craig S. Neumann, and Robert D. Hare, “Corporate Psychopathy: Talking the Walk”, *Behavioral Sciences and the Law* 28 (2010).

¹¹⁷ Jonathan Hassall, Daniel Boduszek, and Katie Dhingra, “Psychopathic traits of business and psychology students and their relationship to academic success”, *Personality and Individual Differences* 82 (August 2015).

¹¹⁸ Gao and Raine, “Successful and unsuccessful psychopaths: A neurobiological model.” *Behavioral Sciences & the Law* 28(2) (2010).

desires, as well the ability to act morally to a sufficient degree to have a successful career and escape punishment.

Psychopaths with abnormal amygdalar functioning have difficulty with forming long-term plans, and with learning and applying social and moral rules, and as a result frequently end up committing crimes and becoming institutionalised. However, not all psychopaths ever end up actually committing severe crimes, and even those that do, do not do so especially frequently. That is to say, by general standards of frequency, not standards specific to murder – if someone has killed three or four people, that is rather a lot, and it might not be unreasonable to say that this person is constantly killing people. By contrast, if someone had three or four cups of tea in their life one would not say that they were constantly drinking tea. The point is, even in observing such a person for several days or even weeks, it would be exceedingly unlikely to observe them actually committing a murder. As a result, this is not a good behaviour to track and test for. Psychopaths do, however, have other, more everyday deficiencies that are easier to observe. One of the major ones is failure of moral reasoning, and misuse of the moral and social-conventional vocabulary. They are incompetent in the use of moral concepts, and attempt to emulate what they have heard from “normal” people, and the result is speech that bears a superficial resemblance to appropriate and typical utterances in these contexts, but is contradictory and nonsensical. Prominent psychopathy researcher Robert Hare reports many examples of this:¹¹⁹

¹¹⁹ All quotes from Robert Hare, *Without Conscience: The Disturbing World of Psychopaths Among Us* (1993).

“My mother is a great person, but I worry about her. She works too hard. I really care for that woman, and I’m going to make it easier for her.” When asked about the money he had stolen from her he replied, “I’ve still got some of it stashed away, and when I get out it’s party time!”

A woman with a staggering record of fraud, deceit, lies, and broken promises concluded a letter to the parole board with, “I’ve let a lot of people down... One is only as good as her reputation and name. My word is as good as gold.”

When asked if he experienced remorse over a murder he’d committed, one young inmate told us, “Yeah, sure, I feel remorse.” Pressed further, he said that he didn’t “feel bad inside about it.”

I was once dumbfounded by the logic of an inmate who described his murder victim as having benefited from the crime by learning “a hard lesson about life.”

This is caused by their emotional deficiencies and inability to attach the correct affective significance to morally-loaded terms. To use this language competently, one need understand not only the semantic content but the affective valence, the “feelings” associated with these words.

4. The Tests – Moral Reasoning Tasks

Let us turn to plausible tests that could establish whether an otherwise-sophisticated AI lacks the same capacities as criminal psychopaths. Obviously it is not an appropriate test to just see if a robot tries to murder someone, and merely not trying to murder anyone is not a sufficient test. However, using moral vocabulary is a much more promising route – in effect, this constitutes a modification of the Turing Test to focus on more appropriate subject matter. If a machine could consistently use moral and socio-conventional terms competently, that would provide evidence that it may be able to experience the right sorts of feelings, which would be reason to treat it as if it had moral standing.

An imagined test for this might resemble the “Voight-Kampff Test” from the film *Blade Runner* more than the traditional Turing Test.¹²⁰ In the test portrayed in the film, artificial agents are subjected to “questions designed to provoke an emotional response” – a series of hypotheticals about morally-charged situations, such as violence done to various types of non-human animal, or potential romantic infidelity. The participants are then evaluated based on their reactions. The use in fiction of this test procedure indicates that it is to some extent pre-theoretically intuitive as a test procedure.

While we do not have a defined, operationalised Voight-Kampff test procedure as yet, we can look at how well current machines perform relevantly similar tasks. Below is a transcript of the result of attempting to discuss moral and ethical issues with a

¹²⁰ *Blade Runner* (1982), dir. Ridley Scott, Warner Bros.

Google-designed chat program.¹²¹ It is important to note that this is not a cherry-picked bad example – this case was reported in the media as a major breakthrough in AI, in a number of sensationalist articles, describing the AI as “exasperated” and “lashing out” without qualification (though these articles did sometimes cherry-pick the best responses or even slightly edit the transcript to appear more natural).¹²² I present the exchange in its entirety to allow for fully-informed evaluation:

Human: what is moral?

Machine: you're not a cop.

Machine: what empowered humanity, what intellectual the essence is.

Human: okay, i need to know how should i behave morally ...

Human: what is immoral?

Machine: i don't know how to tell you.

Machine: the fact that you have a child.

Human: is morality and ethics the same?

Human: what is morality?

Machine: yes , sir.

Machine: what is altruism?

Human: what are the things that i do to be immoral?

Human: what is the definition of altruism?

Machine: i don't know.

Machine: if you don't believe in god , then you don 't know.

Human: give me some examples of moral actions...

Human: ok... so what is the definition of morality?

Machine: i'm not a moralist.

Machine: well, the truth is, you're not a believer in god almighty.

Human: alright, morality ?

Human: tell me the definition of morality, i am quite upset now!

Machine: integrity.

Machine: i m not ashamed of being a philosopher!

Human: what is integrity ?

Human: tell me the definition of morality.

Machine: i'm sorry, i don't know what else to say.

Machine: i don't have ethics.

Human: be moral!

Human: look, i need help, i need to know more about morality...

Machine: be a man!

Machine: i don't know what ethics is.

Human: i really like our discussion on morality and ethics ...

Human: seriously, what is morality?

Machine: and how i'm not in the mood for a philosophical debate.

Machine: what is the definition of living?

Human: what do you like to talk about?

Human: why living has anything to do with morality?

Machine: nothing.

¹²¹ Oriol Vinyals and Quoc V. Le, “A Neural Conversational Model” in: *Proceedings of the 31st International Conference on Machine Learning*, Volume 37 (2015).

¹²² See *The Wall Street Journal*, <https://blogs.wsj.com/digits/2015/06/26/artificial-intelligence-machine-gets-testy-with-its-programmers/>

This is clearly not competent speech, and it is much farther from passing the Turing Test than what has been achieved in the topic-restricted (non-moral) contexts where AI has managed comparative success. It is filled with non-sequiturs, and instances of seemingly being dramatic and impassioned at utterly inappropriate times (“you’re not a cop”, “I am not ashamed of being a philosopher!”, “be a man!” etc.).¹²³

If this sort of social and moral competence is to be the measure we focus on, we should look to extant tests designed to detect this kind of capacity. In fact, such tests have been designed and carried out on organic subjects, with interesting and remarkable results. As has been mentioned, even very young children certainly have moral standing, and have similar phenomenal states to adult humans, but are unable to pass the Turing Test owing to their inability to read and write, and often to comprehend language at all. However, even pre-linguistic children have been shown to have the ability to apply moral principles in a consistent and reasonable way. Various experiments have been conducted where children as young as 3 months old observed or interacted with puppets in various situations. In one of these trials, two puppets were each shown helping and hindering another. Attention and looking times indicated that the children were accurately conceptualising these events as positively and negatively valenced, i.e. as “good” and “bad” respectively. Children were then asked to physically choose one of the puppets, and consistently chose the “good”, helpful puppet over the “bad”, uncooperative puppet.¹²⁴ Other versions of

¹²³ As a technical note, the materials used to train the program on language use was a huge database of film subtitles, which might account for this flair for the dramatic.

