The Benevolent Ideal Observer Theory

By

Michael John Patrick Campbell

Department of Philosophy

Duke University

Date:_____

Approved:

_____

Walter Sinnott-Armstrong, Supervisor

_____

Gopal Sreenivasan

_____

David Wong

_____

Jennifer Hawkins

Dissertation submitted in partial
fulfillment of the requirements for the degree
of Doctor of Philosophy in the Department of
Philosophy in the Graduate School of
Duke University
2018

ABSTRACT

The Benevolent Ideal Observer Theory

By

Michael John Patrick Campbell

Department of Philosophy

Duke University

Date:_____

Approved:

_____

Walter Sinnott-Armstrong, Supervisor

_____

Gopal Sreenivasan

_____

David Wong

_____

Jennifer Hawkins

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Philosophy in the Graduate School of Duke University

2018

Abstract

This dissertation provides an answer to what I call *the central question* of moral philosophy: what, if anything, is moral value? The answer, I argue, is that moral value is the relational property of eliciting a suitable response from a properly informed, rational, benevolent and otherwise minimal spectator. I call this theory the response-dependent benevolent ideal observer theory or BIO$_{RD}$.

Although the way in which I express and argue for BIO$_{RD}$ is original and unique, the core of the theory is old. In chapter 1, I explore these historical roots. The notion that moral value depends, in some sense, upon the reactions of an idealised spectator stretches back at least to Adam Smith and, I argue, to his tutor Francis Hutcheson. I explore how a natural connection between ideal observers, benevolence and utilitarianism has often been assumed without being explicitly argued for.

In chapter 2, I lay out certain crucial meta-theoretical assumptions that help to motivate BIO$_{RD}$. I argue against the methodology of conceptual analysis and in favour of a revisionary approach sometimes called 'conceptual ethics'. I explore the theoretical aims that ought to guide the conceptual ethicists' project before arguing in favour of a response-dependent metaethics, in contrast to other sentimentalist theories such as fitting-attitude accounts. The response-dependent schema states that an object is morally valuable if and only if a particular agent would have a particular reaction to it (in certain circumstances).

In chapter 3, I argue that the agent that ought to fill the response-dependent schema is a properly informed, rational, benevolent and otherwise minimal spectator. I define benevolence as a final care directed towards the welfare of conscious creatures and thus argue in favour of welfarism: the view that welfare is the only essential moral value.

In chapter 4, I discuss which of the benevolent observer's reactions are best thought of as relevant, and which objects are best thought of as the bearers of both essential and non-essential moral value. I suggest that any attitude which lacks the property of being 'truth-oriented' is relevant. Arguments concerning the objects of value leads to a discussion of Parfit's Repugnant Conclusion, which BIO$_{RD}$ entails. I argue that we ought to embrace it. Lastly, I discuss the objects of non-essential value, showing how BIO$_{RD}$ can capture some intuitions that value pluralists might otherwise use against it.

Finally, in chapter 5, I discuss how accepting BIO$_{RD}$ would impact our moral discourse, arguing in favour of a radical eliminativist proposal in which deontic language is abandoned in favour of comparative value-talk. I then discuss how BIO$_{RD}$ can provide us with a theory of reasons.

Dedicated to the memory of Paul McClean,

and to Pete, Irene, Neil and Eve.

"…there is nothing either good or bad, but thinking makes it so." – *Hamlet*

"Benevolence is a world of itself – a world which mankind, as yet, have hardly begun to explore. We have, as it were, only skirted along its coasts for a few leagues, without penetrating the recesses, or gathering the riches of its vast interior." – *Horace Mann*

"If it makes you happy, it can't be that bad." – *Sheryl Crow*

Table of Contents

List of figures

All figures are in the public domain

A note on section numbers and historical references

Below is a list of historical texts referred to throughout this dissertation. Since there are numerous editions in use without common pages numbers, I have referred to them using book and section numbers and the abbreviated titles, given below.

Jeremy Bentham's 1789 *An Introduction to the Principles of Morals and Legislation* (*PML*)

David Hume's 1738 *A Treatise of Human Nature* (*THN*)

——— 1751 *An Enquiry Concerning the Principles of Morals* (*E*)

Francis Hutcheson's 1725 *Inquiry into the Original of our Ideas of Beauty and Virtue* (*Inquiry*)

——— 1728 *An Essay on the Nature and Conduct of the Passions and Affections, with Illustrations on the Moral Sense* (*Essay*)

——— 1755 *A System of Moral Philosophy* (*System*)

John Locke's 1690 *An Essay Concerning Human Understanding* (*AECHU*)

John Stuart Mill's 1863 *Utilitarianism* (*U*)

George Edward Moore's 1903 *Principia Ethica* (*P*)

Bertrand Russell's 1918 *Lectures on Logical Atomism* (*LLA*)

Henry Sidgwick's 1907 *The Methods of Ethics* (*Methods*)

Adam Smith's 1759 *The Theory of Moral Sentiments* (*TMS*)

Ludwig Wittgenstein's 1922 *Tractatus Logico-Philosophicus (Tractatus*)

Section numbers in this dissertation are organised as follows: the first digit indicates the chapter number in which the given section is located. Thus, every section number is unique, and the reader should immediately know what chapter consult if they wish to read a section referenced in some other part of the text. The second digit indicates the chapter section and, where present, the third digit indicates the sub-section.

Acknowledgements

Philosophy can be a lonely affair. Yet, if it weren't for the generosity of others, this dissertation would have never existed, or at least been far poorer than it is. My committee, Walter, Gopal, David and Jennie have all provided valuable input throughout the entire process, as well as offering personal support. My fellow graduate students at Duke have been a source of inspiration and encouragement. Particular thanks are due to Paul Henne, who provided extensive and insightful comments on significant portions of this dissertation, as well as for being a philosophical role-model and vital friend. I would have been lost at Duke were it not for the generous and constant support of Lisa, Janelle and Stacey. Needless to say, all flaws are entirely my own.

I might stop there, but given that this may well be my last as well as my first significant piece of philosophical writing, I hope it is appropriate to extend my thanks to all those who have had some impact on this work, even indirectly, by being vital to my philosophical development. Special thanks in this regard are owed to Karen Davies, Alex Oliver, Nick Crawford, Jacob Trefethen, Ele Gower, Fergus Blair, Wayne Norman, Aaron Ancell, Sara Copic, Christian Coons and Molly Gardner.

After I began writing this dissertation in earnest my oldest friend, Paul, died in tragic circumstances. He was a person of incredible character and ability. Since his death my primary aim has been to finish this work to the standard he would have held himself to

and expected of me. As such, I dedicate it to his memory, as well as to his family, Pete, Irene and Neil, and to his girlfriend, Eve.

Finally, I would like to thank my parents, Barry and Dorothy. I quite literally owe them everything. I hope I can repay them one day, in some small way.

Chapter 1: Introduction – A fresh look at an old theory

1.1. The central question

Many of us tend to think that at least some states, actions or character traits are worth wanting more than others. As such, we sometimes feel a kind of pressure to bring about good consequences, act generously towards others and cultivate decent personalities. This pressure is distinct from that which would promote our own self-interest. It seems to involve a concern with the lives of others, from friends to distant strangers. We sometimes say that the states, actions or character traits worth wanting in this distinctive sort of way are morally valuable. There is a rich vocabulary for talking about the ways in which moral value impacts our lives. We speak of reasons, duties, rights, right action, virtue and goodness to give a few such terms.

One core task of moral philosophy is to provide a more detailed characterisation of moral value. In particular, moral philosophy must answer what I will call *the central question*: what, if anything, is moral value?[1]

[1] This is not to say that there are no other vital questions in moral philosophy. One is epistemological: 'how can I know what moral value is, if at all?'. This epistemological question and the central question are intimately related, in that if there is no moral value, then one cannot know about it, and if moral value, assuming it exists, is the sort of thing we cannot interact with (such as a non-natural property), we might be led towards scepticism. I will largely pass-over epistemological questions such as this.

Broadly speaking, there are two types of answer to this question. The first tells us what moral value is without entailing anything about which things are morally valuable. Typical realist metaethical views are often of this kind. For example, if reasons are the fundamental moral normative unit, and reasons are non-natural properties (Parfit 2011a; Scanlon 1998) it remains an open, and perhaps insoluble, question as to how these properties are distributed over objects or states of affairs (Blackburn 1984, pp. 181-9). A second type of answer tells us what moral value is and which objects or states are morally valuable. This kind of answer is less common since answering the question 'which objects or states are valuable' has traditionally been the preserve of normative ethical theory, which often sets aside the central question, assuming that it has some positive and compatible answer. Let us call the question of what objects are morally valuable *the object question*.[2] Classical utilitarians tell us that states of affairs involving pleasure and only those states are morally valuable. Deontologists tell us that actions bear moral value in that they can be right or wrong. And virtue ethicists tell us that character traits have moral value, in that they can be worth cultivating. The central question, on the other hand, has traditionally been the preserve of metaethics, and normative ethics and metaethics are often conducted in isolation from one another. This

[2] Hereafter, I use the term 'object' as catch-all, designed to cover all possible bearers of moral value, including states of affairs, character traits, persons, actions, etc. I offer my views on which specific kinds of object bear moral value in chapter 4, but I leave the question open until it becomes necessary to address it.

dissertation takes a less-common two-pronged approach. I argue in favour of a metaethical thesis, response-dependence, and refine it in a way that also answers the object question. Ultimately, I will arrive at a non-standard form of utilitarianism and, as such, one might view this dissertation as an argument for (a version of) utilitarianism.

In the interest of transparency and for the sake of future reference, I state the core of my view below in its entirety, with definitions of technical terms provided. The details and arguments are to be given in later chapters.

My primary metaethical commitment is to the response-dependent theory of value. As I understand it, all response-dependent theories of value adopt the following general schema:

RD**:** *x* is morally valuable if and only if and because *x* elicits *R* from *S*.[3]

In chapter 2, I argue that the response-dependent theory of value is our best hope of reconciling the existence of value with the broader aims of moral theory, in particular, naturalism. Adopting RD naturally invites the question: who/what are *x*, *R* and *S*? That is, what are the objects of value, the relevant responses and the agent whose reactions make those objects valuable? I call my final position with respect to these questions the response-dependent benevolent ideal observer theory (BIO_{RD}):

[3] Sometimes circumstances or broader context are added to the schema.

BIO$_{RD}$: an object, *x*, is morally valuable if and only if and because it elicits a non-truth-oriented attitude, *R*, from a properly informed, rational, benevolent and otherwise minimal observer, *S*.

The definitions of technical terms to be discussed further in later chapters are as follows:

(i) An observer is properly informed if and only if they know all the relevant facts.

(ii) An observer is rational if and only if they are both instrumentally and formally rational.

(iii) An observer is benevolent if and only if they have an impartial final care directed towards the welfare of conscious creatures.

(iv) An observer is minimal if and only if they have the least complex psychology required to exhibit traits (i) – (iii).

(v) A truth-oriented attitude is any attitude which has a constitutive standard of correctness directed at truth and this is the only constitutive standard that attitude has. A non-truth oriented attitude is any attitude which is not truth-oriented.

I make the case for including the various attributes of the ideal observer in chapter 3, discussing the principles that ought to guide this process in chapter 2. In chapter 4, I argue that all non-truth-oriented attitudes are relevant and discuss the objects of value at

length. In chapter 5, I suggest that accepting BIO$_{RD}$ should incline us to adopt a radically revisionary, eliminative view of most moral discourse along broadly scalar utilitarian lines.

However, for the remainder of this chapter, I will be concerned with motivating my exploration and defence of BIO$_{RD}$ by way of tracing its curious history. In order to do this, it will be helpful to have some initial conception of benevolence. Thus, let us say, as a purposefully inclusive working definition, that benevolence is an impartial desire (or care or concern) for the welfare of others.

1.2. Historical precedents

With the possible exception of Francis Hutcheson (see sec. 1.2.2), this dissertation is, as far as I'm aware, the only detailed defence of an observer theory where said observer is explicitly characterised by benevolence. It is certainly the only contemporary defence. Yet some form of the benevolent ideal observer theory is often presented as if it were already widely known. This is perhaps due to the influence of Roderick's Firth's famous 1951 paper, *Ethical Absolutism and the Ideal Observer*. Firth tried to give a conceptual analysis of ethical statements in a way that captured the absolutist quality of moral language.[4] He argued that the reactions of an omniscient, omnipercipient, disinterested,

[4] By 'absolutist' Firth means 'not relativistic' and by 'relativistic' he means 'essentially contains egocentric terms whose meaning varies depending on the speaker.'

dispassionate, consistent and otherwise human observer could provide such an analysis. However, there is no mention of benevolence anywhere in Firth's theory. In fact, my observer and Firth's have no characteristics in common whatsoever (except, perhaps, for a certain kind of rationality). Firth does consider attributing something like benevolence to the ideal observer, but ultimately rejects the idea:

> "… my reason for believing that it is not necessary to attribute such virtues as love and compassion to an ideal observer, is not that it would be a logical mistake to do so,[5] but simply that I am not inclined to think that a man is necessarily a better moral judge, however superior as a person, merely because he possesses such virtues. The value of love and compassion to a judge, considered solely as a judge, seems to lie in the qualities of knowledge and disinterestedness which are so closely related to them; and these two qualities, as we have seen, can be independently attributed to an ideal observer." (Firth 1951, p. 341)

Despite his clear statement on the matter, philosophers have sometimes associated his dispassionate observer with a benevolent one. For instance, R.M. Hare writes:

> "The ideal observer theory (as I shall summarize it) holds that in considering what we ought to do, we have to conform our thought to what would be said by a person who had access to complete knowledge

[5] It is interesting to note that Firth claims that this would not be a logical mistake, given the subsequent criticisms of modified versions of his theory that did include these attributes as being viciously circular. For a detailed discussion, see section 3.3.3.

> of all the facts, was absolutely clear in his thinking, was impartial between all of the parties affected by the action, *and yet equally benevolent to them all*." (Hare 1972, p. 168 [italics mine])

Hare only and incorrectly cites Firth as an proponent of this view. Elsewhere, Richard Brandt considers Firth's original theory and, aware that Firth included no benevolence condition, a "second version [which] differs in the addition of 'benevolent' to [Firth's] qualifications" (Brandt 1979, p. 225). In a footnote, Brandt then cites Hume, Adam Smith, C.D. Broad, Jonathan Harrison, William Kneale and himself as examples of philosophers "who have occasionally supported something like this theory" where 'this theory' is ambiguous between Firth's and one with an added benevolence condition. However, none of these philosophers have defended the benevolence condition, at least in print. I will consider Hume and Smith in detail below, but it is worth briefly describing the other views Brandt cites.

C. D. Broad offers an extensive analysis and clear defence of what he calls the 'trans-subjective dispositional form of the moral sense theory'; the view "that such judgments as 'That act is right (or is wrong)' are […] some variant on the formula 'That act would evoke a moral pro-emotion (or anti-emotion) in any human being who might at any time contemplate it.'"(Broad 1944, p. 149) Broad's theory is distinguished from Firth's by its minimal idealisation (Broad does think that the human beings whose contemplation matters must be free from ignorance, not deluded, etc., which requires at least *some*

idealisation from the world as it stands). But yet again, benevolence plays no role in the specification of the agent(s) whose emotions are used to analyse moral statements.[6]

Kneale's only significant ethical work is his short article 'Objectivity in morals' (1950). In it, he defends objectivism against subjectivism by attempting to more accurately characterise the objectivist's position. He rejects Rossian idealism in favour of a more grounded notion of moral law, contrasting moral codes with legal ones. He argues that morality is distinguished from the law by (among other features) the fact that moral agents must, in some sense, endorse the moral law whilst they may reject the laws of their country and, secondly, that a consensus on moral law will be reached by reasonable people upon calm reflection (and that this will guarantee impartiality). In these two ideas, one can see hints of a constructivist metaethic, but Kneale is, very self-consciously, rejecting the sort of subjectivism with which ideal observer theories flirt. He never uses the term 'benevolence.'