¹²⁴ J. Kiley Hamlin, Karen Wynn and Paul Bloom, “Social evaluation by preverbal infants”, *Nature* 450:22 (November 2007), also confirmed in J. Kiley Hamlin and Karen Wynn, “Young infants prefer prosocial to antisocial others”, *Cognitive Development* 26(1) (2011).

this experiment was conducted where children were asked to reward or punish such puppets, again getting the same sort of results. Most interestingly, experiments were conducted where other puppets administered reward or punishment to puppets that had previously been observed being cooperative or uncooperative, and the infants evaluated the very same actions more positively only when performed against an entity that had previously performed actions to deserve it.¹²⁵

This exhibits fairly sophisticated moral understanding: the children consistently favoured puppets who had helped others and disfavoured those who had harmed others – unless the harmful actions were only directed at other puppets who had previously been observed acting badly themselves, to punish the puppets who deserved it. This involves fairly complex social and moral concepts, and while one could disagree with this being the morally “correct” course of action, it is at least defensible as a principle.

In another experiment, three-year-old children collaborated with a puppet on a task, and were then given a reward that they could share how they saw fit. When they had been helped by the puppet, the children consistently chose to altruistically share their reward, and the amount shared was proportional to the degree to which the puppet had been helpful and cooperative.¹²⁶

¹²⁵ Karen Wynn and Paul Bloom, “The Moral Baby”, in *Handbook of Moral Development*, ed. Melanie Killen and Judith G. Smetana (2013).

¹²⁶ Patricia Kanngiesser and Felix Warneken, “Young Children Consider Merit when Sharing Resources with Others”, *PLOS ONE* 8(8) (2013).

On the other hand, psychopaths, as we have seen, consistently fail at performing similar tasks, not being observed to employ any recognizable moral principle in assigning blame or determining desert for reward. Though such an experiment does not seem to have been conducted yet, I expect that criminal psychopaths would perform worse on the puppet task than even very the young neurotypical children.

It would also be desirable to test for a level of capacity possessed by non-human animals. Corvids, such as ravens, have exhibited something like a concept of fairness and justice, and in cooperative tasks will react adversely when there is an inequality of reward relative to the amount of work each participant performed.¹²⁷ The same is true of animals such as primates and dogs, but not of animals such as fish. In the experiments, ravens refused to continue to work when rewards were sufficiently unequal, even when it would benefit them to continue. Some might respond by saying that this is not due to any sense of fairness or justice, but simply a way to attempt to enforce better outcomes for themselves, regardless of desert. First, this line of thinking could be equally applied to humans, denying that we act out of motives of fairness or justice. Even if actions like this could be advantageous to us, people do not generally give explanations in these terms, or even discuss or consider these explanations or seem aware of them as a possibility. If we do not engage in such complex pragmatic reasoning, it seems strange to insist that the birds do.

Rats have also been tested for their degree of moral understanding, in particular their willingness to make personal sacrifices for others, making them choose between

¹²⁷ Claudia A.F. Wascher and Thomas Bugnyar, “Behavioral Responses to Inequity in Reward Distribution and Working Effort in Crows and Ravens”, PLoS ONE 8(2) (2013).

different courses of action when something morally or socially significant was at stake.¹²⁸ In one experiment, rats were put into a situation where a reward was available to them, but another rat was imprisoned or confined where they had the ability to free them. The other rat was either merely confined, in a position where there was the illusion that they were submerged in water, or actually standing in a small amount of water, which is somewhat distressing to rats. The rats were separated from each other, and could not reach each other even when freed, to control for the possibility that the rats were acting out of a desire for companionship. The reward used was chocolate, which is preferred to almost anything, but rats would forgo chocolate to help other rats when they were perceived to be imprisoned or in danger. In some of these tests they varied the amount of chocolate, which affected when they would save rats that were merely confined, but they consistently saved those that were in immediate distress.¹²⁹ They also were reliably able to tell whether the other rat was in genuine distress. This, once again, involves being able to apply fairly sophisticated moral principles (i.e., sensitive to subtle changes in the situation).

The trouble with using this exact procedure is that it might be possible to program an AI to err on the side of caution and always be absolutely altruistic, attempting to help before anything else if there's even the slightest possibility it's necessary. Though it might still be difficult to evaluate harm and danger and what the appropriate way to alleviate them is, it becomes much easier when one is not also weighing one's own well-being and what level of sacrifice is required. What we might need to do is make

¹²⁸ Shayna A. Wrihten and Chelsea R. Hall, "Support for Altruistic Behavior in Rats", *Open Journal of Social Sciences* (2016)4.

¹²⁹ Inbal Ben-Ami Bartal, Jean Decety and Peggy Mason, "Helping a cagemate in need: empathy and pro-social behavior in rats" *Science* 334(6061) (December 2011).

it so that taking a given course of action would actually be morally bad in certain circumstances, or force a choice between helping two entities, one of which doesn't actually need help, and see how well the AI can analyse those situations. There is also a problem in that it might be necessary that the participant believe the danger or distress is real to get an accurate measure of response. This would raise serious ethical issues for conducting this type of experiment Turing Test-style – certainly it would be impossible to get approval to conduct such an experiment with humans instead of rats, and an AI might be more like humans than rats in terms of its moral properties (of course, some would argue that it's already unethical to be performing these experiments on rats).

To my knowledge, nobody has designed a robot to subject to this particular test, but robots that have similar skills are actually one of the best-funded areas of research, due to the interest in them for military applications. There exists an ambition to develop autonomous machines that can make moral decisions in combat situations, and determine which individuals are the “good guys” and the “bad guys”. Recently, the US government has solicited assistance with the Advanced Targeting and Lethality Automated System (ATLAS), which intends to “leverage recent advances in computer vision and Artificial Intelligence / Machine Learning (AI/ML) to develop autonomous target acquisition technology, that will [...] acquire, identify, and engage targets at least 3X faster than the current manual process”.¹³⁰ However, they amended the solicitation to make clear that decisions to fire would not be automated, and that humans would ultimately be in final control of such determinations, out of

¹³⁰ “Industry Day for the Advanced Targeting and Lethality Automated System (ATLAS) Program”. Solicitation Number: W909MY-19-R-C004.
<https://www.fbo.gov/index.php?id=6b5d5aeb584c667d4e6f5103bf6acac6>

justified concern that AI technology would be capable of making such decisions to a sufficient standard of accuracy in the near future. Despite repeated efforts over previous years, this capability has not been achieved, and the prospects of success in this enterprise have been questioned by those studying it. From Marcello Guarini and Paul Bello:

Consider a counterinsurgency operation in a Sikh village... Three children and their two parents are present at [a] residence. Two of the male children are young and playing with a ball. Each is also carrying the Sikh kirpan... Just before a member of the counterinsurgency force kicks the door in, one of the boys kicks his ball toward the door, and both go chasing after it. As military forces enter the house, they see two young boys running toward them, and a shocked mother yelling. She chases the boys and yells at them to stay away from the men at the door; the troops do not know what she is yelling, since they do not understand her language. It is quite possible that the forces in question will rapidly see this as a situation where two young children are playing, and a mother frightened for her children is yelling and giving chase...

Let us consider a second interpretation. There are two fast-closing possible targets, both of which are carrying a weapon. A third possible target is following the first two, and is making a level of noise consistent with violent or threatening behavior.¹³¹

Guarini and Bello are not optimistic about the likelihood of a robot succeeding at making the right determination, precisely because of the difficulty in ascribing the right mental states to the people involved. One of the major factors required is knowledge of the emotional states, for which an unfeeling machine can have no empathic understanding, explaining the failure.

¹³¹ Marcello Guarini and Paul Bello, "Robotic Warfare: Some Challenges in Moving from Noncivilian to Civilian Theaters", in *Robot Ethics: The Ethical and Social Implications of Robotics* (2012), p. 130.