Jonathan Harrison wrote a commentary on Firth original article soon after it was published (Harrison 1956). In it, Harrison raises fourteen concerns about Firth's theory, each of which he suggests need to be adequate resolved before one can fully endorse it (though he expresses sympathy with it). He also offers a number of interesting suggestions including, for instance, the claim that we ought to be concerned with ideal

[6] There is no further evidence of Broad endorsing a benevolent ideal observer theory in his more well-known work *Five Types of Ethical Theory* (1930).

reactions, as opposed to ideal observers. Harrison disagrees with Firth that the observer ought to be dispassionate, noting that sympathy stimulates passion in Adam Smith's ideal spectator, a view Harrison finds more plausible (see 1.2.2 below). Later, Harrison (1971) revisited ideal observer theories, but is ultimately less enamoured with them. His main objection is that it is not possible, without circularity, to include substantive character traits in the specification of the ideal observer (I deal with this objection at length in section 3.3.3). Nowhere does Harrison mention or consider a benevolence condition, but according to his later view, such a character trait would render the ideal observer theory viciously circular.

Finally, Brandt's view in *A Theory of the Good and the Right* is similar in kind to the ideal observer theory. Brandt argues that the good consists in what it would be rational to desire, after one undergoes an extensive process of "cognitive psychotherapy" (Brandt 1979, p. 131). Elsewhere, Brandt offers a sympathetic discussion of ideal observer theories, and notes that he defended a relativistic version of the theory, similar to Westermarck (1932), in an earlier exchange with Firth (Brandt 1959, p. 174). In this exchange, in addition to defending a kind of relativism, Brandt offers a proposal for capturing the meaning of 'relevantly informed' without circularity, thus eliminating the need for Firth's stronger omniscience condition (Brandt 1954a, 1954b) (considered in section 3.1.1). Yet only once does Brandt explicitly consider the possibility of a

benevolent observer.[7] Thus, it is not immediately clear where Brandt got the notion of a benevolent ideal observer from.

Brandt and Hare are not the only contemporary philosophers to posit such a counterfactual observer. The view is sometimes associated with utilitarianism. For instance, Rabinowicz and Österberg write that "the utilitarian attitude is embodied in an impartial benevolent spectator, who evaluates the situation objectively and from the 'outside'" (1996, p.3).[8] The connection was also drawn by Rawls in *A Theory of Justice* in a sub-section titled 'Classical Utilitarianism, Impartiality and Benevolence' (1971, pp. 160-66) where he wrote that classical utilitarianism "is closely related to the concept of the impartial sympathetic spectator" (1971, p. 161). The same is true of Parfit who discusses the relationship between utilitarianism and an impartial observer but does not explicitly consider a benevolent observer (Parfit 1984, p. 94). Yet elsewhere Parfit and others have often drawn a direct connection between benevolence and utilitarianism, albeit lacking any direct reference to a spectator or observer (see, for instance, Parfit's discussion of Sidgwick (1984, pp. 137-40)). In his famous exchange with Bernard Williams, J. J. C. Smart writes that the utilitarian "can appeal to the sentiment of generalized benevolence, which is surely present in any group with whom it is profitable to discuss ethical problems" (Smart and Williams 1973, p. 3). Smart later characterised generalized

[7] For another of his (brief) discussions of ideal observers, see Brandt (1955, pp. 108-9).

[8] See also Rabinowicz (1989, p. 31)

benevolence as a desire for the happiness or good of all sentient creatures, including non-human animals, still claiming that utilitarianism embodies this attitude (Smart 1980, p. 115).[9]

Perhaps the closest thing to a contemporary defence of the connection between benevolence, ideal observer theory and utilitarianism is John C. Harsanyi's Impartial Observer argument (Harsanyi 1977, 1986). Harsanyi's argumentative aim is a technical one. He attempted to justify the classical utilitarian's claim that the summation of individual utility is the overall goodness of a state of affairs. His Impartial Observer argument forms part of this broader argument and concerns the state an individual must be in when forming preferences; namely, that of being totally impartial and facing an equal probability of being any member of that society (that is, behind a Rawlsian veil).[10] But the notion of a rational observer behind a veil of ignorance is distinct from a benevolent one with full information.[11] Behind a veil of ignorance, a rational agent need only consider their own self-interest to arrive at a utilitarian set of preferences. Also

[9] See also Smart (1977, p. 128). One also finds the connection made in Rachels (1986, p. 101) and Hospers (1972, p. 149, 209) (cited in Nesbitt (1992)).

[10] Harsanyi described his theory as "a modern restatement of Adam Smith's theory of an impartially sympathetic observer" (Harsanyi 1977, p. 633) and saw Smith as equating "the moral point of view with that of an impartial but sympathetic spectator (or observer)" (p. 623). I examine Smith's theory below in detail, and argue that benevolence is not a part of Smith's characterisation of the observer.

[11] Although, in section 3.3.4, I argue that benevolence is characterised by impartiality.

important, however, is Harsanyi's aim. His goal was to provide a justification for classical utilitarianism using the notion of an impartial observer.[12] This dissertation attempts to examine the prospects for an ideal observer theory where the observer is characterised by benevolence and, in doing so, arrives at a non-standard form of utilitarianism. It is not part of my aim to justify classical utilitarianism.

1.2.1. Mill, Bentham and Sidgwick

However, there is at least one obvious candidate for the origin of a connection between benevolence, ideal observers and utilitarianism, due to J. S. Mill:

> "I must again repeat, what the assailants of utilitarianism seldom have the justice to acknowledge, that the happiness which forms the utilitarian standard of what is right in conduct, is not the agent's own happiness, but that of all concerned. As between his own happiness and that of others, *utilitarianism requires him to be as strictly impartial as a disinterested and benevolent spectator*." (*U*, chapter 2.) [italics mine]

But despite this remark, there is only one other use of the term 'benevolence' (or any of its cognates) in *Utilitarianism* and Mill does not elaborate on how or why a benevolent spectator would have utilitarian judgements. Although Mill drew a connection between utilitarianism and a benevolent spectator, we must assume that he did not think it

[12] For an argument that he was successful, see Greaves (2017).

important enough to require further elaboration. Interestingly, Bentham's *An Introduction to the Principles of Morals and Legislation* contains seventy-two uses of 'benevolence' and its cognates, but mostly in relation to the motive of benevolence as a virtue. None connect utilitarianism with the attitudes of a benevolent spectator.

Sidgwick, too, was alive to the connection between utilitarianism and benevolence:

> "Especially in modern times, since the revival of independent ethical speculation, there have always been thinkers who have maintained, in some form, the view that Benevolence is a supreme and architectonic virtue, comprehending and summing up all the others and fitted to regulate them and determine their proper limits and mutual relations […] The phase of this view most current at present would seem to be Utilitarianism." (*Methods*, Bk. 3, Ch. 4, Sec. 1)

But aside from remarking on the connection, he is not at any pains to draw it out at length. In book 4, chapter 4, section 1, Sidgwick does briefly discuss Adam Smith's ideal spectator, but this leads him to consider the role of sympathy in our moral imagination, not benevolence. Nonetheless, Sidgwick's famous claim that from the point of view of the universe there is no reason to promote one individual's good over another lends credence to a benevolent ideal observer theory. A benevolent observer, one might say, transforms Sidgwick's metaphorical point of view into a literal one, in which everyone's

welfare is taken into account equally. An ideal observer is one who takes the point of view of the universe.

1.2.2. Hume, Smith and Hutcheson

An idealised observer (or perspective) plays a crucial role in the work of both David Hume and Adam Smith. In his *Theory of Moral Sentiments*, Smith famously placed an idealised sympathetic spectator at the centre of his theory of moral judgement. But for Smith, sympathy is the exercise of an imaginative capacity, not an instance of benevolence:

> "By the imagination we place ourselves in his situation, we conceive ourselves enduring all the same torments, we enter, as it were, into his body, and become in some measure the same person with him, and thence form some idea of his sensations, and even feel something which, though weaker in degree, is not altogether unlike them." (*TMS*, Part I, Sec. I, Ch. 1.2)[13]

[13] Compare this to Firth's discussion of omnipercipience:

> "We sometimes disqualify ourselves as judges of certain ethical, questions on the ground that we cannot satisfactorily imagine or visualize some of the relevant facts, and in general we regard one person as a better moral judge than another if, other things being equal, the one is better able to imagine or visualize the relevant facts. Practical moralists have often maintained that lack of imagination is responsible for many crimes, and some have suggested that our failure to treat strangers like brothers is in large part a result of our inability to imagine the joys and sorrows of strangers as vividly as those of our siblings. These facts seem to indicate that the ideal observer must be characterized by extraordinary powers of imagination." (Firth, 1951, p. 335)

Sympathy is then defined in terms of imagination in the following way: when we exercise our imaginative capacity, we will sometimes come to share "fellow-feeling" with another, upon perceiving (and perhaps understanding the cause of) their emotional state. *The Theory of Moral Sentiments* is rich with examples, such as this one:

> "What are the pangs of a mother, when she hears the moanings of her infant that during the agony of disease cannot express what it feels? In her idea of what it suffers, she joins, to its real helplessness, her own consciousness of that helplessness, and her own terrors for the unknown consequences of its disorder; and out of all these, forms, for her own sorrow, the most complete image of misery and distress. (*TMS*. Sec.1 Ch. 1. 12)

Whenever we come to share fellow-feeling in this way, we are in a state of sympathy. Smith explicitly, and in a self-consciously revisionary way, defines sympathy as this kind of fellow-feeling:

> "Pity and compassion are words appropriated to signify our fellow feeling with the sorrow of others. Sympathy, though its meaning was, perhaps, originally the same, may now, however, without much impropriety, be made use of to denote our fellow-feeling with any passion whatever." (*TMS.,* Sec 1. Ch. 1.5)

The reason sympathy and imagination play such a crucial in Smith's theory is that they are used to explain our judgements concerning the propriety of actions. When an

impartial and sympathetic spectator would come to share fellow-feeling with another, then it is proper to approve of the original sentiment:

> "When the original passions of the person principally concerned are in perfect concord with the sympathetic emotions of the spectator, they necessarily appear to this last just and proper, and suitable to their objects; and, on the contrary, when, upon bringing the case home to himself, he finds that they do not coincide with what he feels, they necessarily appear to him unjust and improper, and unsuitable to the causes which excite them. To approve of the passions of another, therefore, as suitable to their objects, is the same thing as to observe that we entirely sympathize with them; and not to approve of them as such, is the same thing as to observe that we do not entirely sympathize with them." (*TMS*, Sec 1. Ch. 3.1)

Benevolence is simply absent from this picture. Of course, that is not to suggest that Smith says nothing of benevolence – far from it. Smith thinks, for example, that we will have "the strongest disposition to sympathize with the benevolent affections" and that 'the benevolent affections' are a class of the passions encompassing "generosity, humanity, kindness, compassion, mutual friendship and esteem" (Sec. 1, Ch. 4.1). The reason the disposition to sympathize with these traits is so strong is once again due to the force of the imaginative capacity. When we ourselves are the beneficiaries of such traits we cannot help but be pleased. And so, when one imagines oneself in another's

place, benefiting from the benevolent traits, an impartial and sympathetic spectator could only be pleased by their presence. The crucial point is that there is no need for Smith to define the observer as benevolent to begin with, since the observer's fondness for benevolence will follow from their imaginative capacity.

David Hume had similar views with respect to moral approbation. Hume held that sympathy gives rise to feelings of pleasure when we contemplate that which is useful or agreeable in others. However, Hume is careful to note that sympathy can be fickle. For instance, mere distance from ourselves can render the pain of another less unpleasurable than that of someone nearer to us. The vividness of our imagination can also impact our ability to fully sympathise with one another. Thus, Hume suggests that morality takes what he calls 'a general survey of the universe', where the defects of sympathy are corrected by encompassing the viewpoint of multiple individuals. The fact that morality takes this general survey allows us to prevent what Hume calls 'continual contradictions':

> "…every particular man has a peculiar position with regard to others; and it is impossible we coued ever converse together on any reasonable terms, were each of us to consider characters and persons, only as they appear from his peculiar point of view. In order, therefore, to prevent those continual contradictions, and arrive at a more stable judgment of things, we fix on some steady and general points of view; and always, in our thoughts, place

> ourselves in them, whatever may be our present situation." (*THN*, Bk 3, Part III, Sec. 1)

For Hume, taking this 'general point of view' allows us to co-ordinate our actions as a society to prevent disharmony and disunity. These contradictions are failures of societal co-ordination and communication that we all desire to avoid.[14]

[14] Sayre-McCord has noted that Hume's general point of view must be epistemically accessible, otherwise sympathy would entirely lose its appeal. For this reason, Sayre-McCord argues that Hume does not endorse an Ideal Observer theory:

> "Controlling for the variations, and correcting the perceptions, by settling instead on the Ideal Observer's point of view, the point of view of an unbiased, equi-sympathetic person responding with full knowledge to the actual effects on everyone of some particular person's character, will not do, either. Although stable, and presumably univocal in its deliverances, that point of view is not sufficiently accessible. We have neither the psychological equipment nor the knowledge required. Our estimates of the Ideal Observer's view of the effects of someone's character will differ in exactly the way our judgments of the actual effects differ. As a result, an Ideal Observer sets an inappropriate standard, not simply because we cannot take up her position ourselves (though we cannot), but because we cannot begin to anticipate what her reactions might be. Ignorant as we all inevitably are of the actual, subtle, and long-term effects of each person's character on everyone who might be affected, even earnest attempts by all to determine how an Ideal Observer would respond would leave us without a common standard around which to coordinate our actions and evaluations. No longer each speaking from her own peculiar point of view, each would still be speaking from her own peculiar take on a point of view she could not possibly occupy. And this means an Ideal Observer cannot play the role that needs to be filled. (Sayre-McCord 1994, p. 218)

However, Sayre-McCord appears to be taking a narrow view of what an Ideal Observer theory is. He is correct to insist that Hume does not require omniscience nor what Sayre-McCord calls "angelic equi-sympathetic engagement with all of humanity." (1994, p. 203). For sure, Hume's general point of view is not the point of view of a Firthian observer, but it is that of an idealized, if not fully ideal, spectator.

Though Hume is more cautious than Smith in accounting for failures of sympathy, both theories are alike in that the sympathetic spectator in Smith's case, and the general point of view in Hume's, are not characterised by benevolence. Benevolent motives are, for Hume, something that one who occupied the general point of view would consider virtuous.[15] They are not, however, constitutive of that point of view. Our moral sense stems from our sympathetic capacity once it takes the general survey. It is from this point of view that benevolence appears to us as a worthy and admirable trait. But neither Hume's general point of view, nor Smith's spectator, can be described as 'benevolent' to begin with.

Of the early moral sentimentalists, it is only Frances Hutcheson who places benevolence at the centre of his moral theory. Hutcheson, along with Shaftesbury, is one of the earliest philosophers to endorse a 'moral sense' theory. Moral sense theorists hold, roughly, that morality constitutes a distinct internal sense which causes us to take

[15] Virtues, for Hume, appear to be traits that are either useful to the beholder or others, or agreeable to the beholder or others:

> "Every quality of the mind is denominated virtuous, which gives pleasure by the mere survey; as every quality, which produces pain, is call'd vicious. This pleasure and this pain may arise from four different sources. For we reap a pleasure from the view of a character, which is naturally fitted to be useful to others, or to the person himself, or which is agreeable to others, or to the person himself." (*THN*, Bk. 3. Ch. II. Sec. 1)

It is not clear which of these criteria Hume takes benevolence to meet, but he clearly considers it a virtue (Radcliffe 2004).

pleasure in things that are good, and displeasure in things that are bad. A moral sense, for Hutcheson, explains why

> "…as soon as any Action is represented to us as flowing from Love, Humanity, Gratitude, Compassion, a Study of the good of others, and a Delight in their Happiness, altho it were in the most distant Part of the World, or in some past Age, we feel Joy within us, admire the lovely Action, and praise its Author. And on the contrary, every Action represented as flowing from Hatred, Delight in the Misery of others, or Ingratitude, raises Abhorrence and Aversion." (*Inquiry*, II. 1. ii.)