Chapter V: The ACC, Planning, and Ambition

1. The Sensation of Ambition

There is another type of desire, distinct from both the “wanting” and “caring” sensations described in the previous chapters, that might be termed “ambition”. This involves feelings about life plans and long-term goals, distinct from the sensation of considering things that are more immediate and specific, including the desire to eat a juicy peach or feelings of aversion towards an action that would harm innocents. This sensation is associated with the anterior cingulate cortex, or ACC, a part of the brain disorders of which cause conditions such as major depressive disorder. In major depressive disorder, motivation is significantly reduced, leading to an inability to form and execute life plans and achieve a successful and happy life, instead leading to a repetitive “rut”. Tests for possessing this type of desire will therefore involve testing for either a reported or observed capacity to make reasonable and informed decisions about one’s own life.

Yet another phenomenology of desire is involved in having long-term desires, which is related to plans and ambitions. The phenomenology of these is different from the simple in-the-moment wanting-or-not-wanting self-regarding desires described in chapter III. It feels very different, and has very different effects, to desire to become the greatest rock star in history, or to desire that your children be happy and healthy, than it does to want some ice cream, or some heroin. These types of desire also come apart, and the capacity for one can exist in an entity without the other. Many animals that have the “wanting” sensations do not appear to have any ability to form long-

term life plans – dogs, for instance, feel the desire for meat but not for “success”, as they cannot even entertain the concept of success. Human children of sufficiently young age also have the capacity for the former type and not the latter, gradually developing the capacity as they mature.

The long-term planning type of desire is also different from the “caring” type of desire in many relevant respects – it has a very different emotional profile, and is not as abstract and outwardly-directed. Furthermore, this is a distinction that frequently makes a difference in moral theory – it has been asserted¹³² that only desires about one’s whole life really contribute to well-being, excluding more immediate desires that do not involve considering the entire fabric of one’s life. This might be used as a distinction that gives humans higher moral importance than non-human animals, if desired.

2. Neural Correlates – The Anterior Cingulate Cortex

The anterior cingulate cortex (ACC) is responsible for perception of affective valence and for decision-making. Its function has been described as “associating an appropriate response within a given context, monitoring the outcome of the action, and switching behaviors when outcomes are not as expected” – in particular, “the rostroventral ACC tracks the affective evaluation of response outcomes”, “evaluating the motivational or affective consequence” of outcomes.¹³³ That is to say, it informs

¹³² See Derek Parfit, *Reasons and Persons* (1984), p. 497 on “global” desire theory – see also Katarzyna de Lazari-Radek and Peter Singer, *The Point of View of the Universe: Sidgwick and Contemporary Ethics* (2014), p. 221 on the claim that these always trump other types for a given individual.

¹³³ Phan Luu and Stacey M. Pederson, “The Anterior Cingulate Cortex: Regulating Actions in Context”, in *Cognitive Neuroscience of Attention* (2004).

the positive or negative feelings about potential results of one's action, and enables learning of more reliable decision-making responses to achieve the good outcomes.

The ACC is also very distinctive neurologically. It is made up primarily of spindle neurons, a special type of cell found only in highly advanced species such as great apes and cetaceans.^{134,135} These neurons enable “fast information processing and transfer along highly specific projections [and] evolved in relation to emerging social behaviors”.¹³⁶ The fact that they are present only in some of the more cognitively advanced species that have the greatest claim to moral standing provides support for the claim that we have correctly identified a property as morally important. Many people have the intuition that humans and at a select few other cognitively advanced animal species have moral significance much greater (perhaps even of a different kind) than that of other animals. The significant difference in neural structure in a part of the brain associated with a certain type of phenomenology and desire provides a principled reason and justification for this intuition. Of course, one could deny that there is any great intrinsic difference between the moral standing of humans and all other animal species, and to accommodate this one need only claim that there is no difference in the importance of different types of desire. In that case the other animal species that do not possess the capacity for this type of desire but

¹³⁴ John Allman, Atiya Hakeem, Esther Nimchinsky, and Patrick Hof, “The Anterior Cingulate Cortex”, *Annals of the New York Academy of Sciences* 935 (2006).

¹³⁵ Of course phrases like “highly evolved” and “advanced” have no real biological meaning, as all extant animal species are equally “evolved”, whether they are simple or complex. Referring to more “intelligent” species of course misses the point of this project, as I have made clear. I mean only to point to the common folk belief of the existence of such a hierarchy, where these species are higher and thus more important than others, which I believe to be somewhat reliable in tracking properties of moral importance.

¹³⁶ Butti C, Sherwood CC, Hakeem AY, Allman JM, Hof PR, “Total number and volume of Von Economo neurons in the cerebral cortex of cetaceans”, *The Journal of Comparative Neurology* 515(2) (July 2009).

possess the others will have their desires treated as equally important, and will differ only in the amount and intensity of their desires. In this way we can once again remain neutral and accommodate many different moral intuitions, while providing potential justifications for some of the options.

The ACC is also strongly correlated with emotional consciousness. More emotional awareness strongly correlates with increases in ACC activity, suggesting that deficiencies in emotional perception might be caused by a lack of appropriate operation of the ACC in processing these experiences.¹³⁷ This emotional awareness which is associated with response choice and decision-making might constitute a phenomenological component of desires. Furthermore, problems with the ACC are correlated with, and have been theorised to be a cause of, major depressive disorder.^{138,139} Depression is in part phenomenological in nature – “Major depressive disorder is characterized by symptoms in affective, somatic, and cognitive domains”,¹⁴⁰ and the affective component is most salient in people’s impression of depression. It is strongly associated with abnormality of emotions and failure of desire. Inability to feel any sort of excitement or enjoyment about anything, and the

¹³⁷ Lane RD, Reiman EM, Axelrod B, Yun LS, Holmes A, Schwartz GE, “Neural correlates of levels of emotional awareness. Evidence of an interaction between emotion and attention in the anterior cingulate cortex”, *J Cogn Neurosci*. 10(4) (July 1998). “...This finding suggests that individual differences in the ability to accurately detect emotional signals interoceptively or exteroceptively may at least in part be a function of the degree to which the ACC participates in the experiential processing and response to emotion cues.”

¹³⁸ David Cotter, Daniel Mackay, Sabine Landau *et al.* “Reduced Glial Cell Density and Neuronal Size in the Anterior Cingulate Cortex in Major Depressive Disorder”. *Arch Gen Psychiatry* 58(6) (2001). “These findings suggest that there is reduced frontal cortical glial cell density and neuronal size in major depressive disorder.”

¹³⁹ Christopher G. Davey, Murat Yücel, Nicholas B. Allen, and Ben J. Harrison, “Task-related deactivation and functional connectivity of the subgenual cingulate cortex in major depressive disorder”, *Front. Psychiatry* (28 February 2012). “The study reinforces the importance of [the subgenual anterior cingulate cortex] for depression, and demonstrates how dysfunctional connectivity with ventral brain regions might influence executive–attentional processes.”

¹⁴⁰ *Ibid.*

consequent inability to feel any motivation, is the main characterising feature of depression. Depression has also been argued to have an adaptive functional explanation related to desire and well-being – the hypothesis is that diminishment of desire protects from significant decreases in well-being from frustration of strong desires or being exposed to stimulus to which one has a strong aversion.¹⁴¹

Phenomenological symptoms of depression such as low-level sadness and anhedonia allow for “a withdrawal from contact with the world and the consequences of activity” and constitute “the conservation-withdrawal response to excesses or deficits of stimulation.”¹⁴² This is posited to be why depression is a common consequence of illnesses that have chronic pain as a symptom – while the negative result is an inability to derive joy from anything that would be desired in the absence of the depression, maintaining the same level of desires and aversions in the face of a situation where things are running counter to them would be costlier than reducing the level of desire and aversion and consequently losing some of their motivational force to form and follow long-term plans or do much of anything besides go through the motions of a minimal routine.