At the core of his 1725 work *Inquiry into the Original of our Ideas of Beauty and Virtue* is a division of the moral sense into the dual affections of love and hatred; so much so that Hutcheson claims that all other affections "seem but different Modifications of these two original Affections." Love is then divided into two further affections: "Love of Complacence or Esteem, and Love of Benevolence"[16] (*Inquiry*, II. 2. ii). Of course, by complacence, Hutcheson is not praising laziness or self-satisfaction. Complacence, for him, means something like 'that which is pleasing.' He also refers to this form of love as 'esteem' or 'good-liking.' More importantly for our purposes, is what Hutcheson says about 'Love of Benevolence.' Above, I defined benevolence as an impartial final care

[16] Hume appears to be mirroring Hutcheson's division when he claims that the virtues are divided into those which are pleasing and those which are useful (see fn. 12).

directed towards the welfare of others. In this respect, I am following Hutcheson's example when he claimed that benevolence was 'disinterested' or impartial, and that a benevolent act could not flow from self-interest or self-love:

> "If there be any Benevolence at all, it must be disinterested; for the most useful Action imaginable, loses all appearance of Benevolence, as soon as we discern that it only flowed from Self-Love or Interest." (*Inquiry*, II. 2. iii)

However, it is possible to overstate the importance of benevolence for Hutcheson. Its role is critical, but not fundamental. If anything takes this crown, it is love of complacence:

> "…so Benevolence seems to presuppose some small degree of Esteem, [and, in addition] Benevolence supposes a Being capable of Virtue. We judge of other rational Agents by our selves. The human Nature is a lovely Form; we are all conscious of some morally good Qualitys and Inclinations in our selves, how partial and imperfect soever they may be: we presume the same of every thing in human Form, nay almost of every living Creature: so that by this suppos'd remote Capacity of Virtue, there may be some small degree of Esteem along with our Benevolence, even when they incur our greatest Displeasure by their Conduct." (*Inquiry*, II. 2. iv)

Benevolence, therefore, presupposes our esteem of those to whom we act benevolently. [17] This passage is not entirely straightforward, since in one breath Hutcheson suggests that benevolence requires virtue and that it merely requires some esteem of a 'Form' (or forms) possessed by "every living creature." This presumably includes animals not capable of acting virtuously.[18] If, however, we adopt the more restrictive interpretation and read Hutcheson as saying that benevolence, or acting out of care for another, requires us to possess a certain degree of esteem for that individual, his view begins to look a great deal like Coons' Ideal Care theory, the closest contemporary moral theory to BIORD (Coons 2012).[19]

Aside from claiming that it cannot flow from self-interest, the *Inquiry* gives no explicit definition of benevolence. For this, we must turn to the less well-known and posthumously published *System of Moral Philosophy*. There, Hutcheson divides 'acts of will' into two classes:

> "… according as one is pursuing good for himself and repelling the contrary, or pursuing good for others and repelling evils which threaten them. The former we call *selfish*, the later *benevolent*. Whatever subtile debates have been to prove that all motions of the *will* spring from one fountain, no man can deny we often have a real internal undissembled

[17] Indeed, this section of the *Inquiry* is titled 'Benevolence presupposes Esteem.'

[18] Another possibility is that Hutcheson was considering angels, but it is simply not clear from his writing.

[19] I will have more to say about Coons' theory in section 4.3.

> desire of the welfare of others, and this in very different degrees." (*System*, Vol. 1, Bk. I, Ch. I, Sec. 5).

Slightly later in the same passage Hutcheson describes benevolence as a "determination alleged […] toward the universal happiness of others." (*System*, Vol. 1, Bk. I, Ch. I, Sec. 5). Once again, the conception of benevolence as a desire or care directed towards the welfare of others defended in later chapters comes to us directly from Hutcheson.

Like Smith, Hutcheson is deeply interested in the role of sympathy in exciting our passions. Unlike Smith, Hutcheson seems to assign to benevolence a primary role in the exercise of sympathy:

> "This sympathy seems to extend to all our affections and passions. They all seem naturally contagious. We not only sorrow with the distressed, and rejoice with the prosperous; but admiration, or surprise, discovered in one, raises a correspondent commotion of mind in all who behold him. Fear observed raises fear in the *observer* before he knows the cause, laughter moves to laugh, love begets love, and devout affections displayed dispose others to devotion. *One easily sees how directly subservient this sympathy is to that grand determination of the soul toward universal happiness* [i.e., benevolence]." (*System*, Vol. 1, Bk. I, Ch. 2, Sec. 3) [italics mine]

The claim that sympathy is subservient to benevolence suggests that the capacity for fellow-feeling exalted by Smith requires benevolence for Hutcheson. That is to say,

sympathetic engagement with others only leads to fellow-feeling when it is coupled with a determination towards universal happiness (i.e., benevolence). This is clearly not the case for Smith, whose picture is almost an inversion of Hutcheson's. One ought to be careful about reading too much into this, however, since Hutcheson does not say a great deal more about the way in which sympathy is subservient to benevolence. The point does not appear to have been of great concern.

The passage quoted above also contains a reference to an 'observer' with respect to the passion of fear. Yet references to observers or spectators in Hutcheson's *Inquiry* and *System* are few and far-between, and, when they do occur, are mostly given in passing, without great emphasis. The notable exception is Hutcheson's 1728 *Essay on the Nature and Conduct of the Passions and Affections with Illustrations on the Moral Sense* where references to observers and spectators and commonplace and substantial.[20] Mostly notably, Hutcheson uses the concept of a spectator to define obligation:

> "When we say one is obliged to an Action, we either mean, 1. That the Action is necessary to obtain Happiness to the Agent, or to avoid Misery: Or, 2. That every Spectator, or he himself upon Reflection, must approve

[20] See also Hutcheson's *A Short Introduction to Moral Philosophy*:

> "The proper arbiters are persons of wisdom, under no special attachment to either side, and who can gain nothing by the decision of the cause in favour of either party. Such men influenced by no interest or passion, tho' they be neither wiser nor better men than the parties contending, yet will more easily discern what is just and equitable." (Ch. XVII, Sec. 2)

> his Action, and disapprove his omitting it, if he considers fully all its Circumstances. The former Meaning of the Word Obligation presupposes selfish Affections, and the Senses of private Happiness: The latter Meaning includes the moral Sense." (*Essay, Ill. Ch. 1*)

Hutcheson seems to be expressing the duality between selfish and moral motivations that was, historically, the primary concern of most British Enlightenment philosophers and others well into the 20th century. Obligation, he suggests, can be understood in a self-regarding sense, and in a moral one. The moral one, being of most interest to us, involves the approval of a spectator upon reflection. If the clause stating that this meaning 'includes the moral sense' is taken at face value, what Hutcheson seems to be saying is that one is obliged to perform an action if a spectator, possessed of functioning moral sense, would approve of it (and disapprove of its omission).[21] And, as we have seen, the moral sense is composed of two distinct forms of love: love of esteem (or complacence) and love of benevolence. Thus, insofar as benevolence is required for the moral sense, Hutcheson defines obligation in relation to the reactions of a benevolent spectator.

This may be the beginning of the association between a benevolent spectator and utilitarianism, since Hutcheson was arguably the first philosopher to openly endorse a

[21] It is also interesting to note that Hutcheson uses the observer define 'obligation', a concept which does not seem to concern him in the *Inquiry*.

form of utilitarianism, famously writing in the *Inquiry* that "that Action is best, which procures the greatest Happiness for the greatest Numbers; and that, worst, which, in like manner, occasions Misery" (*Inquiry*, II. 3. viii).[22] The link is an obvious one if, as I have claimed, a chief component of the moral sense is an impartial desire for the happiness[23] of others on behalf of a spectator whose approval and disapproval determines obligation.

Thus, despite never using the phrase 'benevolent spectator' or 'benevolent observer', Hutcheson appears to have endorsed a theory very similar to BIO_{RD} and can, I think, rightly be regarded as the originator the elusive benevolent ideal observer theory. Of course, Hutcheson does not explicitly develop his moral theory in these terms, and his interests are somewhat distinct from those of contemporary philosophical theory. There is also no real discussion of the idealisation required on behalf of the spectator. The position he staked out in the early 18th century certainly merits revisiting.

[22] It appears, however, that Hutcheson did not coin the phrase 'the greatest happiness for the greatest number.' That honour appears to belong to Leibniz (Hruschka 1991). Hutcheson's utilitarianism is particularly interesting since he does not endorse the principle of utility as the standard of right action, but the degree of virtue. This makes his older version of the theory most similar to Norcross' scalar utilitarianism, which I discuss in sectionn 5.2.

[23] And it's clear that, with perhaps one qualification, Hutcheson means aggregate happiness. See *Inquiry*, II. 3. xi.

1.3. A roadmap

In what follows, I defend the view originated by Hutcheson with the benefit of almost three hundred years of philosophical insight. Ultimately, I hope that my defence provides a satisfactory answer to the central question, as described at the start of this chapter. However, I remain sceptical that any entirely satisfactory answer to the central question can ever be given. There will always remain disagreement among reasonable people about certain crucial cases which determine the important fault-lines in one's moral theory. Thus, at best, I hope that what follows achieves at least two aims.

The first is that it articulates what those who are already sympathetic to some of the ideas presented here have found themselves unable to articulate before. I cannot help but be reminded of the following remark of Wittgenstein's in his preface to the *Tractatus*:

> "Perhaps this book will be understood only by someone who has himself already had the thoughts that are expressed in it – or at least similar thoughts. – So it is not a textbook. – Its purpose would be achieved if it gave pleasure to one person who read it and understood it." (*Tractatus,* Preface.)

Naturally, this dissertation is nothing like the *Tractatus*, in subject matter or philosophical insight. Neither do I believe that it can only be *understood* by those who have already had the ideas contained in it. But I believe that part of what Wittgenstein

hoped for was for readers to think "Ah ha! I suspected that all along but could never quite say it." I have the same ambition.

The second aim I hope to achieve is to honestly locate the source of disagreement between myself and possible interlocutors. I believe that an unfortunate amount of philosophical labour is spent talking past one's opponents. I have tried to be forthright about the benevolent ideal observer theory's strengths, but, more importantly, its shortcomings. Quite honestly, although I sometimes wake up fairly convinced it is true, I often do not. I am, however, quite sure that is a powerful theory, of great interest historically, and potentially helpful to contemporary ethicists. If, in articulating it, I can pin down where some of the disagreement between it and its most compelling rivals lies, I will consider myself to have made progress.

Thus, as an example of this approach, I believe it will be helpful to state some of my metaethical presuppositions upfront. It is possible to endorse an ideal observer theory from mutually incompatible metaethical starting points. Yet, given that I rely on certain metaethical and meta-philosophical assumptions in arguing for a benevolent ideal observer theory, I hope it will help the reader if I am up-front about them from the beginning, and can offer some plausible reasons as to why they might be true. This is the subject of the next chapter.

Chapter 2: The aims of moral theory and response-dependence

In the previous chapter, I traced the historical roots of the benevolent ideal observer theory. Despite a significant historical pedigree, it lacks a robust contemporary defence. In this chapter, I begin constructing that defence. My aims are twofold: (i) to make clear the general principles guiding the construction of BIO$_{RD}$ (and moral theories more generally); and (ii) to argue that the response-dependent theory is the most attractive metaethical position with respect to moral value.

2.1. The central question

Let us return to what, in section 1.1, I called *the central question* of moral philosophy: what, if anything, is moral value? How should one go about answering it?

We might begin with a conceptual analysis of MORAL VALUE.[1] Conceptual analysis is the attempt to describe some concept as *ordinary* and *competent* users of that concept employ it, most often revealed through their language (and often this is limited to the English language). Conceptual analysis has been an important philosophical tool since philosophy's beginnings. What was Socrates doing if not probing the conceptual scheme of his Athenian contemporaries to try and understand JUSTICE, COURAGE, VIRTUE and the

[1] Here and throughout, I adopt the following convention:. 'MORAL VALUE' designates the concept moral value, wheras '<moral value>' designates the property moral value.

like? Importantly, starting with an ordinary concept does not preclude one from encountering puzzles and complexity either. Bertrand Russell, in his lectures on logical atomism, described his method of analysis thus:

> "I am trying as far as possible […] to start with perfectly plain truisms. My desire and wish is that the things I start with should be so obvious that you wonder why I spend my time stating them. This is what I aim at, because the point of philosophy is to start with something so simple as not to seem worth stating, and to end with something so paradoxical that no one will believe it." (*LLA*, lecture 1)

And contemporary philosophy is rich with examples too. One only needs to consider Edmund Gettier's (1963) paper '*Is Justified True Belief Knowledge'* and the literature that arose in its wake (Shope 1983). What was Gettier doing if not providing counterexamples to a conceptual analysis of the ordinary concept KNOWLEDGE?

Although conceptual analysis has been historically (and continues to be) an influential research paradigm, it has serious flaws, particularly in the case of ethical theory. It is simply not clear that there is something we can legitimately call 'the ordinary concept MORAL VALUE as used by competent speakers of English.' There is wide-spread disagreement about what our ordinary moral concepts are. Some consider RIGHTNESS and WRONGNESS to be fundamentally relative or subjective, while others do not. Some consider these concepts to make no sense in the absence of a divine law-giver,

while others do not. It is not clear how we are to define 'ordinary' in light of such disagreement. Is the ordinary concept the one that over half the population hold? What if there is no such concept? How about the one that largest number of members of the population hold? What if the concept that most of us have turns out to be held by only a small overall proportion of the population? It is not clear how to answer these questions.

Scepticism about ordinary concepts should not be taken as total. Clearly, we are often capable of communicating about value with other members of our linguistic community without talking past one another, and this wouldn't be possible if our concepts had nothing in common.[2] Even if our concepts are not identical, they may be able to tolerate some degree of difference. It's unlikely that a devout Catholic and a cultural

[2] Interestingly, in his aforementioned lectures, Russell thought otherwise:

> "The whole question of the meaning of words is very full of complexities and ambiguities in ordinary language. When one person uses a word, he does not mean by it the same thing as another person means by it. I have often heard it said that that is a misfortune. That is a mistake. It would be absolutely fatal if people meant the same things by their words. It would make all intercourse impossible, and language the most hopeless and useless thing imaginable, because the meaning you attach to your words must depend on the nature of the objects you are acquainted with, and since different people are acquainted with different objects, they would not be able to talk to each other unless they attached quite different meanings to their words." (*LLA*, lec. 2)

This strange conclusion is the result of a combination of Russell's knowledge by acquaintance/description distinciton and the theory of definite descriptions, neither of which are important for my purposes.

anthropologist would have identical moral concepts, yet they still seem to be capable of substantive disagreement at least some of the time.[3]

Still, it is an open question how much difference is tolerable in order for two concepts to count as 'similar enough' and whether, in light of the fact that there may be a great deal of difference, it continues to make sense to call one concept 'ordinary' and the other 'non-ordinary' or 'deviant'.

But supposing that one had satisfactory answers to these questions, it remains an open question as to how we could go about discovering what this ordinary concept is. Perhaps if one knew that one's own concept was ordinary, one could simply reflect on the nature of that concept. But how could one come to know that one's own concept was ordinary in the relevant sense? One would have to establish something about how other competent speakers conceptualise moral value in order to establish what was ordinary *before* one could recognise that one's own concept mirrored that conception. Just because we can sometimes engage in an activity which looks and feels a lot like disagreeing with other seemingly competent English speakers does not guarantee that our concept is 'ordinary'. And since we do not have any privileged access to the mental states or ideas of others, one seems forced to investigate the matter through some empirical means. A

[3] One might appeal to the concepts/conception distinction here (Gallie 1955; Hart 1961; Rawls 1971; Dworkin 1986). According to this distinction, the Catholic and the anthropologist share the same concept, but have competeing conceptions. Concepts are individuated by the functional rule they play whereas conceptions are individauted by what the concept is taken to apply to.

rudimentary suggestion is that we conduct a survey asking people to "describe the concept: MORAL VALUE." Though certainty possible, and perhaps fruitful, it neglects to consider the fact that people sometimes use concepts without being able to fully articulate them. As I understand the project of experimental philosophy, its aim is to probe our ordinary concepts, insofar as they exist, by asking us to apply them in certain complex cases designed to tease apart any latent ambiguities. This seems like a worthy project, if one's goal is to figure out what our ordinary concepts are. And, insofar as philosophers rely on ordinary concepts in their theorizing, they should pay deference to experimental results.[4]

Since this is not a work in experimental philosophy the reader may infer that I do not think that the central question is best answered *via* conceptual analysis. Suppose that, after the data is in, it's clear that ordinary speakers tend towards using a unified single concept MORAL VALUE. What reason would we have to care about this concept? The fact that ordinary speakers reliably converge on a single concept (still an open question) does not entail that this concept is coherent, consistent or in any way suited to the functions

[4] The counter-argument often given is that philosophers bring a kind of expertise to conceptual analysis lacking in ordinary folk (Williamson 2007). Even if this is true, it does not justify the total absence of empirical results in much contemporary conceptual analysis, understood as an attempt to clarify an *ordinary* concept. The folk may be confused and lack the relevant expertise, but surely, we require some empirical information as to what ordinary speakers mean by a particular term before we can apply our philosophical expertise (assuming there is such a thing)?

we would want our moral concepts to fulfil. The notion of conceptual function is particularly important, as it forms the starting point for a rival approach to answering the central question.