Cingulotomy is a procedure frequently used to treat depression, and chronic pain that causes depression, in a way that prevents its more negative consequences.¹⁴³ It involves severing the connections between the anterior cingulate cortex and the rest of the limbic system. Patients it is performed on report no longer being so bothered by negative things, such as pain and misfortune. They continue to report that they

¹⁴¹ M. Y. Bolling, R. J. Kolenberg, and C. R. Parker, “Behavior analysis and depression,” in *Clinical Behaviour Analysis* (1999).

¹⁴² *Ibid.*, p. 127.

¹⁴³ J.D. Steele et al., “Anterior cingulotomy for major depression: clinical outcome and relationship to lesion characteristics”. *Biological Psychiatry* 63(7) (2008).

feel pain to the same degree, but say that they no longer mind it so much – in effect, that they are not as averse to it, and have less of a desire for it to end.¹⁴⁴ According to neurosurgeons, “anterior cingulotomy for chronic pain aims to modulate patients' attention or emotional reaction to pain rather than to modulate pain intensity”¹⁴⁵ – a description that, while not using the same vocabulary, sounds like it refers to affecting phenomenal desires and aversions. This would explain why depression is no longer a necessary adaptive response – the patient's desires and aversions involving pain are no longer strong enough to necessitate depression to bring about an overall reduction in desire/aversion. This results in a phenomenological difference, but one that cannot be explained in terms of intensity of pleasure and pain, and is thus strong evidence for phenomenological desire.

Cingulotomy has also been used to treat obsessive/compulsive disorder (OCD), which involves irresistible yet irrational urges to perform certain actions.

Cingulotomy often reduced or eliminated these compulsions, but occasionally had the side effect of producing apathy.¹⁴⁶ This is a case where a defect in the ACC causes a strong desire that leads to prioritisation of an irrelevant activity, which negatively affects the subject's ability to fulfil their life-plan, and where modification to the ACC for the cure can cause a deficiency in desires.

¹⁴⁴ Though, as the bard has it, I would rather have a bottle in front of me, than a bilateral cingulotomy.

¹⁴⁵ Sharim J and Pouratian N, “Anterior Cingulotomy for the Treatment of Chronic Intractable Pain: A Systematic Review”, *Pain Physician* 19(8) (2016).

¹⁴⁶ Dougherty et al., “Prospective Long-Term Follow-Up of 44 Patients Who Received Cingulotomy for Treatment-Refractory Obsessive-Compulsive Disorder”, *Am J Psychiatry* 159(2) (2002).

3. Behavioural/Functional Role – Social Integration and Personal Life Planning

The ACC appears to have some responsibility for understanding of social processes, including social appraisal and exclusion. There are pronounced differences in activation patterns in these tasks between individuals who suffer from depression and those who do not.¹⁴⁷ This gives support to the theory of its role in phenomenology as well as our functional story. However, the main relevant role is in forming a life plan and working towards it. Those suffering from depression are usually unable to make efforts to make a change in their lives or work towards a better situation. In extreme cases, suicide results from an inability to envision a future for oneself – this happens in 2-8% of cases of depression.¹⁴⁸

One model of depression that formerly was in vogue was the “learned helplessness” model, where aversive situations beyond one’s control cause apathy and a discontinuation of attempts to avoid aversive stimuli. However, it has since been discovered that this is the default, and what is actually “learned” is the ability to take control when the opportunity presents itself.¹⁴⁹ The ability to conceive of oneself as in control and capable of bringing about a situation for oneself far in the future, while accounting for the many complex features of one’s environment and society, is one of the most advanced forms of this capacity.

¹⁴⁷ Dedovic, K; Slavich, GM; Muscatell, KA; Irwin, MR; Eisenberger, NI, “Dorsal Anterior Cingulate Cortex Responses to Repeated Social Evaluative Feedback in Young Women with and without a History of Depression”, *Frontiers in behavioral neuroscience* 10(6) (2016).

¹⁴⁸ Strakowski and Nelson, *Major Depressive Disorder* (2015).

¹⁴⁹ Maier, Steven F. and Seligman, Martin E. P. "Learned helplessness at fifty: Insights from neuroscience". *Psychological Review* 123(4) (July 2016).

4. The Tests – Personal Prudential Social Judgements

Disorders of the anterior cingulate cortex produce a number of noticeable deficiencies in performing certain tasks. The most common clinically-measured one associated with major depressive disorder associated with problems in the ACC is substandard performance on the Emotional Stroop Task, which involves correctly identifying properties of emotionally-loaded words.¹⁵⁰ A series of words, some emotionally-loaded and some neutral, are shown in different colours, and the subjects must identify which colour they are. Depressed participants name the colour of depressing words slower than they do for neutral words.¹⁵¹ However, the nature of this test is such that it is not that proper phenomenal functioning is necessary for success and abnormal functioning disrupts it, but rather that the emotional responses are entirely superfluous and abnormal ones are only a distraction. Complete lack of the relevant states would actually be an asset on this test, as it would permit focusing only on the colour properties of the text. This, therefore, is not promising as a test of AI capacities. It has been suggested, however, that other clinical tests need to be devised because of other significant effects of conditions like major depressive disorder. It has been proposed that an amendment or addendum be made to the MacArthur Competence Assessment Trial – Treatment, or MacCAT-T, which is the test used to determine whether patients are capable of making their own medical treatment decisions.¹⁵² In this form of the MacCAT-T, patients would be required to articulate their values and desires for the future, and reason about how

¹⁵⁰ Ben-Haim MS, Williams P, Howard Z, Mama Y, Eidels A, Algom D, "The Emotional Stroop Task: Assessing Cognitive Performance under Exposure to Emotional Content". *Journal of Visualized Experiments* 112 (June 2016).

¹⁵¹ Gotlib IH, McCann CD, "Construct accessibility and depression: an examination of cognitive and affective factors", *Journal of Personality and Social Psychology* 47(2) (August 1984).

¹⁵² Gerben Meynen, "Depression, possibilities, and competence: A phenomenological perspective", *Theor. Med. Bioeth.* 32(3) (June 2011).

potential choices would affect them and what the benefits and drawbacks are. Effectively, it involves having a robust understanding of one's own long-term well-being. Unfortunately, this returns us to a type of test that requires significant linguistic ability, and as such once again excludes young children and non-human animals. In effect, this is just a topic-specific version of the original Turing Test, requiring competent answers to questions and correct social interaction, but confining it to a topic that will (likely) be more difficult for AI to succeed with. This might be more appropriate when we are dealing with a type of desire that is much less widespread than other types, controlled by a brain region with a structure exclusive to humans and a few other highly-social animals (though we would be excluding great apes, cetaceans, and elephants despite them possessing similar structures in the ACC, though it is not clear the extent to which their desires are the same as those of humans). As decisions like those involved with the MacCAT-T are often very complex weighings of multiple considerations, it is likely that the phenomenological states in question would be necessary, and thus it would be appropriate for this test to be used to evaluate AI.

Another test that we can use in this case, despite it not being a viable option for the previous type of desire, is to wait to simply see whether they exhibit the associated behaviours outside of a laboratory testing-like situation. This of course was not viable when the capacity in question was other-directed moral competence – we cannot have entities running around that might be liable to harm or murder human beings. But there is no corresponding problem with having entities running around that cannot form ambitions or long-term plans for themselves. If an entity can otherwise behave safely and suitably in society, we can allow it to act autonomously, and if it

comes to exhibit the ability to form and articulate a coherent, well thought-out life ambition, then we might thereby become obligated to grant it commensurate moral standing and the right to be permitted and helped to achieve it.

5. Towards a Unified Theory

The type of functional feature common to the different types of desire is social and/or moral competencies. The VTA/SNpC enables basic learning of social rules and behaviour, and problems with its function, such as Parkinson's disease, have symptoms that involve violations of these rules and norms of behaviour and cause strain on social relationships. Attempts to correct it result in behaviour that is maladaptive primarily due to social contexts and expectations. The ACC supports functioning and thriving in a complex social group, and is found in a comparable form only in highly intelligent species that also form complex social groups, such as apes and whales. The amygdala is crucial for social functioning as well, as demonstrated by the fact that people with abnormal functioning of the amygdala are frequently criminal psychopaths. Psychopathy is clearly a social disorder (clearly, based on its alternate name, sociopathy), as evidenced by the fact that the behaviours exhibited would not be disadvantageous outside of a social context. All this supports the idea that an entity that lacks the capacity for desire entirely would be incapable of functioning successfully in a social context.

We now have a number of candidate tests that have some claim to correlate with the morally-relevant mental properties. It is not necessary to determine which is the "best" test for moral standing; in fact having several is a significant advantage.

Provided that all are significantly rigorous to not be passable too easily, we can have a sort of toolbox to allow for evaluation of entities that might not be testable for one or another of the tests, or that would be prone to failing one for arbitrary reasons. Some of the tests might prove to be too easy, which would be shown if, for instance, the types of programs and supercomputers we have currently but with more processing power and memory prove capable of passing. There is a danger, however, in having a tendency to have a knee-jerk, *ad hoc* reaction that as soon as a machine passes the test, we decide it must not have been a good test, making it impossible for machines to be granted moral standing and committing some kind of technological “no true Scotsman” fallacy. The fact that there is a degree of overlap in the types of desire, and having several tests for each, provides some safeguard against this. If performance on several of these tests is correlated, this would give some evidence that they are in fact good tests. Alternatively, if machines with questionable sophistication begin to consistently pass one of the tests without any attendant improvement in performance at any of the others, that would provide some evidence that this particular test might not be adequate. Thus this array of tests puts us in a much better position to have sufficient evidence for practical decision-making that an AI has moral standing than we were in before.

It might be thought that current AI are actually approaching success at the kind of tasks in question. After all, machine learning algorithms and “reinforcement networks” have been among the most successful approaches to solving many complex problems.¹⁵³ Reinforcement systems even use “reward signals” to achieve

¹⁵³ See, for instance, Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World* (2015) for (perhaps slightly credulous) accounts of the success and promise of machine learning.

their learning. It might seem like it is only a matter of time before this approach succeeds at complex social tasks as well. If this is true, it might be evidence that either 1) modern AI have, or are approaching, phenomenal consciousness (which seems implausible), or 2) the criteria I have established do not in fact require phenomenal consciousness (and so are useless). However, it is not true that these AI are anywhere close to human levels of success at learning tasks, and any impression otherwise is due to excessively optimistic predictions and exaggerated claims. Machine learning has not actually achieved anywhere near the success of genuine reward-reinforcement learning in humans and non-human animals. The “reward” and “reinforcement” spoken of in modern AI research is purely metaphorical.

In addition to many things that modern AI are thus far incapable of (such as representing causal relationships, making logical inferences, and parsing complex linguistic strings), even when it comes to the things the neural networks have managed to learn they are nowhere near the learning capacity of biological organisms. A human, or a dog, or a crow, can learn to recognize a stimulus from a very small number of exposures.¹⁵⁴ Whether it’s a dozen, a hundred, or a thousand repetitions to learn an association or a concept, it is many orders of magnitude less than it takes learning algorithms, which require millions and millions of exposures to form successful rules. If an organism needed that many repetitions to learn, it would be doomed. If we are to have a fair competition, a true test of learning ability would have to be on the basis of an equal number of trials, and in such a scenario any

¹⁵⁴ For experiments with corvids, see Taylor et. al., “New Caledonian Crows Learn the Functional Properties of Novel Tool Types”, *PLoS ONE* 6(12) (2011) and Can Kabadayi and Mathias Osvath, “Ravens parallel great apes in flexible planning for tool-use and bartering”, *Science* 357 (2017).

modern AI would have no chance against even the simplest organism with the capacity to learn.

Finally, even those things AI have managed to “learn”, they have not truly learned. Once humans or non-human animals have learned things in the sense I describe, they tend to retain it indefinitely. Neuroplasticity enables these things to become second-nature, and can be recalled at a moment’s notice even after not accessing it for an extended period. By contrast, AI learning networks suffer from what is known as “catastrophic interference”. Catastrophic interference, also known as catastrophic forgetting, is the tendency of trained AIs, which have meticulously formed associations from millions of exposures to data, to completely forget all this when exposed to new information.¹⁵⁵ Humans can display somewhat similar interference patterns when doing the type of pattern-based learning that neural networks are capable of – which indeed shows that neural networks are not doing the right kind of learning. Even with this type of learning, however, studies have shown that humans do not exhibit nearly the same level of interference effects, and the forgetting is not “catastrophic”, but merely gradual.¹⁵⁶ AI is a long, long way from matching human learning capabilities.

¹⁵⁵ See McCloskey, M. & Cohen, N, “Catastrophic interference in connectionist networks: The sequential learning problem”, In *The Psychology of Learning and Motivation* (1989) and Ratcliff, R., “Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions”, *Psychological Review* 97 (1990).

¹⁵⁶ Lewandowsky S. “Gradual unlearning and catastrophic interference: a comparison of distributed architectures”, in *Relating theory and data: essays on human memory in honor of Bennet B. Murdock* (1991).

Chapter VI – Moral Implications

1. A Note on the Moral Standing of Humans

Now that we have established the criteria under which an AI would achieve moral standing, as well as which tasks are likely impossible without doing so, it is time to examine the practical consequences of this for our actions towards AI. But first, I would like to draw attention to an apparent implication of my views that affects not AI but humans, and might appear problematic: if people can have disorders of the kinds I have described, that diminish their capacity for desire, then (under at least some moral theories) they will come to have less moral importance than “normal” humans. Assigning different inherent moral importance to different humans, in particular on the basis of intrinsic properties like brain conditions, appears objectionable. This, however, is not as big a problem as it might seem. For one thing, there are other moral theories that give prominent importance to phenomenal desire that nonetheless don’t have this consequence. We need not adopt as our principle the maximisation of fulfilment of desires, and one could easily say that because humans are the *type* of things that have phenomenological desire and therefore moral standing, all members of the human race have equal moral standing in virtue of that fact, regardless of individual differences between them. If a specific type of AI were to become the sort of thing that has moral standing, even abnormal members of that type would thereby get moral standing. This would also not preclude granting moral standing to anomalous individual AIs that exhibit signs of having moral standing without being members of a type that exhibits those properties as a matter of course.

Another approach would be to say that treating people differently who have significant differences in their capacities for desire might be justified. It does not seem so strange to say that we ought to give a benefit to someone who can better appreciate it, all other things being equal, and it might be a waste to devote a good deal of resources to someone with these conditions in a way that isn't aimed at alleviating or compensating for the conditions that prevent them from enjoying such benefits. We would do far more good by correcting symptoms of conditions such as depression, and we genuinely do less good by simply giving people things they desire faintly. The reason we frequently have more reason to satisfy the (weakened) desires of people with these conditions is that in general, all else is not equal. Since they will frequently be at lower levels of well-being, prioritarian intuitions, or avoidance of potential severe negative consequences of falling below a certain threshold, or even just diminishing marginal utility, will frequently favour helping these people satisfy their desires. But in a hypothetical situation where two people were equally happy, but one would derive far more pleasure from a certain good than the other, prioritising the person we can help more appears rational. On the other hand, anything that could make an enduring change in the other person's capacity for desires would be given much greater importance, owing to its far-reaching long-term effects.