Instead of conceptual analysis, I suggest that we ought to investigate MORAL VALUE *via* what has come to be known as 'conceptual ethics'. Although this rival approach is fairly common philosophical practice, it has only recently been explicitly conceived of as a distinct research project (Burgess and Plunkett 2013a, 2013b). Broadly speaking, conceptual ethics is the *prescriptive* project of determining what concepts we *ought* to use, instead of *describing* those that we *in fact* use. The project itself is not systematic in that it does not recommend that we cease conceptual analysis altogether in favour of conceptual ethics. But it does suggest that, where appropriate, we ought not merely consider what our concepts are, but what they ought to be. Conceptual ethicists may disagree about the sense in which we 'ought' to revise our concepts. I take no stance on concepts in general, but with respect to the moral domain, I suggest that our concepts ought to be revised so as to best fulfil the *functions* we want those concepts to fulfil.

This characterisation of the distinction between conceptual analysis and conceptual ethics is not universally accepted. Frank Jackson (1998) offers an extensive defence of conceptual analysis and suggests that it does, in fact, have a prescriptive element.

> "… conceptual analysis has a prescriptive dimension. We may decide that, say, "free action" as used by the folk embodies some kind of confusion or perhaps an out-dated metaphysics. In such a case, we may in part prescribe the division the term effects. I think this is how we should view compatibilist analyses of free will. Our ordinary notion of free will is, at best, nowhere instantiated and, at worst, confused and impossible of instantiation. But something along the lines sketched by compatibilists serves the worthwhile purposes of our defective folk notion." (Jackson 2001, p. 618)

As an advocate of conceptual ethics, I do not wish to begrudge Jackson his use of the term 'conceptual analysis'. Yet, I think it is helpful to distinguish between philosophical work that takes its primary task to be understanding concepts as they are in fact employed by competent users, and philosophical work which does not. Jackson's view is located somewhere in the middle. He does not take conceptual analysis to be disregarding or ignoring ordinary concepts, yet he permits some degree of revision. In fact, this middle-position is useful since it allows us to distinguish what we might call *moderate* conceptual ethics from *radical* conceptual ethics. Consider the following question: how should the post-revision concept relate back to the pre-revision, 'ordinary' concept? One answer is that all competent speakers should be able to recognise the post-revision concept as similar enough to the pre-revision concept and perhaps even be able, without great difficulty, to adjust their everyday talk in order to

accommodate this change. I take it that this is roughly the position of Jackson, and, by stipulation, it is also the position of the moderate conceptual ethicist. The radical conceptual ethicist, on the other hand, has no regard for how ordinary folk view the post-revision concept. They may fail to recognise it and fail to accommodate the prescribed alteration in everyday discourse. The radical conceptual ethicist simply does not care, since he holds that the new concept is interesting enough on its own to merit our attention, regardless of its relationship to the prevailing conceptual scheme. They claim that the mere fact that ordinary speakers may lack a word or concept for some idea does not mean that this idea is not interesting or worthy of consideration.[5] However, the radical conceptual ethicist must be careful. If *no one at all* thought it reasonable to call the revised concept an instance of the pre-revision concept, even a strange one, they would seem to have changed the subject, rather than to have revised a familiar concept. As David Lewis wrote:

> "In trying to improve the unity and economy of our total theory by providing resources that will afford analyses . . . I am trying to accomplish two things that somewhat conflict. I am trying to *improve* that theory, that is to change it. But I am trying to improve *that* theory, that is

[5] One might again appeal to the concepts/conceptions distinciton here to help illuminate matters (see fn. 4 of this chapter). The conceptual ethicist's project can be seen as improvoing upon one conception of a concept or as an expression of scepticism that there is anyting like an 'ordinary' conception.

to leave it recognisably the same theory we had before." (Lewis 1986, p. 134)[6]

One can view the moderate and the radical as at two ends of a spectrum. The archetypal moderate only cares about improving the concept if all speakers could recognise the post-revision concept and adapt their discourse in-line with this revision. The archetypal radical does not care if *anyone* recognises the post-revision concept or changes their discourse. It seems to me that the radical has the edge, if they can show that their concept fulfils certain important theoretical functions (more on this function later in this section). Yet, I happen to think that the theory on offer here, though at the radical end, still arrives at a concept of MORAL VALUE that is recognisable by most people, even if they would reject it or refuse to adapt their discourse in line with it.

Though the reader may be uncomfortable with this approach, I ask them to suspend judgement, at least temporarily. I remind them that I am not alone in requesting a

[6] Lewis is talking about theories rather than concepts, but the point stands. He continued, expressing moderate, even conservative, tendencies:

> "For it is pointless to build a theory, however nicely systematised it might be, that it would be unreasonable to believe. And a theory cannot earn its credence just by its unity and economy. What credence it cannot earn, it must inherit. It is far beyond our power to weave a brand new fabric of adequate theory *ex nihilo*, so we must perforce conserve the onen we've got. A worthwhile theory must be credible, and a credible theory must be conservative. It cannot gain, and it cannot deserve, credence if it disagrees with too much of what we thought before. And much of what we thought before was just common sense. Common sense is a settled body of theory – unsystematic folk theory – which at any rate we *do* believe; and I presume that we are reasonable to believe it. (*Most* of it.)"

temporary suspension of disbelief. To take one example from another philosophical discipline, there has been a resurgence in epistemology in the 21st century thanks to Williamson's *Knowledge and its Limits* (2000) and the 'knowledge-first' approach. Williamson famously argues that knowledge is the broadest factive mental state and that all other epistemological notions should be analysed in term of knowledge, not vice-versa. This project is best viewed as an example of conceptual ethics.[7] In it, Williamson asks the reader the judge the project by its fruits (2000, p. 2). He notes that epistemological projects are not subject to proof or disproof. Rather, it is a matter of judging a theory in its entirety, once it has been laid out clearly for the reader to inspect. I make the same offer here.

Before moving on, it is worth mentioning two prominent examples of work in moral philosophy that make their commitment to conceptual ethics clear (although, neither uses the term). The first is Brandt (1979, pp. 1-24). Brandt argues against what he calls 'the method of linguistic intuitions' in which

[7] One may reject the characterisation of Williamson (2000) as an example conceptual ethics, if one thinks that Williamson has simply been successful at describing what was, all along, the ordinary concept (Cappelen 2018). If one thinks that Williamson's concept of knowledge is not ordinary, then one will view the work as an example of conceptual ethics. If all Williamson achieved was describing the ordinary concept, then it rather speaks against the idea (ironically, put-forward by Williamson in his (2007)), that philosophers possess a particular expertise in conceptual analysis, given their utter failure to recognise such a basic conceptual fact for so long.

> "we identify the questions of normative ethics by finding which questions people actually raise when they pose the traditional moral questions. It is thought that if we note sufficiently carefully what people (including ourselves) are doing when they raise these questions, we shall be able to paraphrase them in a perspicuous way. Then […] having got the 'logic' of normative concepts right, we can see what kind of argument is a reasonable argument in normative discourse." (Brandt 1979, pp. 3-5)

As I understand it, this is a variety of conceptual analysis. Brandt offers similar reasons for rejecting this approach as those offered above, namely: (i) normative words are vague so as to often yield no definite results, and; (ii) even if they did, it is not clear why we should rely on these results in reflection. Suppose that we develop a sophisticated ethical theory, yet the concept MORAL VALUE yielded by the theory differs from the ordinary one. Brandt summarises his critique pithily by asking "[w]here is the sting of being denied the use of a certain English expression?" (1979, p. 9).

Brandt offers in place of the method of linguistic intuitions what he calls 'the method of reforming definitions' according to which it's not what moral concepts actually mean that matters. Instead, "the serious question is what they might helpfully be employed to mean" (1979, p. 193).

Another instance of conceptual ethics in moral philosophy comes from Railton, who explicitly endorses Brandt's method of reforming definitions:

> "Since almost any notion […] found in natural language will draw its meaning from multiple sources, and will have taken on diverse functions […] at various points in its evolution, it is to be expected that any rendering of that concept intended to make it sufficiently clear to suit the purposes of theory-construction will be, to some degree, revisionist. So it was with water and H_2O, and so it will be with any significant evaluative concept."
> (Railton 1989, p. 158)

Here, Railton is explicit about the importance of paying attention to the function of a concept in constructing a theory of that concept. The notion of a conceptual function is important because it allows us to anchor the radical conceptual ethicist's project to something concrete. Since the radical does not care whether ordinary folk recognize his concept, he may only persuade others to adopt his new concept if there is some other reason he can give to persuade us that his new concept is still worthy of concern. Functions allow the radical to do just this, since he can tailor his concept to fulfil a particular function which we all (or enough of us) care about. So long as we all care about there being some concept that is able to fulfil an important function, then who cares if ordinary folk are initially inclined to reject it? Once again, where is the sting in being denied the use of a particular word? The radical conceptual ethicist answers: there is none.

2.2. Theoretical aims

Suppose, then, that the way to answer the central question is by revising the concept MORAL VALUE so as to better to fulfil its function. We must answer two questions: (i) can we find some concept to use as a starting point for our inquiry? and (ii) what is the function of MORAL VALUE?

With respect to (i), the radical conceptual ethicist is simply entitled to start with their own concept. Since he is not much motivated by how widely his concept is or could be shared, there is no obligation to investigate how similar his own concept is to others. Though, as a competent member of a linguistic community in which the concept is employed, his concept is unlikely to be radically different from those of other speakers (yet it wouldn't matter if it was).[8]

Still, given that I will start with my own concept, it is worth briefly stating three features of that concept that, if at all possible, I wish to retain. The first feature is *cognitivism*. It seems to me that ethical judgements express beliefs about certain features

[8] What ends up as a theoretical constraint and what ends up as simply being part of the starting concept is determined by what one is willing to sacrifice along the way. Theoretical aims should only be sacrificed as a last resort. Concepts, however, can be more easily altered. For instance, the theory I present is cognitivist in kind. Cognitivism is part of the nature of my concept of moral value, and the same is true for many others. However, I have not included cognitivism as a theoretical aim because I would be willing to give it up in favour of expressivism, say, if it was necessary to capture the relevant theoretical aims. Thus, there is no sharp boundary between conceptual starting-points and theoretical aims. It is a matter of a degree and, for the conceptual ethicist, unproblematically personal.

of the world, as opposed to being expressions of emotions or desires (Blackburn 1998), or endorsements of norms (Gibbard 1990). The second feature is *realism*. This is the view that ethical statements are sometimes substantively true and moral discourse is not in any systematic error. The third feature is *absolutism*. As I will understand it, absolutism is the view that ethical statements can be true for everyone, regardless of the community one inhabits. There are some ethical statements which are not true *for* one community and false *for* another. We will see that this doesn't prevent there being some statements that do have this feature, but the most fundamental ethical truths are absolutist.[9] Although MORAL VALUE, as I understand it, has these features, they are not entirely egocentric. The number of moral theories committed to each of these features are too numerous to count. Although the conceptual ethicist's project has the potential to be a lonely one, I don't believe this to be the case here.

The far more important question is (ii). Fulfilling a particular function is what the radical conceptual ethicist aims to do in revising his concept and the only basis he has for persuading us to adopt it after it has been revised. It is also a far more difficult question to answer. Let us call it *the functional question.*

One way to answer the functional question is by offering some account of the evolutionary role of moral concepts. On this view, we might discover that the concept

[9] It is worth stating that a closely related thesis is that of *ethical necessity*. The most fundamental ethical statements express necessary truths. This is also a feature of my concept **MORAL VALUE**.

MORAL VALUE survived because those who believed in such a thing were better adapted to their environments and thus tended to survive and reproduce. We often observe extreme forms of altruism in other species and it is not implausible to suggest that, in mammals capable of language and reflection, a concept like MORAL VALUE might arise so as to make sense of our evolutionarily acquired disposition towards altruistic behaviour.

This is the approach taken by Philip Kitcher in *The Ethical Project* (2011). Kitcher attempts to "reconcile ethics with a Darwinian picture of life" (p. 411) by stressing the importance of the evolutionary function of morality:

> "… I propose that socially embedded normative guidance [i.e., morality] is a social technology responding to the problem background confronting our first full human ancestors. […] Moved by a sense of the fragilities and tensions of their social life, they first guided their behavior by regularities to help them avoid trouble and later discussed with one another rules to govern conduct, to be applied in increasingly explicit systems of punishment. […] Ethical codes serve the function of solving the original difficulties, dimly understood by these ancestors." (Kitcher 2011, p. 221)

Kitcher argues that the problem background morality has evolved to overcome are unstable social structures. Thus, "the original function of ethics" (p. 8) is to remedy altruism failures, since failures of altruism in a fragile social community can have high costs for everyone. Importantly, this is still the primary function of morality, although

'the ethical project' consists in revising morality to better fulfil this function (and some additional subsidiary functions) in new and more complex environments.[10]

Kitcher anticipates that the primary criticism of his view is its rejection of the centrality of truth to moral theory. Moral progress is not, for Kitcher, a matter of accumulating more truths, but refining the 'social technology' of morality in light of its functional aims. The question of truth is important, and I will consider it below, but there is a more fundamental reason to question Kitcher's approach than its rejection of truth.

Kitcher's concern for failures of altruism is the result of his emphasis on Darwinian natural selection. But what if the reason we care about morality has nothing to do with its survival value? Kitcher may be entirely correct when he says that morality is an evolved technology designed to prevent failures of altruism. This may be its 'original function'. But what's to stop us abandoning our commitment to this function and divorcing morality from its Darwinian origins altogether? Although we are evolved creatures, and our evolutionary past has surely constrained and influenced our actual values in a host of interesting and important ways, evolution also provided us with a

[10] We can accept Kitcher's claims about the evolutionary function of morality for the sake of argument, but in reality, it's very difficult if not impossible to thoroughly test such claims. Why not think that morality is an evolutionary spandrel, some additional side-effect created by self-conscious creatures that has no additional survival value (or disvalue)? I do not say that this is true, it's simply difficult to know how one could show that it's false (as Kitcher needs it to be).

mind capable of reflecting on the merits of the functions of practices that, in some way, shaped that mind. Kitcher may not be wrong in saying that the evolutionary function of morality is to deal with failures of altruism, nor even to say that we should reject the idea that this is one of morality's aims, but it is a mistake to say that the reason why we should care about some function is that it pertains to our evolutionary past. In other words, the fact that morality has a particular evolutionary function does not seem like the sort of thing that should sway us when answering the functional question. The point is perhaps made clearer once we notice that the function Kitcher assigns to morality, if true, is only contingently true. Morality (or something like it) could have evolved to fulfil some other, less noble, function and should this have happened, we would have been perfectly entitled to rid ourselves of accommodating this function when constructing the concept MORAL VALUE. The biological role of the concept is simply neither here nor there when it comes to theory building. So, how are we to answer the functional question? I suggest that there are five functions we ought to care about.