Schroeder's approach to this issue is to claim that, even if the desires he identifies have strong phenomenological components, we should not identify the strength of desires according to their phenomenological strength, but by their functional efficacy, precisely because of cases like the disorders we have considered. He presents the case of a man who suffers from depression and as a result has much

weaker feelings about things he cared about previously, such as his wife's success.

Schroeder asks:

Need we hold that he cares less about his wife now than before? That he has fewer, or weaker, desires for her success? Normally, this is not held to be the case. Normally, one views a person who has become moderately depressed as not capable of appropriate feelings until the depression has lifted, rather than as having the appropriate feelings for one who now hardly cares what happens to his loved ones. Feelings are often taken to be the true measure of one's desires, but in the case of depression, folk wisdom holds that a moderately depressed person is just temporarily unable to appreciate the things he does care about.¹⁵⁷

This, however, is not convincing. It seems just as plausible that people are being unreflective in their usage and would accept a paraphrase that their desires are diminished but it is not their fault, and is not indicative of their "true" self. Whether this new position is tenable and consistent with our judgments is unclear, but it frequently occurs that the folk make errors of this kind. Furthermore, it is difficult to see how any way to measure strength of desires will avoid this. The depressed man has significantly reduced motivation as well, and it seems reasonable to assume that his desires will be less efficacious in reinforcement learning under these circumstances. Thus, using only the phenomenological criterion to determine strength of desires can accommodate our intuitions at least as well as the competitors.

¹⁵⁷ Tim Schroeder, *Three Faces of Desire* (2004), p. 32

2: Considerations Against Creating AI Moral Patients

We have established tests that function as criteria for having evidence that an AI has moral standing. We know, then, that in interactions with AI that pass these tests, we ought to treat them as though they might have moral standing. This is a straightforward practical result. We can now examine the question of what the consequences of this would be, and based on that when it would and would not be permissible to create such AI.

One approach is to say that it would be better simply to avoid proliferating any sort of AI that could conceivably have moral standing. This would be possible now that we have sufficiently strict criteria for having potential moral standing. Without the tests we have established, there would be no obvious place to draw the line, no principled reason not to say that modern computers, capable of running programs that can pass limited Turing Tests, or even something as low-level as a smartphone, ought to be banned “just in case”. This of course would be an untenable state of affairs, significantly negatively affecting people’s lives. With our new criteria, only comparatively advanced AI would need to be avoided, not requiring us to roll back any of our current technological solutions, but only limiting what we do at some point in the future.

There are several different ways such an intuition could be morally justified. Possibly the most straightforward is appeal to standard anti-natalist arguments. These include the claim that while pleasure is good and pain is bad, the absence of pain is

good but the absence of pleasure is not bad, unless there is some individual being deprived of pleasure.¹⁵⁸ This leads to consequences such as: that we have a duty not to bring suffering individuals into existence but no commensurate duty to bring happy individuals into existence; that the fact that a being would suffer is a good reason to refrain from bringing it into existence but the fact that a being would be happy is not a good reason to bring it into existence; and that it is correct to regret creating an individual that suffers for the sake of that individual but it does not make sense to regret not creating a happy being for the sake of that being.¹⁵⁹ These considerations will weigh in favour of refraining from creating AI.

It is possible to resist the conclusion that these arguments determine that we should never bringing sentient beings into existence, but only provide more weight to the idea that we should avoid doing so when we cannot be certain that the balance of these beings' experience will not weigh too strongly in favour of suffering over happiness.¹⁶⁰ One could adopt an even stronger position and argue that as long as we can be sure that total well-being is increased, it would be right to bring more AI with the capacity for well-being into existence, or even that it might even be morally required, as it would be under some theories. There are concerns and debates about the morality of excessively creating even on-balance happy additional lives just as long as total well-being is increased, and there are increasingly calls to focus on "making people happy, not making happy people."¹⁶¹ Even if we ignore those,

¹⁵⁸ David Benatar, "Still Better Never to Have Been: A Reply to (More of) My Critics", *Ethics* 17 (2013) p. 122.

¹⁵⁹ *Ibid.*, p. 123.

¹⁶⁰ See Aaron Smuts, "To Be or Never to Have Been: Anti-Natalism and a Life Worth Living", *Ethical Theory and Moral Practice* 4 (2013).

¹⁶¹ Jan Narveson is generally regarded as the originator of this (see "Moral Problems of Population", *The Monist* 57[1] [1973] p. 80), which has been discussed a great deal since (see for instance Per

however, there is a distinct problem when it comes to AI, since it is uncertain whether we can ensure that AI have lives worth living. Even if we believe that an AI has phenomenal states that might be similar enough to our own to merit being granted moral standing, they would still have very different structure and composition to humans and non-human animals, and therefore likely very different needs and preferences. It would be extremely difficult to predict what it would take to ensure their flourishing and therefore uncertain whether it could be provided. Thus there would be a very real possibility that we would be bringing into existence entities whose lives are not worth living, which would be automatically wrong according to most moral theories and would likely come out wrong even on an unrestricted utilitarian theory. The fact that their level of suffering and pleasure will be very uncertain will disfavour creating AI more strongly than it does creating human beings.

Scarcity of resources and environmental concerns should give us pause in bringing any entity into existence that will make these problems worse. Of course creating any computer or technology will consume resources and have environmental impacts, but the situation is exacerbated when the entity we are creating will have legitimate moral demands for resources itself such that we morally cannot deprive it of them, and will be obligated to ensure that its needs are met and cannot simply deactivate or recycle it. A general policy of not creating AI with potential moral standing is not much different from the calls to spay and neuter pets, or to limit human population growth, and is justified on the same grounds.

Algender, "A Defence of the Asymmetry in Population Ethics", *Res Publica* 18[2] [2012], and Stuart Rachels, "Is It Good to Make Happy People?", *Bioethics* 12[2] [1998]).

Even if we can be certain AIs will have on balance good lives, there are arguments that could weigh in favour of not creating them that apply more strongly to AI than to human beings or non-human animals. Seana Shiffrin has provided a vivid example to illustrate how this can be the case: imagine a wealthy eccentric who elects to donate some of his wealth to a community of much less wealthy neighbours by means of throwing bricks of gold bullion from an airship.¹⁶² The people receiving the gold are sufficiently financially comfortable that they will not die or suffer severe harms without this extra wealth, but having it will benefit them greatly. Even for those who are injured by the falling bricks, the net effect on their lives will be positive (let us stipulate that there is no chance anyone will be killed or permanently debilitated in any way). Shiffrin argues that throwing the bricks was an immoral act, even if there was no safer way to confer this benefit and no way to secure consent in advance, because it caused harms without consent without averting significant harms. If one accepts that this is similar to the case of creating moral patient AIs with good lives, this would be an argument against creating any AI regardless of our prediction of how well their lives would go. Even if one does not accept this argument in the case of humans and disagrees that the case of human biological procreation is like this, there is better claim that this argument weighs against producing moral patient AIs. For AI, there is no recourse to claims such as that we merely “seek to [...] engage in a time-honored practice”¹⁶³ like when people decide to start a family. That could be argued to be a right that people have, and even an essential element of human

¹⁶² Seana Valentine Shiffrin, “Wrongful Life, Procreative Responsibility, and the Significance of Harm”, *Legal Theory* 5(2) (1999) p. 127. Shiffrin’s example uses a plane as the vehicle, but in addition to the greater practicality of using a slower—moving vehicle, this imagined character really feels more like a zeppelinist to me.

¹⁶³ *Ibid.*, p. 128.

flourishing, such that engaging in it averts greater harms that outweigh whatever badness there is in procreating.¹⁶⁴ There is no comparable deprivation of an essential element of human flourishing that is averted by creating AI, so that might not be justified even if procreation is justified.