Firstly, a moral theory ought to be capable of guiding our conduct. This is not to say that we must be able to appeal directly to our theory to resolve any decision. There is an important difference between theories designed to provide standards of conduct and those designed to provide useful decision-making procedures in moments of moral reflection. For example, classical utilitarianism, properly conceived, provides a standard of rightness against which actions may be judged. Yet no sane utilitarian would

recommend calculating the consequences of each potential course of action and acting in accordance with what one judges to produce the best consequences. Instead, they would likely recommend adopting a decision-procedure which is practicable and, on the whole, tends to produce the highest aggregate utility. The theory presented in this dissertation is intended to provide a standard against which to judge conduct, rather than a decision-procedure. In what sense, then, is it capable of guiding action? There are at least two ways. Firstly, when fixing upon a decision-procedure, one must be able to appeal to a standard of conduct to determine the aim of that procedure. The idea of a decision-procedure without a standard of conduct is empty. Secondly, there may be occasions when it is appropriate to appeal directly to a standard of conduct when making moral decisions, for example, when one has a great deal of time to reflect on the relevant features of the case. We will be able to express this action-guiding function of a moral theory through normative language, since this is (at least partly) the function of such language. Our theory may issue in strong deontic claims about what our duties are or what 'right action' consists in. Alternatively, it may be merely evaluative, in telling us which states would be good or better to try and realize. Perhaps it may be both. We ought not decide the question in advance.

It is important to note, however, that a standard of conduct need not yield a definite result about what to do in every case in order to guide conduct. There may be no fact about what it is best to do in certain cases, or no way that one can fulfil one duty without

violating another. The very fact that there is a standard of conduct characterizable in evaluative terminology is sufficient for it to be action-guiding, in my sense of the term. One might worry that this is a very weak conception of what it is for a theory to be action-guiding. And it is true that one may desire something stronger from a moral theory, as do motivational internalists. Yet, the gap between the evaluative and the descriptive is one of kind, not of degree. Any theory which falls on the evaluative side of this divide is markedly different from that which falls on the descriptive side, and a mere description could never guide our action.

The second function a moral theory ought to fulfil is impartiality. Impartiality, as I will understand it, is constitutive of the moral point of view. It is part of what makes moral value different to other kinds of value, like prudential value, which is radically partial. Of these four aims, this is perhaps the most controversial. I will discuss impartiality in more detail in section 3.3.4.

Thirdly, a moral theory ought to account for the importance of welfare. Welfare is central to our moral lives. Bad actions make lives go worse by impacting negatively on our welfare. The same is true when we harm others. Conversely, praiseworthy actions are often those which (at least tend to) improve welfare for someone or some group. It is debatable whether other properties play such a central role in our thinking, but it is not

incompatible with this aim that other things matter besides welfare.[11] This theoretical aim only mandates that, whatever else may matter, welfare must matter too.

Fourthly, a moral theory ought to be as simple as possible. Simplicity can be characterized in a number of distinct ways. As I will understand it, a simple theory is one which makes the fewest assumptions required in order to capture the relevant phenomena. This principle is somewhat similar to Occam's razor, which states, roughly, that one should not multiply entities without necessity. However, simplicity differs from Occam's razor in that it goes beyond the requirement that one not multiply entities without necessity. Some theoretical assumptions may render a theory more complex without requiring us to posit additional objects or properties in the world. For example, BIO$_{RD}$ is a monistic theory of value in that it claims that there is only one essential value: welfare (see section 3.4). Though it would not be multiplying entities to insist that there was more than one essential value, it would require a further theoretical assumption, and thus render the theory less simple. Of course, the simplest possible theory may be extremely complex if the relevant phenomena are impossible to capture without a wide-

[11] Notice how these theoretical aims complement each other, allowing us to distinguish moral value from other kinds of value. For instance, a theory which was only action-guiding and impartial might guide us towards a theory of aesthetic value, where beauty is impartial and demands our attention, but need not contribute to our welfare. A theory which was action-guiding and concerned our welfare, but not impartial, could adequately characteristic prudential value, but not moral value.

range of assumptions. Simplicity should not be pursed fetishistically, but a simple theory that can explain the relevant phenomena should, *ceteris paribus*, be preferred to a more complex one.[12]

And finally, our moral theory ought to be naturalistic. There is no universally agreed upon view of what makes a theory naturalistic. Nonetheless, there are at least two senses of the term that are relevant here. The *ontological naturalist* holds that no theory can permissibly posit the existence of any entity which is not already posited by the natural sciences. The *methodological naturalist* holds that philosophy has no distinctive form of inquiry separate from the natural sciences.[13] Railton writes:

> "Methodological naturalism holds that philosophy does not possess a distinctive, *a priori* method able to yield substantive truths that, in principle, are not subject to any sort of empirical test. Instead, a methodological naturalist believes that philosophy should proceed *a posteriori*, in tandem with – perhaps as a particularly abstract and general part of – the broadly empirical inquiry carried on in the natural and social sciences." (Railton 1989, pp. 155-6)[14]

---

[12] It is perhaps more natural to characterise simplicity as a way in which a theory fulfils its theoretical functions, rather than a theoretical function itself, but the distinction does not appear to be of great importance, whichever way we cash it out.

[13] This view is perhaps most famously associated with Quine (for instance Smart (1968)).

[14] This viewed is echoed by Huw Price:

In constructing BIO$_{RD}$, I follow Railton in suggesting that methodological naturalism be our guiding principle, as opposed to ontological naturalism. As it happens, BIO$_{RD}$ (and Railton's own view, for that matter) are both ontologically natural, in that neither posits the existence of anything outside what the sciences already posit. The important point is not to assume, in advance, that only such natural entities exist, but instead to be guided by the method of inquiry which led us to posit the existence of these entities in the first place. My commitment to methodological naturalism simply arises from the fact that the sciences have undoubtedly been the most impressively powerfully explanatory tools we have. If we can reason in-step with them, we ought to.

These five functions will not be endorsed by every theorist. Some may reject one or more of them, whereas others may add additional functions. There at least two such additional functions I have omitted that it is commonly hoped a theory of MORAL VALUE will meet. The first is an explanatory aim with respect to motivation. Michael Smith argues that *the* moral problem centres on explaining why it is that "if someone judges that it is right that she ϕs then, *ceteris paribus*, she is motivated to ϕ" (Smith 1994, p. 12).

> "What is philosophical naturalism? Most fundamentally, presumably, it is the view that natural science properly constrains philosophy, in the following sense. The concerns of the two disciplines are not simply disjointed, and science takes the lead where the two overlap. At the very least, then, to be a philosophical naturalist is to believe that philosophy is not simply a different enterprise from science, and that philosophy should defer to science, where the concerns of the two disciplines coincide." (Price 2013, p. 3)

Motivational internalism, as this view is often called, is endorsed by many theorists (Hare 1952; Nagel 1970; McDowell 1985; Korsgaard 1986). The rival view, externalism, claims that there is no necessary connection between judging that ϕ is right and being motivated to ϕ, a position endorsed by many other theorists (Foot 1972; Railton 1986; Brink 1986; Copp 1997; Svavarsdottir 1999). According to internalists an adequate moral theory must explain moral motivation. But an externalist moral theory cannot be deemed a failure (at least on its own terms) if it fails to explain moral motivation, since it never tried to explain it in the first place. An extended discussion of the merits of internalism/externalism would take us too far afield. My sympathies, however, are with the externalist, hence why I have omitted the aim of explaining moral motivation from the five theoretical aims outlined above. Although it seems perfectly plausible to insist that, in most cases, we tend to be motivated to act morally, it seems simply false to insist that this connection is a conceptual necessity. I have (more frequently than I would care to admit) at least seemed to myself entirely unmotivated to take what I consider the right course of action. So much so that I would consider it a theoretical vice if BIO$_{RD}$ entailed motivational internalism.

The second function omitted above is *categoricity*. Kant famously argued, and contemporary Kantians often agree (e.g., Korsgaard (1996)), that morality must be authoritative in a way that makes it a requirement of reason. Kant expressed this thought *via* the idea that morality was composed of a series of categorical imperatives,

that is, requirements unconditioned upon anything else, such as our desires, or even human nature. It is sometimes said that categorical requirements are requirements of reason 'as such.' (e.g., Cohen (1996)).[15] However, it is hard to say precisely what the authority of morality might consist in, and it is even more contentious how morality might achieve this authority. To make matters more complicated, the term 'categoricity' is often conflated between two different theses in contemporary metaethics. The first understanding is that which I have outlined above; morality must be a requirement of reason or rationality, and thus authoritative for creatures like us. The second understanding is that we can have a moral reason for some action regardless of our desires, or, as Bernard Williams suggested, in the absence of 'sound deliberative route' from one's motivational attitudes to the conclusion that one has a reason in favour of that action (Williams 1979).[16] We will see that BIO$_{RD}$ is categorical in the latter, but not the former, sense.

Let us now return to the issue of truth, a concern raised in light of Kitcher's emphasis on biological function. So far, my focus has been entirely with the conceptual role of MORAL VALUE. I suggested that MORAL VALUE should guide our action impartially, with

[15] The authoritative nature of morality is closely linked to motivational internalism, for one way in which morality could command us is by compelling us to act in certain ways. However, the two can come apart. It does not seem incoherent to suggest that something could be authoritative without necessitating obedience. Laws are authoritative in some sense, but there are criminals.

[16] Also sometimes referred to as 'reasons externalism' (Finlay and Schroeder 2017).

an emphasis on welfare, simplicity and naturalism. Yet, unless MORAL VALUE denotes some property, <moral value>, then MORAL VALUE will be an empty concept. We can revise our concepts all we like, but what's the use if they fail to denote anything?

Moore famously argued in his *Principia Ethica* that 'goodness' was a non-natural property, only capable of apprehending *a priori* (if at all). Moore's arguments and the objections to them are well-known, and so I pass over them noting only that Moore's *a priori* approach is not available to me in light of my commitment to methodological naturalism. Finding a satisfying answer to this question as a naturalist has led some metaethicists to despair of finding a moral theory that is both realist and cognitivist in kind. This despair takes two forms. The first is error theory (Mackie 1977; Olson 2014). The error theorist argues that moral claims presuppose the existence of properties (likely of the Moorean kind) but that there are, in fact, no such properties. It follows that moral discourse makes a massive presupposition failure and that all moral claims are either false, or at least not true. The second form of despair (though its advocates are likely to resent the characterisation) are the varieties of expressivism (Ayer 1936; Stevenson 1944; Gibbard 1990; Blackburn 1998). The expressivist argues that, despite appearances to the contrary, our moral discourse is not cognitivist. This picture is slightly more complicated in that one could interpret the expressivist as offering a conceptual analysis of our moral discourse, albeit a surprising one, or as engaging in a form of conceptual ethics in which, although our discourse is in fact cognitivist in kind, it can be re-interpreted in

expressivist terms. As a general rule, it seems that earlier emotivists such as Ayer and Stevenson were engaging in conceptual analysis, whereas later expressivists are engaging in conceptual ethics. For instance, Blackburn's quasi-realist project has the aim of providing a logic of attitudes in order to account for the surface-level cognitivist appearance of moral discourse. These interpretive matters aside, there is sufficient reason to reject, at least for the time being, the expressivist approach. I have already claimed that a central feature of (my concept) MORAL VALUE is that it manifests itself in our discourse cognitively. One must only turn to expressivism if it turns out that we cannot take this cognitivist appearance at face value. I believe, however, that we can and so there is little reason to turn to expressivism, unless one thinks that their story is the more plausible one to begin with. As such, I will have no more to say about expressivism, although one can read my arguments as an attempt to show the expressivist that their despair of finding a satisfying cognitivist moral theory was too hasty. The error theorist thus provides the foil, since we share the view that moral language is, at its root, cognitive. The burden is on me to provide an account of <moral value> that satisfies the revised concept MORAL VALUE.

2.3. Towards response-dependence

The response-dependent[17] theorist makes the following general claim:

RD**:** *x* is morally valuable if and only if and because *x* elicits *R* from *S*.[18]

The crucial thing to notice about RD is that, in addition to a conditional claim, the 'and because' makes an explanatory priority claim. Anyone may accept that there is a hypothetical agent whose reactions are always morally appropriate. But, if RD is correct, such an agent would not simply be tracking the independent moral truth with their attitudes. Rather, their attitudes determine what's valuable in the first place.

This defining feature of response-dependence is often seen by its opponents as a fatal weakness. In Plato's *Euthryphro,* Socrates asks the titular interlocutor whether what is

[17] A brief note on nomenclature is in order, since my use of 'response-dependence' is somewhat non-standard. The term 'response-dependence' is due to Johnston (Smith, Lewis, and Johnston 1989). Johnston and others use the term to refer to a feature of concepts, rather than properties. The term 'dispositional theory of value' is often used to refer to properties that fit the schema, RD. I prefer to use the term 'response-dependence' for properties too, since we should not suggest that *x* must elicit *R* from *S* due to some disposition *S* has. A property may be response-dependent even if *S*'s reaction is not due to some stable disposition. Perhaps not many interesting properties will fail to result from stable dispositions, but they should not be ruled-out of contention by the very name of the theory. There are, however, those who insist that there has always been an ambiguity about whether response-dependence is a theory about concepts or properties (Wedgwood 1997; LeBar 2013).

[18] Sometimes this schema includes an additional variable, *C*, for context. I omit it here only because it is somewhat irrelevant for my purposes.

holy is holy because it is loved by the Gods, or whether it is loved by the Gods because it is holy. This is a dilemma for Euthyphro, since the first option ties holiness to the whims of the capricious Olympians and their kin, but the second option makes holiness prior to the Gods, in a way that restricts their power (perhaps impiously so). One can ponder an analogous dichotomy with respect to the good. Is the good good because we judge it so (or desire it, or desire to desire it, etc.) or do we judge it to be good because it is good? However, unlike the case of piety, there is now no sting in embracing the second option, where goodness is prior to our judgements (or reactions). For those of us engaging in secular ethics there is, apparently, no problem in assuming that the good is prior to our reactions because its priority does not threaten the omnipotence or prestige of any sacred cow. And given the apparent absence of any pressure to accept the first option, why does the response-dependent theorist do so?

The answer is that despite this initial stumbling-block, response-dependence arises from an extremely attractive naturalistic way of understanding ethics. Response-dependence takes as its inspiration Locke's distinction between primary and secondary qualities (*AECHU,* Bk. II, Ch. 8). For Locke, primary qualities are those features of objects which it must retain, no matter what changes or alterations it undergoes. Secondary qualities, however, are 'powers' or dispositions of objects to produce what Locke called 'sensations', or what I am calling 'responses', in subjects. Locke does not discuss value in relation to secondary qualities. His primary example is colour. But the

analogy between colour and value has been emphasized more recently, particularly by McDowell (1985).[19] According to McDowell, conceiving of value of as a secondary quality preserves the 'common-sense' phenomenology of value, which is one of sensitivity to certain objective features of the world. Falsely assuming that the phenomenology of value demands a conception of moral properties as primary qualities is what mistakenly motivates error theory. Although secondary qualities, including value, are not in the objects themselves, the dispositions of those objects to produce certain responses most certainly are features of the objects (being, according to Locke, reducible to primary qualities). Thus, we can preserve the appearances of sensitivity without having to make the erroneous claim that value is a feature of objects in the same way that physical properties, such as solidity, are.

As already noted, Locke's primary concern was not value. Yet there is a historical tradition out of which this concern naturally emerges; sentimentalism.[20] We can understand sentimentalism, broadly, as the view that morality is 'grounded-in' or dependent upon our psychology. The following from Hume illustrates the general idea well:

[19] See also Blackburn (1985, 1993) Smith (1998), Pettit (1991), Jackson and Pettit (2002), Railton (1998), Wiggins (1987).

[20] I do not wish to suggest that there is any necessity or sufficiency relationship between sentimentalism and response-dependence.

> "The final sentence […] which pronounces characters or actions admirable or odious, praiseworthy or blameworthy […] depends on some internal sense or feeling which nature has made universal in the whole species." (*E*, Sec. 1)

Today, sentimentalism still has a large number of adherents.[21] What unites them is a vision of morality as dependent upon creatures for whom things matters, and not as an independent and distant realm of independent truths to be intuited or discovered. Moral theory is not, according to the sentimentalist, a matter of describing an external world of moral facts. It is about squaring our cares and commitments with our nature as conscious creatures, capable of experiencing a range of emotions, and with the natural world both as it appears to us in conscious experience, and as it is revealed by the methods of the natural sciences. Moral theory is a matter of describing a normative universe one can, on reflection, accept in light of these constraints. Response-dependence takes as its starting point the fact that value arises as a result of our (perhaps counterfactual) interaction with the world. The process of reflecting on what we really care about is what guides the processes of filling in the schema RD.