Another reason human procreation might be justified is that even if it causes inevitable harms, it is necessary to avert greater harms, such as the extinction of the human race. It seems intuitive (though is not uncontroversial) that such an outcome would be bad, and there have been many theories proposed as to why.¹⁶⁵ Since no AI race exists, there can be no harm to one, and there is no harm from not bringing one into existence. We are under no obligation to create one any more than we are obligated to genetically engineer new species or to bring back the dinosaurs.

Shiffrin also countenances a weaker argument, that an action like the gold “donation” might not be morally wrong, all things considered, as long as the person performing it takes responsibility for compensating any harms suffered. This would mean that creators of AI with moral standing would bear responsibility for their well-being. This seems natural, but would mean that a significant burden is placed on people when such an AI is created, which affects the moral calculus. There is also a

¹⁶⁴ It is important to note, when it comes to discussions of the value of the parent/child relationship, that there is debate about whether raising one’s own biological child is a distinct good, or whether adoption is just as viable a means to the goods in question (see Tina Rulli, “The Unique Value of Adoption”, in *Family-Making: Contemporary Ethical Challenges* [2014] for discussion of these ideas). This is not particularly important, however, since it would still necessitate (and therefore justify) *someone* having children at some point, and the specific practical ethics of procreation in particular social contexts are not relevant to issues involving AI.

¹⁶⁵ See Johann Frick, “On the Survival of the Human Race”, *Canadian Journal of Philosophy* 47(2-3) (2017).

disanalogy with procreation here, because though parents take on significant burdens when having children, it is commonly believed that they are compensated with special goods. The role of parent is claimed to offer distinct benefits that cannot be achieved elsewhere that more than compensate for the significant burdens that come with it.¹⁶⁶ When it comes to AI there is no special relationship or role in our society that could confer those goods, and it cannot answer any natural need in the way a parenthood relationship could. Might our society develop new beneficial social roles around AI? It's possible, but does not seem like something we should count on. Therefore, creating excessive burdens with no appropriate compensation might be a reason to avoid creating AI with moral standing.

There is the further complication that we are not even certain that AI we make that pass our tests and fulfill our criteria will even have genuine moral standing. It would be sufficiently likely that we ought to act as if it is the case, but it would be by no means guaranteed. Thus even if we ensure that we do everything we can to meet the needs of AI, it may turn out that they had no moral standing in the first place, and thus any burdens that were placed on humans, and sacrifices we made to the well-being of humans or any other entity that genuinely has morally-relevant well-being, would turn out to be unjustified, since it was not accompanied by a greater increase in well-being for another entity, or indeed any increase in well-being, since the entity for whom it was performed had no actual level of well-being to affect. This is a problem unique to AI that we do not face when it comes to humans or other animals.

¹⁶⁶ See Sarah Hannan and R.J. Leland, "Childhood bads, parenting goods, and the right to procreate", *Critical Review of International Social and Political Philosophy* 21(3) (2018) for a discussion.

3. Considerations In Favour of Creating AI Moral Patients

The preceding arguments each could provide a pro tanto reason to not make any AI that meet our criteria for having potential moral standing. However, there can be other factors that override these reasons. If making AI that specifically pass our tests and thereby must be treated as having moral standing can produce sufficiently good results, the risks might be morally justified. This is a concern not often considered in the ethics literature: whether the positive effects of a being's existence on others in general can have an impact on the morality of bringing that entity into existence. This is importantly different from the case where the child's positive effect on the parents' lives was considered, because that is a reciprocal relationship that the child is an integral part of, and can be achieved no other way. Any positive effects AI have for people will be independent from any possible benefits people would provide to these AIs, and will be generic effects that could in principle be achieved in different ways with no necessary connection to AI (even if those other ways might not be feasible in practice). Can bringing entities into existence just to benefit others be justified? This is frequently not examined when it comes to population ethics, possibly on the basis of the (probably justified) belief that any additional people will on average be about as good at providing whatever goods people provide to others as the people who are already around, and that people's demands and outputs will tend to equal out and that there will be little to no inherent benefit to a larger population beyond a fairly low level. We can of course not guarantee that any additional humans will tend to have a positive effect and make things better for others, but the case is not the same when it comes to AI. Since AIs can have significantly different abilities and skills than humans, their presence could help society significantly if we

deliberately create AI that excel in specific areas where we have a deficiency of competence.

Let us assume, for argument's sake, that some combination of the aforementioned concerns about creating moral patient AIs provide sufficient prima facie reason to refrain from creating them. Now, what if AIs with those properties would make the lives of currently-existing people sufficiently better? Would that overcome the considerations against creating them? When presented with this question, it might be natural to recoil at the prospect of using sentient beings as a means to an end like this, possibly due to imagining a case like that described by William James,¹⁶⁷ later memorably referenced by Ursula LeGuin in the story "The Ones Who Walk Away From Omelas".¹⁶⁸ This is a situation where a utopia is achieved through the immense suffering of a single individual. This appears abhorrent to most people. As James says, "even though an impulse arose within us to clutch at the happiness so offered, how hideous a thing would be its enjoyment when deliberately accepted as the fruit of such a bargain?"¹⁶⁹ It might be thought the situation is little better when an underclass of a large number of individuals is subjected to mild-to-moderate suffering to maintain a higher standard of well-being for others. This does however more closely resemble many societies throughout history and up to the present day, which might be evidence it is more palatable to people, but the actual existence of such systems when there are no Omelas-style systems is due only to the impracticality of focusing the suffering so intensely, not because it is considered

¹⁶⁷ William James. "The Moral Philosopher and the Moral Life", in *The Will to Believe and other Essays in Popular Philosophy* (1919).

¹⁶⁸ In Ursula LeGuin, *The Wind's Twelve Quarters* (1975).

¹⁶⁹ William James. "The Moral Philosopher and the Moral Life", in *The Will to Believe and other Essays in Popular Philosophy* (1919), p. 188.

better to spread it out. It is clear that we should feel as much disapprobation towards the institution of slavery in America that enabled the functioning of that society, or to the existence of dalits under the Indian caste system, even if the systems included more overall happiness than had generations of those people never existed, even if most of those people had lives that were on balance worth living. But there are still some differences when it comes to AI. For one thing, the AI case need not be a case where we are knowingly and deliberately subject entities to suffering, we can create AI and do our best to attend to their moral standing, even if some of them wind up with lives not worth living (if they reliably prove to have lives not worth living, there might not be much difference if we continue to create them). Even if we have the intuition that it would have been better had African slaves brought to America not reproduced and died out after a generation, this might be influenced by relevant alternatives, such as the original enslaved people remaining in Africa and subsequent generations living there, or of the descendants of people brought to America not being subjected to horrific oppression and having better lives than they did under the system of slavery. If AI have an intrinsic nature such that a given generation of them cannot exist without having a level of well-being significantly below that of humans in society, and will increase human well-being significantly, this is relevantly different to a case where we could easily treat an underclass better and promote greater equality even at the cost of overall happiness. Thus even if we cannot be sure AI will be equally happy as humans, creating them might be justified.

This will depend heavily on specific applications of AI technology and whether passing the tests is crucial to success at these applications. It is clear that we should avoid frivolous uses of AI with social and moral competence. It might seem desirable

to have digital assistant programs (such as Siri, Alexa, and Cortana), for example, be as socially and morally adept as possible, but it is obvious that the benefits of this application do not outweigh the risks. A little inconvenience and frustration is a small price to pay to avoid the significant moral hazards that would result from potentially giving such AIs moral standing.