This all means that response-dependence is well-suited to satisfy the relevant theoretical aims discussed in section two of this chapter. It does so whilst also preserving as much of the concept of MORAL VALUE as possible. Most notably, response-

[21] For an overview, see Kauppinen (2017).

dependence has impeccable naturalistic credentials in both the ontological sense and the methodological. Ontologically, all that response-dependence requires is the ability of objects to produce reactions in a hypothetical observer. There is nothing naturalistically suspect about such properties. Response-dependence coheres equally well the more foundational methodological naturalism, though this takes some more effort to show.

We can formulate a version of RD designed to cover only the relevant concept:

RD$_{\text{CONCEPT}}$**:** *x* is MORALLY VALUABLE if and only if and because *x* elicits *R* from *S*.

Some advocates of RD assert that, when fleshed out, RD$_{\text{CONCEPT}}$ must have the status of a necessary *a priori* conceptual truth (Smith 1994, 1998). There are at least two reasons why one might think this. The first is that moral properties are non-natural and can only be known *a priori*. Therefore, it must be known *a priori* what the content of the concept MORAL VALUE is. This argument is particularly unconvincing, however, in light of the fact that a primary motivation for accepting any form of RD is a commitment to naturalism. The second reason is that one simply conceives of RD$_{\text{CONCEPT}}$ as saying something true about the ordinary concept. RD$_{\text{CONCEPT,}}$ when fleshed out, just is a non-obvious but nonetheless analytical truth about the ordinary concept MORAL VALUE. The conceptual ethicist, however, has no reason to assert this, since one requires RD$_{\text{CONCEPT}}$ to have the status as a necessary *a priori* truth only if one believes that one is revealing a common necessary conceptual connection between the left and right-hand side of the

biconditional. Though some believe this to be important, I have already argued in section 2.2 that it is not. We ought to care more about revising our concepts than describing them. Instead, let us focus for the moment, not on the concept but on the property of moral value and consider the following:

RD$_{\text{PROPERTY}}$**:** *x* has the property <moral value> if and only if and because *x* elicits *R* from *S*.

This claim, when fully fleshed out, is a necessary but *a posteriori* truth.[22] Though it is not an identity claim, it is in the same semantic family as the identity 'water equals H_2O.'[23] The thought is that one requires some interaction with the world to understand the value of certain states of affairs. It is easier to see this if we flesh out RD$_{\text{PROPERTY}}$ along broadly similar lines as will occur in subsequent chapters. Let us say that *R* is a desire and *S* is someone fully informed, rational and who cares about welfare. In that case, the truth of RD$_{\text{PROPERTY}}$ is a discovery made when one comes to see the value in promoting welfare by having experienced such a promotion oneself. But, once the discovery is made, the term 'moral value' acts like a rigid designator, for what it is to be morally valuable simply is to elicit a desire from the appropriate kind of agent. None of this requires any unique epistemological faculty, elusive to investigation from the natural

[22] The possibility of which is famously defended by Kripke (1980).

[23] See also Brink (1989, pp. 157, 176).

sciences. It is of a piece with scientific inquiry, even if the subject matter is distinct. With respect to the additional theoretical aims, meeting these will largely depend on how RD is fleshed out in subsequent chapters.[24]

2.4. Fitting attitudes

Within this sentimentalist tradition, response-dependence is rivalled by what is sometimes known as the 'fitting-attitude' account of value. Fitting-attitude accounts, as I will understand them, adopt the following schema:

FA: x is morally valuable if and only if it is fitting for S to value x.

Fitting attitude accounts arguably begin with Brentano[25] and Ewing but they find their contemporary origin, as does response-dependence, in the work of McDowell and Wiggins. Consider the following from McDowell:

> "The idea of value experience involves taking admiration, say, to represent its object as having a property which (although there in the object) is essentially subjective in much the same way as the property that an object is represented as having by an experience of redness – that is, understood

[24] It is worth noting, however, that response-dependence, though it makes value mind-dependent, does allow us to be mistaken about moral value. If the reactions of an idealised agent is what matters, then our reactions can be mistaken insofar as we fail to be ideal.

[25] See Chilshom (1986).

> adequately only in terms of the appropriate modification of human (or similar) sensibility. The disanalogy, now, is that a virtue (say) is conceived to be not merely such as to elicit the appropriate 'attitude' (as a colour is merely such as to cause the appropriate experiences), but rather such as to *merit* it. And this makes it doubtful whether merely causal explanations of value experience are relevant…" (McDowell 1985, p. 118)[26]

Whereas the response-dependent theory suggests that eliciting an attitude from the appropriate agent is enough to explain that object's value, McDowell suggests that value is best thought of, not just as eliciting a response, but as *meriting* it.

The most popular account of what it is for a response to be fitting (or for an object to merit a response, in McDowell's terminology) is Scanlon's *buck-passing* theory of reasons (Scanlon 1998). According to Scanlon, a response is fitting if there is sufficient reason for it.[27] But Scanlon's account faces a well-known problem: the 'wrong kind of reasons problem' (D'Arms and Jacobson 2000; Rabinowicz and Ronnow-Rasmussen 2004). Suppose that it is fitting that I admire you whenever there is sufficient reason to admire you. There is sufficient reason to admire you when, let us say, you have a generous

[26] See Wiggins' claim that subjectivism ought to take the following form: "*x* is good if and only if *x* is such as to arouse/such as to make appropriate the sentiment of approbation." (1987, p. 190)

[27] Scanlon's account is called the 'buck-passing' theory since he does not believe that the fact that there is a reason in favour of an action itself provides any additional reason to perform it. The reason to perform the action is given by the features of the world that ground the reason. This aspect of his account is not relevant for my purposes.

spirit or display courage in the face of adversity. On the face of it, these are 'the right kinds of reasons' to admire someone. Yet, what if the devil offers to make you rich if you admire him? Most of us believe that it would not be fitting to admire the devil under any circumstances. Yet, there does appear to be sufficient reason to admire him, given the significant payment on offer, even if it is a reason of 'the wrong kind.' The trouble for the buck-passing account, then, is to say when a reason is of the right kind and when it of the wrong kind in a non-circular way.[28] In despair of finding an adequate solution to this problem, McHugh and Way have recently argued that fittingness should be seen as the fundamental normative relation between attitudes and objects (McHugh and Way 2016).

It would be fairly straightforward to adapt the specific version of RD to be argued for in subsequent chapters, BIO_{RD}, and give it a fittingness gloss:

> BIO_{FA}: an object, x, is morally valuable if and only if it is fitting for a properly informed, instrumentally and formally rational, benevolent and otherwise minimal observer to value it.

[28] The most well-known attempt is to draw a distinction between reasons for an attitude and reasons for causing oneself to have an attitude (e.g., Parfit, 2011, Appendix A). I will not survey the merits of this or other attempts here.

If the reader is persuaded by my arguments regarding the benevolent observer but is otherwise committed to fittingness-style metaethics, then they ought to feel free to adopt BIO_{FA}. Much of what follows this chapter does not depend on the truth of RD over FA. Yet, without being able to provide a decisive argument, I believe that there is reason to prefer RD to FA.

Suppose that BIO_{FA} is true. What explains why it is fitting for a benevolent observer to value a certain object? According to the *buck-passing* account, the explanation bottoms out in the existence of reasons, for which there is no further normative explanation. According to the fittingness-first account, fittingness, not reasons, are primitive. BIO_{RD}, on the other hand, claims that the fact that a benevolent observer would have a non-truth-oriented attitude towards an object is, itself, a sufficient explanation of its value. The question is this: do we need the addition of 'fittingness' to explain why it is that some object is valuable? The benevolent observer will have the same reactions to the same objects according to both BIO_{RD} and BIO_{FA}. Yet, BIO_{FA} adds an additional normative layer, 'fittingness', on top of BIO_{RD}'s explanation. McDowell suggests that this additional, 'non-causal', layer is needed because the phenomenology of value is that objects of value present themselves as meriting responses, not just as eliciting them. But this is not so important to the conceptual ethicist whose aim is not to capture the typical phenomenology of value (assuming that McDowell is even correct). And, furthermore, it seems compatible with our phenomenology of value that objects present themselves as

meriting responses, and to have value determined only causally by the reactions of a highly idealised benevolent version of ourselves. We are not ideal, and so our phenomenology may well be distinct.

In fact, this last point reveals how strange BIO$_{FA}$ in fact is. The theory tells us only what it is fitting for an ideal observer to value, and whatever it is fitting for them to value will be, due to the way we construct the observer (see chapter 3), what they *de facto* value. But we care about what it is fitting for *us* to value. It is better, then, to say that it is fitting for us to value something and specify the standards associated with fittingness. Yet, if the arguments of subsequent chapters are compelling, then those standards will be the standards that result from the psychology of the benevolent observer. In other words, suppose that, in the schema FA, *S* is any agent whatsoever. Then, still assuming subsequent arguments are compelling, the standards for what is fitting must be determined by how a benevolent observer would, in fact, react. But this is exactly the 'merely causal' claim that the response-dependent version of the theory claims. 'Fittingness', then, appears to be doing little work.

2.5. Summary

In this chapter I paved the way for my defence of BIO$_{RD}$. I did so by laying out the kind of project that I take myself to be engaged in: conceptual ethics. I discussed the aims I thought should constrain the project, and described the concept I will be revising

in light of those aims. I then moved on to discuss response-dependence as a metaethical theory. A full defence of response-dependence would itself be dissertation-length, but I have attempted to lay out some of its charms, particularly with respect to the theoretical aim of naturalism. Thus, for the rest of this dissertation, I will assume that the following schema offers the best way to understand moral value:

RD**:** *x* is morally valuable if and only if and because *x* elicits *R* from *S.*

In the next chapter, I begin filling-out this schema by discussing the relevant agent, *S.* In chapter 4 I discuss the variables *x* and *R*, or the objects and relevant reactions. I then take stock before moving on to discuss the impact of the benevolent ideal observer theory on our moral discourse in chapter 5.

Chapter 3: The ideal agent

This chapter begins the process of filling out the response dependent schema RD:

RD**:** *x* is morally valuable if and only if and because *x* elicits *R* from *S.*

More precisely, this chapter discusses the idealised agent, *S.* In short, I propose that *S* is best conceived of as a properly informed, rational, benevolent and otherwise minimal spectator. I defend each of these attributes in turn, beginning with 'properly informed' in section 3.1, rational in section 3.2, benevolent in section 3.3 and minimal in section 3.5, where I also consider the related question of whether there are many benevolent observers or just one. For ease of exposition, up until that point (and elsewhere in this dissertation where number is irrelevant), I will speak as if there is only one benevolent observer.

Some brief methodological remarks are in order. Given that I am engaging in radical conceptual ethics, the criteria outlined below are only constrained by the starting concept and theoretical aims. As a reminder, those theoretical aims were (i) the ability to guide action; (ii) impartiality; (iii) accounting for the importance of welfare; (iv) simplicity; and, finally, (v) naturalism.

3.1. Properly informed

The benevolent observer ought to be properly informed. This criterion is interesting, in that it is not immediately entailed by any of our theoretical aims. Yet, it is needed to preserve most of what would be required by anyone familiar with some concept of MORAL VALUE. Suppose that an observer was ignorant of some fact. Then they may desire that I perform some action, which, if they were aware of this fact, they would strongly desire I *not* perform. Thus, we require full-information for the theory to issue in plausible guidance.

To be properly informed is, by definition, to know all the relevant non-moral facts. In section 3.1.1, I show how to determine membership of the set of relevant non-moral facts. In section 3.1.2, I discuss a subset of the relevant facts – facts about the experiences of others. In order to know these facts, Firth (1951) argued that an ideal observer must be omnipercipient. That is, they must possess perfect (or as close to perfect as possible) imaginative capacities. Unlike Firth, I do not think that the ideal observer must be omnipercipient. Firth includes omnipercipience as a separate attribute of his ideal observer, but since omnipercipience is a capacity for gaining a particular kind of knowledge, I discuss it in in relation to the requirement of being properly informed.[1]

[1] Firth hints that this is so, writing "The imaginal powers of the ideal observer, to be sure, are very closely related to his omniscience." (Firth 1951 p. 335)

3.1.1. Relevant knowledge

Let us label the set of relevant non-moral facts $\{F_R\}$. Firth's ideal observer knows *all* the non-moral facts. This is merely because Firth thinks that, although there is some set of relevant facts, there is no way to independently specify $\{F_R\}$ without vicious circularity.

> "…it is evident that a concept of relevance cannot be employed in defining an ideal observer. To say that a certain body of factual knowledge is not relevant to the rightness or wrongness of a given act, is to say […] that the dispositions of an ideal observer toward the given act would be the same whether or not he possessed that particular body of factual knowledge or any part of it. It follows, therefore, that in order to explain what we mean by "relevant knowledge," we should have to employ the very concept of ideal observer which we are attempting to define." (Firth, 1951, pp. 333-4)

Firth does, however, consider the following proposal for determining $\{F_R\}$. He attributes it (rather generously) to Brandt (1954b) and so I shall follow suit and call it 'Brandt's proposal' (Firth 1954).[2] Brandt's proposal is as follows: take an ideal observer

[2] Here is Brandt's original proposal:

> "For what a person needs to be vividly conscious of, in judging or reacting to an ethical situation, is simply all those facts vivid awareness of which would make a difference to his ethical reaction to this case if (to use Firth's other qualifications) he were a disinterested, dispassionate but otherwise normal person. This is all an "ideal observer" needs." (Brandt 1954b, p. 410)

who lacks any beliefs. Add true beliefs to the observer's belief set and if one of these beliefs would make a difference to the observer's relevant reactions, then the fact which is the object of that belief is a member of $\{F_R\}$.[3]

Firth rejects Brandt's proposal. He worries that "[f]or any given fact […] we could find some set of true beliefs which would be sufficiently incomplete to make that fact seem irrelevant." (Firth 1955, p. 418) Firth invites us to consider a case in which the observer knows that a man and a woman are not married, but is unaware that they are engaged. If the observer learns that the man has married another woman, the observer's reactions wouldn't be affected by the man's act of promise-breaking (since he does not know that a promise has been broken). Firth then concludes that, according to Brandt's proposal, facts about promise-keeping will be irrelevant (in this situation) when they are, in fact, relevant.

However, Brandt's proposal can be amended to avoid this problem. Recall, our task is

---

And here is Firth's gloss on this proposal:

> "If I understand this argument correctly, Brandt is proposing, in effect, that we begin with the concept of a person (an ideal observer,) who is disinterested, dispassionate, and normal, but whose knowledge and factual beliefs remain unspecified. Then, with respect to a given ethical situation (S) we define a body of non-ethical facts (F) as "those facts vivid awareness of which would make a difference to the ethical reactions of an ideal observer,." We then define a more determinate kind of ideal observer (an ideal observer$_2$) for case S as a person who is disinterested, dispassionate, and normal, but who also knows F. And finally we define "right" and "wrong," as applied to situation S, in terms of the reactions of an ideal observer of this more determinate kind (an ideal observer)." (Firth 1954, p. 417)

[3] More precisely, the fact corresponding to the proposition that is the object of one's belief is a member of $\{F_R\}$. The same proposal is considered by Harrison (1971, pp. 152-3)

to determine the set of relevant non-moral facts, $\{F_R\}$. We can define another set {F} of all non-moral facts of which $\{F_R\}$ will be a subset. In order to determine the content of this subset, take the power set of {F}, $\wp\{F\}$. Then, take every permutation of $\wp\{F\}$ and number each one $p_1, p_2, \ldots, p_n$. This gives us every distinct possible way of ordering the members of the set {F}. Now, as in Brandt's proposal, we imagine an ideal observer who lacks any specific beliefs. Take the permutation, p_1, and add it to the observer's empty belief set. Then, remove a belief in each element of p_1 in some fixed order from the observer's belief set.[4] If any of the observer's relevant reactions change when a belief is removed, the fact or proposition that was the content of that belief is relevant, and a member of the set $\{F_R\}$. Repeat for every p_n. I will call this *the modified Brandt proposal.*

The modified Brandt proposal works because every single fact is considered by the observer in every distinct possible order. Consider the following example.[5] A husband gives his wife lilies. This has two effects. It makes her smile, as a sign of her husband's care, and it makes her laugh, since lilies were present on some humorous occasion in their past. We can assume that smiling and laughing are both pleasant experiences for

[4] The precise order does not matter so long as it remains the same. The most natural order is simply to take the first element, then the second, then the third, and so on.

[5] It is possible to use Firth's original counter-example to illustrate the success of the modified Brandt proposal, however, doing so is perhaps more confusing since it contains a negative fact: 'that this pair of persons is not married.' Including negative facts in {F} means that {F} will be an infinite set (see fn. 7 of this chapter).

the wife and that there is some positive moral value in her husband's actions. The fact that lilies make the wife smile (call this fact *S*) and the fact that they make the wife laugh (call this fact *L*) are both relevant to assessing the husband's action. If we add the permutation (*L, S*) to the observer's belief set, then remove each belief in reverse order, we may find that removing the observer's knowledge of *S* may not alter their relevant reaction. Suppose that the relevant reaction in this case is the observer's desire that the husband give his wife flowers and that the observer acquires this desire when (*L, S*) is added to their belief set. In this case, when we remove *S* from the observer's belief set, the observer still knows *L* and their desire that the husband give his wife flowers remains unchanged. *S* appears not to have made a difference, and thus, by Brandt's original proposal, *S* is irrelevant. However, according to the modified proposal, it is relevant. Continuing with the procedure described above, we would next remove *L* from the observer's belief set. Removing *L* causes (we can suppose) the observer's desire that the husband give his wife flowers to change, and thus *L* is relevant.[6] But we must then add the permutation (*S*, *L*) to the observer's belief set. In this case, removing each belief in reverse order gives us no change after removing *L*, but a change after removing *S*. Therefore, *S* is relevant. Since the observer's relevant reactions changed after removing

[6] For a mental state to 'change' suggests, in ordinary usage, that the mental state remains, but some of its properties alter. In the sense of 'change' used here, it also encompasses the acquiring and losing of a mental state altogether.

both *S* and *L* on at least one occasion, both facts are relevant, and members of the set $\{F_R\}$.[7] A properly informed observer may know all the non-moral facts, if all those facts are relevant. Thus, being properly informed is not incompatible with omniscience.

It is worth noting that the modified Brandt proposal yields the set $\{F_R\}$ in such a way that it cannot be mistaken. Because moral value is determined by the reactions of a benevolent agent, the facts that are relevant are simply those that a benevolent agent would respond to. The modified Brandt proposal is, therefore, best seen as an attempt to determine the elements of this set, 'after the fact', as it were, by noting which ones change the relevant responses of the benevolent observer.

3.1.2. Omnipercipience

Most advocates of an ideal observer theory have included some form of imaginative capacity in their specification of the observer. As discussed in section 1.2.2., imagination

[7] One complication of the modified Brandt proposal concerns the cardinality of the set {F}. {F} is either finite or countably infinite. If it is finite, then $\wp\{F\}$ is also finite and so is the number of permutations of $\wp\{F\}$ (The cardinality of the power set $\wp\{F\}$ is $2^{|F|}$. The number of permutations of $\wp\{F\}$is $2^{|F|}!$). If {F} is countably infinite, then $\wp\{F\}$ is uncountably infinite, and so is the number of permutations of $\wp\{F\}$. This may seem worrying, as no agent, not even an ideal one, could ever complete the task of determining $\{F_R\}$. However, so long as $\{F_R\}$ can be well-defined, then the modified Brandt proposal serves its purpose. It was never intended to outline a procedure that one could realistically follow in order to determine $\{F_R\}$. It was merely designed to show that $\{F_R\}$ can be defined without circularity. Since $\{F_R\}$ is well-defined even when {F} is infinite, the modified Brandt proposal remains fit for purpose.

plays a critical role in Adam Smith's impartial spectator theory. More recently, Firth endorsed the strongest possible imaginative capacity, omnipercipience. The term suggests a perfect, all-encompassing ability to imagine every fact, but Firth merely wrote that "the ideal observer must be characterized by extraordinary powers of imagination." (Firth 1955, p. 335). Extraordinary imaginative capacities, it seems, may fall short of perfection. When motivating the inclusion of omnipercipience, Firth wrote:

> "Practical moralists have often maintained that lack of imagination is responsible for many crimes, and some have suggested that our failure to treat strangers like brothers is in large part a result of our inability to imagine the joys and sorrows of strangers as vividly as those of our sibling." (1955, p. 335)[8]

A more detailed argument is not forthcoming. Yet, I take it that the thought behind the inclusion of an imaginative capacity is something like the following: I am aware of my own experiences. When someone acts kindly towards me, it is almost always a

[8] Firth, at times, suggests that the imaginative capacity is focused not on the experiences of others, but about facts more generally:

> "The ideal observer must be able, on the contrary, simultaneously to visualize all actual facts, and the consequences of all possible acts in any given situation, just as vividly as he would if he were actually perceiving them all." (Firth 1951, p. 335)

There is a distinction between imagining what it is like for someone to experience something and imagining that thing. Whilst the former does seem important for moral understanding, it is not clear what relevance, if any, the latter has.

pleasant experience. When someone is overtly hateful towards me, it is almost always an unpleasant experience. When I exercise my imaginative capacity, I notice that the experiences of others are like mine. They too would welcome kindness and avoid hatred. We might then think that because I am no more important than any other person, the pleasant experiences of others are just as important as my own. Less drastically, we might conclude that if I can cause others to have similar pleasant experiences without much effort on my part, I ought to do so.

But imaginative acquaintance with the experiences of others cannot, by itself, *require* us to have altruistic, other-regarding desires.[9] Our ability to imagine the experiences of others may only lead a vindictive observer to better understand how to harm people. In the story outlined above, it was only the thought that I am not more important than others, or that I may as well help others if it does me no great harm, which lead to altruistic desires. But this further thought does not seem required by the capacity of imaginative acquaintance.

Yet, even if our imaginative capacities do not demand altruism, they may frequently, as a matter of empirical fact, cause us to have such desires. Those of us who do care about the welfare of other people have probably, at some point, imagined what it must be like to live someone else's life. People who donate money to charities in order to

[9] Motivational internalists may disagree, e.g. Hare (1981, pp. 88 - 106).

improve the lives of distant strangers whom they will never meet may have imagined what it would be like to live a life without stable access to clean water or nutritious food. Some go further in refraining from consuming animal products after imagining what it is like to be a non-human animal kept in the appalling conditions common in today's commercial agriculture. A plausible hypothesis, then, is that one function of imagination is to create some kind of care or concern for the well-being of other sentient beings. Imagination does not rationally require such a concern, nor does it necessarily lead to one, but it may, in general, cause such a concern in most ordinary humans.[10]

A benevolent observer, I want to suggest, does not require an extraordinary imaginative capacity. If, as has been argued, the function of an imaginative capacity in regular people is to provide knowledge of a particular kind of fact, an other-regarding 'what it is like' fact (which are supposedly relevant and thus merit inclusion in $\{F_R\}$), then this capacity is surplus to requirement in an observer who *already* cares about the welfare of others. Though we may never encounter such an agent, it is not incoherent to suppose that someone could care about the welfare of other people without having imagined what it is like to live their lives. This may seem strange, but notice that we do not think that imagination is required to *sustain* care. My local grocery store sells

[10] It is not important for this dissertation that this particular empirical hypothesis is true, and it would require more than armchair speculation to verify it. It is merely an intuitive attempt to explain the apparent connection between altruism and imagination.

beetroot chips. I like them, but my friend does not. Nonetheless, when my friend goes to the store alone, he will sometimes buy them for me. He desires to buy the chips for me, not because he is imagining what it is like for me to eat beetroot chips at the time he purchases them. He simply cares about me and that is enough to generate his other-regarding desire.

But we can go further. My friend may *never* have imagined what it is like for me to eat beetroot chips, the mere taste of them being so repulsive to him, given his intense loathing of all things beetroot. Yet, he desires to buy them for me because he cares about my welfare, and knows that beetroot chips will improve it. My friend may have, at some point in his past, imagined what it is like for me to live in the world in a general sense, and as a result, come to care about me. Or, he may, as a child, have 'learned to see the world through other people's eyes', and since then developed a measured care for other people. But once the care is fixed, imagination is not required to exhibit care. In other words, the etiology of his cares is unimportant to their sustained existence or exercise. Once someone has a care for the welfare of other beings, they may continue to care without exercising their imaginative capacities.[11] Since the agent at the centre of this

[11] Of course, one may come to care less and less about others over time, in which case imagination might help reinstate the care. The benevolent observer, however, is defined so that they never lose their care for the welfare of others.

theory is idealized, and the etiology of our cares is non-essential, we can separate impartial care from the imaginative capacity that usually accompanies or causes it.

Importantly, this does not mean that facts about what it is like for another person to experience something are not relevant. The fact that I enjoy beetroot chips is crucial to my friend's decision buy them for me. But knowing 'that my friend enjoys beetroot chips' and imagining the enjoyable experience of your friend eating beetroot chips are different. More generally, facts about the welfare of conscious creatures are distinct from the capacity to imagine the experiences of conscious creatures that contribute to their welfare.

The advantage that BIO$_{RD}$ has over its rivals in excluding omnipercipience is significant. The source of much antipathy towards heavily idealised observers often stems from the fact that they require an extraordinary imagination. For instance, Connie Rosati argues that persons necessarily inhabit particular psychological perspectives and that there is no good reason to think that there is one privileged, idealised, perspective from which one can accurately assess the value of certain experiences of others who inhabit their own perspective (Rosati 1995).[12]

BIO$_{RD}$ avoids these objections because it does not require the observer to inhabit the perspectives of everyone at once *via* any extraordinary imaginative capacity. The

[12] Similar concerns are raised by Sobel (1994), Loeb (1995), and Velleman (1988).

observer only needs to know to what degree certain experiences impact welfare. Since the observer knows all the relevant physical facts, including those about welfare, the observer will have this information. To suggest otherwise, one would have to make one of two argumentative moves.

The first would be to claim that imagination is epistemically necessary in order to access what are essentially subjective facts to do with welfare. But, to repeat, facts about welfare are distinct from one's ability to imagine these facts. Neither, it seems, can one insist that imagination is required in order to appreciate the significance of these facts, since we have already stipulated that the observer cares about our welfare (and below I argue that there is nothing troubling about such a stipulation). All BIO$_{RD}$ requires is that there is some route to understanding how someone's welfare is affected without undergoing the relevant experience oneself. So long as we are committed physicalists, then we must suppose that there is such a method and that a completed neuroscience can, in the long run, inform us about our welfare.

This objection, does, however, highlight something important. It does sound strange to insist that the agent whose reactions determines moral value is completely devoid of any understanding of *what it is like* for another person to experience pleasure and pain, for example. But BIO$_{RD}$ does not say this. All that I claim in this section is that the observer need not possess any kind of imaginative capacity *insofar as this capacity is necessary for being properly informed*. In section 3.3.1, I suggest that a certain level of

imagination may be required in order to exhibit benevolence, defined as a final care for the welfare of conscious creatures, since imagination is required in order to exhibit care. Yet, the level of imagination required for care is nowhere near as strong as is usually attributed to an ideal observer whose imagination is thought necessary for being properly informed.

The second, less plausible, option is to argue that since facts about welfare are physical, and the considerations that Rosati raises show that one cannot know all the facts about welfare, no one agent can know all the physical facts. Notably, neither Rosati or those who argue in a similar fashion make this argument. It would seem to prove too much, by ruling out the possibility of omniscience with respect to the physical facts.[13, 14]

[13] Another interesting thought is that anomalous monism may be true, in which case the observer could know all the physical facts without being able to predict the psychological states of welfare subjects (Davidson 1970). Full consideration of this possibility is beyond the scope of the present work.

[14] Wiland (2017) has recently argued that some morally relevant facts can only be ascertained or appreciated by individual groups, such as marginalised members of a particular race or sexual orientation. I do not believe that my argument is incompatible with such views if the claim is that, in actual fact, there are certain morally important facts that are extremely difficult for individuals outside of a particular group to fully appreciate. However, the ideal observer gives every welfare-subject equal consideration and knows facts about their experiences just as well as they know facts about the experiences of everyone else. In this respect, they are better than most of us. BIO$_{RD}$ is only incompatible with the claim that there is some morally relevant fact that it would be impossible for anyone outside of a particular group (to which the benevolent observer would not belong) to appreciate. I think this much stronger claim is far from likely.

3.2. Rationality

The benevolent observer is rational. In the sense in which I wish to use the term, the benevolent observer is both formally and instrumentally rational. The observer is formally rational in that their reasoning is closed under deduction. For instance, if the observer knows that an action will bring about 100 units of welfare to one individual and 100 units to another, they also know that an action will bring about 200 units of welfare overall.[15]

As it is usually defined, one is instrumentally rational when one desires the best means to one's end, where 'best' indicates that the means is most efficient, or cost-effective. But in this case, merely desiring the best means is not sufficient. One could have *some* desire for the most efficient means whilst also having the *strongest* desire for the least-efficient means of achieving the desired end. The benevolent observer must be instrumentally rational in that the strength of their desire for any means is proportional to the efficiency of that means to achieve their desired end.[16] So, the observer's strongest desire is for the most efficient means to their desired end, their next strongest desire is for the next best means, and so on.

[15] As such, the observer will also be able draw infinite conclusions.

[16] And, as we will see, the observer only has one end.

Why must the benevolent observer be instrumentally rational? Consider a benevolent observer who was not instrumentally rational. This observer would be either instrumentally irrational or instrumentally arational. Let us say that an instrumentally irrational observer is one for whom the strength of their desire for some means is inversely propositional to the effectiveness of that means for achieving their end. Thus, their strongest desires are always for that which would best frustrate their ends. An ideal observer who was instrumentally irrational would, therefore, always desire agents perform actions to achieve goals benevolence would not recommend. To act in accordance with the desires of this observer would, by definition, fail to achieve that observer's ends. We can safely assert that the benevolent observer is not instrumentally irrational in this way. An instrumentally arational observer is one who is neither instrumentally rational nor instrumentally irrational. Some of their desires may be for that which would frustrate their ends, whilst some may be for that which would promote their ends, and the strength of these desires will bear no tight correlation to the effectiveness of the means. Since the arational observer will sometimes fail to desire what would be the best means to achieve their benevolent ends, the benevolent observer cannot be instrumentally rational. Since these options (instrumental rationality, irrationality and arationality) are mutually exclusive and exhaustive, the benevolent observer must be instrumentally rational.

Instrumental rationality is sometimes contrasted with what we might call *substantive rationality*. To be substantively rational one must act in accordance with the substantive reasons. The benevolent observer is not substantively rational.[17] But it is appropriate to note how the benevolent ideal observer theory is able to accommodate some of the key intuitions that motivate theories of substantive rationality. A particularly famous case, due to Derek Parfit (1984), is as follows:

> *Future Tuesday Indifference*: John is just like you and me in desiring to avoid pain. However, unlike us, John is indifferent to pain when it occurs on a future Tuesday. If there is a dental procedure which John could undertake next Monday and experience mild discomfort, or next Tuesday and experience unbearable agony, John will elect to undergo the procedure on Tuesday.