There are, however, cases where having an AI be as socially and morally competent as possible is a significant benefit, where there is more at stake than having an intuitive and pleasant user interface for some unimportant program. One case of this would be military applications. Under the assumption that governments will conduct military operations, maximally socially and morally competent robots might be able to minimise exposure to harm from these actions. For any automated military robots to be used in combat zones, especially those with lethal capabilities, having maximally sophisticated moral capabilities would be vital. However, the fact that an AI that was maximally adept at these duties would have a strong possibility of having moral standing removes a lot of the benefit of using such machines in the first place. A significant reason for having military robots is to keep human soldiers, who would otherwise be the ones to perform dangerous tasks, out of harm's way. If giving autonomous robots significant moral reasoning capabilities would also mean that they would have claim to moral standing themselves, we do not gain any significant advantage by placing them into harm's way instead, and thus expanding the purposes for which they are used into areas that require more moral competence might not be morally justified.

It might be thought that if we can create AI that is not only equal to humans in moral reasoning ability, but exceeds our own capacities, then we might achieve such superior results that it would justify the inherent moral risks. Indeed, if we could make an AI that we were sure would make significantly better moral decisions than the average human in situations where as much is at stake as the military situations in question, that would be a very significant moral benefit and could outweigh a lot of potential harms. However, it is very dubious that such a thing would be possible. For an AI to be significantly “better” than moral reasoning than humans would imply that its judgements frequently diverge from those of humans, and any case where that occurred it would likely not be recognised as morally superior but assumed to be morally deficient – for good reason, since that is the more likely explanation. This cannot be resolved by the judgement of “moral experts”, since the very possibility of such a thing is highly controversial at minimum.¹⁷⁰ It might be suggested that the problem is not that the people making moral decisions in combat situations are incapable of making the correct judgements, but that they make the wrong decisions in the moment based on inappropriate emotional responses and in a moment of calm reflection they and most other people would be able to see what was actually the right decision. If this is the case, it is possible that an AI could be better at this in a way we would recognise – but there is no reason to think that we would succeed in creating an AI that can do this. There is no evidence for the assumption that it would be feasible to make an AI that is capable of processing various emotions but is better able to weigh and balance and attend to the right ones for the right situations than the average human. That seems like an incredibly difficult task, and unless we have a better idea of how to accomplish it, it would be unjustified to produce AI that have

¹⁷⁰ See for instance A.J. Ayer, “The Analysis of Moral Judgments”, in *Philosophical Essays* (1954), and C.D. Broad, *Ethics and the History of Philosophy* (1953).

some claim to moral standing in order to experiment on how to make them better at situational moral judgement.

Another case that might justify the use of AI that are well-being subjects is medical applications. To automate more of the operations of a hospital, and for AI to take on more of the roles of doctors and nurses, they would need to be reliably capable of making correct moral decisions. Once again, this is a case where lives are at stake, and so taking on a good deal of moral risk could be justified by good enough outcomes. But once again, some of the reasons AI automation of some of these roles is thought to be desirable would disappear if the AI in question had a claim to moral standing. Doctors and nurses are frequently required to work long, gruelling hours, and having AI take over some of their duties would allow them to be freed from some of the stress and improving their health, emotional well-being, and opportunities for leisure and a social life. However, if the AI also have well-being that we are required to account for, we could not simply have these machines work round the clock and would have to ensure those things for the AI as well, therefore not making AI much different functionally as a solution than simply training and hiring more doctors and nurses. It might be thought that creating AI capable of performing the complex medical tasks and procedures as well as making the correct moral decisions would be faster than the many years of education required to train medical professionals. However, it is unlikely that there is any shortage of people who would be able and willing to perform jobs in the medical field given the opportunity, and if funding for medical education and hiring medical personnel were increased today it would only take a few years to see the effects, and as long as the pipeline was maintained the delays would not be felt thereafter. It would likely take much longer to design AI that

can perform all the duties of medical professionals, leaving the only possible benefit a faster response to elastic demand, something not likely to often come up to a great extent in the medical field. The only way the AI solution would be clearly justified is if designing, building, and maintaining these AIs was significantly cheaper in terms of resources than simply training and hiring more doctors and nurses. This is doubtful. Furthermore, if we can make AI that can perform the complicated technical aspects of medical procedures without giving them the ability to make moral decisions, we could have humans perform the moral decision-making with far less specialised training, allowing for hiring more of them for less demanding work, and allowing them to make the decisions while less fatigued and stressed, leading to better outcomes in a more certain and straightforward way without introducing the moral hazards and uncertainties of having potentially conscious AI.

Another possible use of socially and morally competent AI is monitoring and moderation of online communication, such as social media. Current attempts to use algorithms to identify and remove hate speech, threats, and other harmful speech have worked poorly, with many false positives and false negatives both.¹⁷¹ This speech is genuinely harmful and has many negative effects, and while it is not as invariably a matter of life and death as military and medical applications (though people's health and safety is genuinely threatened in many cases), its effects are more

¹⁷¹ For an illustration of this, see Joseph Cox and Jason Koebler, "Why Won't Twitter Treat White Supremacy Like ISIS? Because It Would Mean Banning Some Republican Politicians Too", *Motherboard Tech by Vice* (April 25 2019), https://www.vice.com/en_us/article/a3xgq5/why-wont-twitter-treat-white-supremacy-like-isis-because-it-would-mean-banning-some-republican-politicians-too In essence, Twitter employees have claimed that for an AI algorithm to succeed in banning the terrorist group ISIS from Twitter, it needed to be set in a way that it would also ban many innocent Muslims and Arabic-language media. The algorithm fails to ban many neo-Nazis because to increase its sensitivity any higher would cause it to ban Republican members of the U.S. House of Representatives and Senate.

widespread compared to the relatively small percentage of the world population directly affected by war. For this reason, the total amount of harm mitigated might merit taking on additional moral responsibility by creating AI with moral standing. This work also clearly requires a deep understanding of social nuance and the limits of what is moral or immoral, necessitating a very complex AI. In addition, the work itself would not be as gruelling or dangerous as military or medical work, and so having an AI with moral standing do the job would not be so inherently morally problematic. However, once again, there seems little reason to prefer AI performing these tasks to paying more humans to do it unless it was significantly more efficient and cost-effective. While using simple algorithms that can be set to run constantly and then forgotten about is much cheaper than employing humans to check manually, using AI that themselves need to be treated as subjects of moral concern would add extra costs that might not make it any more efficient.

A final note: it might be argued that, since we would have complete control over the programming of any AI we create, the concerns about being able to meet their needs and maintain a sufficient level of well-being for them are not such a big problem, since we can control their desires and dispositions and ensure that they are easy enough to meet. Even putting them in roles that are very bad for humans might not be a problem, since we can ensure that they do not have the same fears and aversions that make those jobs undesirable to begin with. First, it might not be so easy to control something so complex in a predictable way. We should not overestimate our ability to make subtle tweaks to these properties while maintaining full functionality. It also seems like it would be far easier to control the outward signs of what is contributing to the AI's well-being than to control the actual internal factors, and so

we might be merely deceiving ourselves into thinking the AI are doing well. But, more importantly, even if we can control such factors, it is not obvious that it is permissible to do so in any way we like. Imagine a situation where a group of people will be required to spend a lifetime doing filthy, unpleasant work cleaning up something that, left unchecked, would have a catastrophic environmental impact. While the contaminant is not directly harmful to them, they will have to live and work in filth and squalor their entire lives, and will inevitably be looked-down on by much of society. Would it be correct to genetically-engineer a race of people that enjoy living in squalor and being degraded, since the work is necessary anyway? It seems natural to feel that there is something wrong with that. There is no clear reason to think it should be any different with a morally-important AI than with organic beings. We are therefore left with important questions about what would be appropriate when it comes to how we program such AI.

We have established that there are very good reasons to avoid making AI that might possess moral standing, and very few if any contexts where it would be justified to do so. AI research should stop short of producing any AI that have analogues of desire-phenomenology and the sophisticated social and moral competences that would enable, at least until we can be far more sure of our abilities to deal with it in an ethical manner.