According to certain non-substantive views of rationality and some subjectivist views of morality, John only has reason to do that which would best promote his aims. That is to say, given John's aim of avoiding pain, except for that which will occur on a future Tuesday, he has no reason to prefer the overall less painful procedure to the more painful one. In fact, given the mild discomfort John will experience on Monday and his desire to avoid such discomfort, he has most reason to prefer the agony that will take

[17] The reason is that I doubt that there are substantive reasons of this kind, and one does not require the existence of such reasons in order to defend BIO_{RD} as a theory of moral value. Thus, they are surplus to requirement.

place on Tuesday. Parfit sees this conclusion as absurd. John has overwhelming reason to avoid the operation on Tuesday, regardless of his present aims. The reason is constituted by the excruciating pain the Tuesday procedure would cause him. So, Parfit concludes, John has reasons not given by the theory of instrumental rationality.[18]

BIO$_{RD}$ avoids the sting of this example. John's aims are not relevant to determining whether or not he has a *moral* reason to perform a certain action. In section 5.2 I discuss in great detail exactly how BIO$_{RD}$ provides us with a theory of moral reasons. For now, let us say that because the benevolent observer would most desire that John undergoes the operation on Monday as opposed to Tuesday, there is a moral reason for John to prefer treatment on Monday as opposed to Tuesday, regardless of John's desires.[19] In short, even though the benevolent observer may be only instrumentally rational, and not substantively rational in Parfit's sense, there may still be moral reasons that agents have regardless of their present desires. Since this is sometimes taken to be impossible

[18] For an argument that John doesn't have reason to avoid the operation on a future Tuesday, see Street (2009).

[19] It may sound odd to say that John has a *moral* reason to avoid pain. Isn't what's at stake whether or not John has a prudential reason to avoid this pain? Although it may sound odd, it is not difficult to see why this might be. Moral reasons, being impartial, apply to others but also to ourselves. It is perfectly possible for there to be a moral reason and a prudentail reason for the same action. In cases where only my own welfare is at stake, we will more often speak of prudentail than moral reasons, but this does not preclude the existence of an additional moral reason.

without adopting a theory of substantive rationality, BIO$_{RD}$ undermines one motivation for accepting such a theory.

Nevertheless, proponents of substantive theories of rationality will insist that this is no threat. The reasons with which these theories are concerned are reasons *simpliciter*, lacking any qualifier like 'moral' or 'prudential'. A theory of moral reasons may assert that there are moral reasons for actions regardless of one's present aims, but it may be that there is no reason to act in accordance with one's moral reasons unless doing so already coincides with one's aims, or one is failing to act (substantively) rationally. **BIO$_{RD}$** is not a theory of reasons *simpliciter*, nor do I offer one in this dissertation. The primary reason is that I doubt that there are reasons *simpliciter*, but I cannot offer a full defence of that claim here.

3.3. Benevolence

The benevolent observer is, of course, benevolent. But what is benevolence? As I will understand it, benevolence is an impartial final care directed towards the welfare of conscious creatures.[20] By 'final care', I mean to indicate caring about something for its own sake. Thus, the benevolent observer cares, impartially, about the welfare of

[20] For an alternative definition of benevolence in terms of desire-satisfaction, see Sainsbury (1980).

conscious creatures for its own sake. In what follows, I discuss each component of benevolence.

3.3.1. Benevolence as care vs. desire

I wish to suggest that benevolence is, most fundamentally, a care. There are many competing conceptions of care, particularly within the feminist ethics of care tradition.[21] What tends to be agreed on, however, is that care is a richer and more nuanced attitude than desire. For instance, Nel Noddings defines the attitude of care as, at bottom, "engrossment" (Noddings 2013, p. 17). Stephen Darwall also makes this point clear:

> "Caring for someone involves a whole complex of emotions, sensitivities, and dispositions to attend in ways that a simple desire that another be benefited need not. If someone about whom I care is miserable and suffering, I will be disposed to emotional responses, for example, to sadness on his behalf, that cannot be explained by the mere fact that an intrinsic desire for his welfare is not realized. Taken by itself, all that would explain would be dissatisfaction, disappointment, or frustration." (Darwall 2002, p. 2)

Here, Darwall is focusing on the more specific attitude of caring for an individual, rather than on care more generally. And, in fact, there is some dispute about whether or

[21] For an overview, see Held (2005, pp. 29 - 43).

not one can only care for individuals or one can also care about the realization of more abstract states, such as welfare, or justice or fairness.[22] For now, let us assume that it makes sense to speak of caring about welfare. One can care about a certain object and not merely desire it, or its continued existence, and this care renders one liable to a wider range of emotions as a result. Benevolence is a care directed towards welfare.

I think we can say at least two important things about this kind of care. The first is that caring about welfare involves some imaginative exercise, at least at some time in one's past. In the case of caring about someone in particular, it is clear that the imaginative exercise is putting oneself in their place. To care about welfare more generally, one must have imagined what it is like to have certain experiences that one has not had oneself, without imagining being anyone in particular.[23] And, secondly, caring about welfare opens one up to a range of emotional responses in the same way that Darwall described. It is difficult, and I think not necessary, to give an exhaustive list of the range of emotions to which care renders us liable. BIO$_{RD}$ itself is not greatly affected by the conception of care we adopt. Yet, there is a minimalist alternative worth discussing:

[22] I argue that it is appropriate to care about such states in section 4.3.

[23] Similarly, caring about injustice might involve imagining what it is like to be the victim of some unjust act.

Benevolence as desire: benevolence is an impartial desire directed towards welfare for its own sake.

Arguably, thinking of benevolence as a desire fulfils the theoretical aim of simplicity better than the conception of benevolence as care. Care, after all, is a more complex attitude than desire. However, as I argued in section 2.2, our construction of the ideal observer is not just governed by simplicity. A crucial aim is to preserve as much of what we value (or what the conceptual ethicist himself values) in our moral reactions as possible (without keeping what is unnecessary). It seems to me that we would care, not just about what a better version of ourselves would desire with respect to our welfare and the welfare of others, but how their broader emotional sensitivity would be impacted by changes in welfare too. Removing these reactions from the observer, and hence from the set of relevant attitudes, may be an unnecessarily drastic revision. I did state in section 2.2 that simplicity should not be pursued fetishistically out of any over-bearing love for Quinean desert landscapes. The desire view tends in this direction and one may wish to reject it for this reason.

However, I will not take a stand on which view of benevolence we ought to adopt. My own inclination is to accept a richer view of care because this preserves more of the

concept that I care about without, I think, violating the theoretical aim of simplicity.[24] And how much of our concept these rival conceptions of benevolence preserve, is, I want to suggest, the only method which we could use to adequately decide the matter.

3.3.2. *Final* cares

The ideal observer's benevolence is a final care. An agent has a final care directed towards an object when they care about that object for its own sake. Caring for its own sake is to be distinguished from caring about something for the sake of something else. A non-final care is the sort of care a Formula 1 driver might have for going fast: their final care is directed towards winning and they care about going fast as a means to this end. Non-final cares are not, however, always instrumental cares. Schroeder and Arpaly (2013) have pointed out that some desires are neither final, nor instrumental, but 'realizer' desires. An artist may have a final desire that the world contain beautiful things and, as such, may desire to paint beautiful landscapes as a way of realizing their final desire. There may also be realizer cares of this kind. Regardless of how we choose to divide up non-final cares, final cares should be a relatively familiar psychological state for healthy adult humans. Indeed, caring about something for its own sake (other than oneself) may be crucial to our psychological well-being.

[24] The reader must wait until section 5.1 for more detail about how a richer conception of care affects our understanding of moral value.

An adequate explanation of why the benevolent observer must have a final care directed towards the welfare of conscious creatures and not a non-final care requires further detail, to be given in subsequent sections of this chapter. The general strategy, however, is to argue (i) that the observer should only have final care(s) directed to that which is of essential value and (ii) that welfare is the only essential value[25]. The second part of this strategy is presented in section 3.4. Now is an appropriate place to outline the reasons for (i).

In devising our ideal observer, we may choose whether to specify a final care in the first place. If we do not specify a final care, then the observer will either be allowed to devise their own final cares, or, if they do not, will fail to generate any non-final cares. Non-final cares, I suggest, are dependent on final-cares, since non-final cares are always ways of bringing about what one finally cares about.[26] Non-final cares are, in the broadest sense, cares for something 'for the sake' of something else. If one lacks any final cares, then there is nothing 'for the sake of which' one could have a non-final care. An ideal observer who had no final or non-final cares would be, at best, a strange theoretical

[25] For reasons given in section 3.4, I prefer 'essential value' to 'intrinsic value.'

[26] The phrase 'ways of bringing about' is deliberately all-encompassing. As I suggested above, instrumental cares do not exhaust the range of non-final cares; realizer cares are also an example. However, there may be other kinds of non-final cares. This question doesn't concern or much effect the present argument, so I wish to leave the matter open. The important claim is that one would not have any non-final cares in the absence of a final care.

posit. Thus, if we do not specify a final care, there is no reason to believe that an ideal observer will, once we have specified their other attributes, generate final cares of their own unless we specify some further trait. Firth's theory is an example of this approach. Firth does not provide his ideal observer with any final cares or desires, yet he does state that his observer is 'otherwise human' and so we can presume that Firth's ideal observer will generate final cares much as you or I do, bearing in mind their other idealised attributes.

There are, I suggest, two primary problems with Firth's approach. The first is that arguments in favour of psychological minimalism tell against the inclusion of an 'otherwise human' constraint. Those arguments, and their relation to the present discussion, must wait until section 3.5. The second argument was first suggested by Richard Henson (1956). The problem Henson identified is that, if the observer is 'otherwise human', they may be a poor arbiter of moral matters, in either lacking the typical 'moral' reactions we would want the theory to capture, or, even worse, in desiring their opposite:

> "What bothers me is that I can see no reason to suppose that an observer of the sort Firth mentions would approve, say, of people's being happy, other things being equal, rather than of their being unhappy; and yet, if the ideal observer's reactions were to determine the very meanings of ethical

> predicates, we should surely have to constitute him so that he would take a stand on happiness." (Henson 1956, p. 393)

On Firth's behalf, we might reply that since people (typically) do approve of others being happy, won't an otherwise human ideal observer? Perhaps not, for there are also individuals who take a general pleasure in the misery of others, and many more take great pleasure in the misfortunes of their enemies. Neither can we forget that there are psychopaths who take no pleasure or pain in either the happiness or suffering of others. Firth is curiously silent on the question of how many observers there are, but if we assume that any agent who meets his criterion is an ideal observer, then sadists and psychopaths must be among the ideal observers whose reactions we ought to take into consideration.

The easiest way for the Firthian to meet this concern is to insist that 'otherwise human' means that the observer must be otherwise 'normal', that is to say, non-sadistic or psychopathic. But this response is also unsatisfying, for the extension of the term 'normal human being' is ever-changing. Though we might insist that sadists and psychopaths are not, at present, 'normal', if, by some freak accident, all other humans are wiped out and only the sadists and psychopaths are left behind, then these individuals become the norm, and an observer who is 'otherwise human' will begin to look at best amoral and at worst immoral. Our theory of moral value should not be subject to changes in human population.

A more sophisticated response is to specify that the 'otherwise human' condition refers to statistically normal humans and then rigidify the reaction to the one that is statistically normal at the actual world at the present time.[27] Though this might adequately rule out the possibility of amoral or immoral ideal observers, it remains unsatisfactory for at least two reasons. Though it is a common feature of response-dependent accounts of value to rigidify the reactions of the relevant agents to those that they have at the actual world, this saddles value with a certain amount of contingency that is in no way fatal, but nonetheless undesirable. Even when we rigidify, we rigidify the term 'statistical normal human', but the reference of that term may well have been different. Lewis articulated this anxiety well:

> "We might have been disposed to value seasickness and petty sleaze, and yet we might have been no different in how we used the word 'value.' The reference of 'our actual dispositions' would have been fixed on different dispositions, of course, but our way of fixing the reference would have been no different. In one good sense – though not the only sense – we would have meant by 'value' just what we actually do. And it would have been true for us to say 'seasickness and petty sleaze are values'. The contingency of value has not gone away after all; and may well disturb us. I think it is the only disturbing aspect of the dispositional [i.e. response-dependent] theory." (Smith, Lewis and Johnston, 1989, pp. 132-3)

[27] See Wright (1988) and Lewis in Smith, Lewis, and Johnston (1989)

Secondly, we must consider why we would want to preserve the responses of ordinary human beings to begin with. It seems that the only reason is that there must be something important about the psychology of healthy adults which is lacking in that of a sadist or psychopath. But as soon as we acknowledge this we have implicitly admitted that there is something about the character traits of otherwise ordinary humans that is vital to providing the correct account of value. The question has then shifted. Once we accept that there is something important about normal human beings required to make the ideal observer analysis at all plausible, we need only ask what this attitude(s) could be? I suggest that it is a final care directed towards welfare.

I do not, however, think that this consideration is conclusive. The anxiety about this line of thought that Lewis articulated is just that: anxiety. It is not a decisive argument.[28]

3.3.3. The circularity objection

If we do not allow the observer to generate their own final cares (and they cannot have no final care) then we must specify what their final cares are. Since, as I will argue in section 3.4, welfare is the only essential moral value, the ideal observer's only final care should be directed towards welfare. This line of argument may already seem troubling: the reactions of the ideal observer were supposed to determine moral value,

[28] After all, Lewis accepted a dispositional theory of value despite his anxiety.

yet, in specifying the attributes of the ideal observer, I have relied on the claim that something is morally valuable. Indeed, it is because the observer has a final care towards welfare that welfare is valuable in all possible worlds and thus essentially valuable.

I'll pre-emptively pause here to consider this objection which has been made in similar contexts by Harrison (1971), Wright (1988) and Enoch (2005).

Harrison puts the point against Firth, and for its clarity and vividness, is worth quoting at length.

> "Why should we suppose that an ideal observer should be all these things [dispassionate, consistent, etc.]? Is it not because we approve of these characteristics, and think that it is wrong not to possess them? But if this is why we think that an ideal observer ought to be consistent, and so on, then we are supposing that we already know what is right and what is wrong, which is just what we are not supposed to know until we have discovered, by whatever method turns out to be appropriate, what an ideal observer would approve of. In other words, though the feelings of an ideal observer are supposed to be the ultimate court of appeal on moral matters, we are now rigging the election, so to speak, in such a way that if the ideal observer turns out not to approve of what we already approve of, we will redefine the expression 'ideal observer' until we find a definition which is such that something which is an ideal observer, in this sense, does approve of what we approve of already." (Harrison 1971, p. 154)

Enoch objects that all theories which claim that value is determined by the responses of a particular subject cannot motivate the idealization of said subject. He writes that what's needed in order for any response-dependent idealizing theory to be attractive "is some rationale distinct from its purported extensional adequacy." (Enoch 2006, p. 767). Similarly, Wright argues that

> "… the extension of the truth-predicate among ascriptions of moral quality may not be thought of as determined by our best beliefs. […] The reason, as with judgements of approximate shape, is because whether such a belief is *best* depends on antecedent truths concerning shape/moral status." (Wright 1988, p. 24)

The complaint, then, is that in arguing that the ideal observer should have some final care for an essential moral value, the attitude one decides upon will be responsible for determining the extension of value predicates. This is not only theoretically unsatisfactory, but viciously circular. The reactions of the observer were intended to *determine* what is valuable, but we are using our prior judgements about what is valuable to determine the nature of the observer and thus the extension of moral predicates.

This is a serious worry, but it can be dispelled. First, we need to be clear about the way in which BIO_{RD} is circular, if at all. To refresh the reader's memory:

BIO$_{RD}$**:** an object, *x*, is morally valuable if and only if and because it elicits a non-truth-oriented attitude, *R*, from a properly informed, rational, benevolent and otherwise minimal observer, *S*.

Neither the term 'morally valuable' or any if its cognates appears on the right-hand side of this biconditional, so BIO$_{RD}$ cannot be circular in the straightforward sense in which the analysandum appears in the analysans. The problem, then, appears to be the circular way in which BIO$_{RD}$ must be argued for or motivated, as Enoch explicitly says. Yet the project undertaken here, as outlined in chapter 2, is one in conceptual ethics. Enoch himself admits that the circularity objection does not apply to revisionary accounts, and thus as Sobel (2009) points out, "[g]iven that many prominent champions of such views have offered their accounts in just such a spirit, this does diminish the strength of Enoch's conclusion." (p. 342)

The reason that such an objection does not apply to revisionary accounts is that the revisionist is permitted to outline a theory and, so long as it is absent any internal contradictions or incoherence more broadly, it may be judged on its merits alone. This judgement is made on the basis of considerations outlined in section 2.2. In short, how well does the theory on offer preserve my concept of moral value whilst also complying with the relevant theoretical functions? That its motivation is somewhat circular is

relatively unimportant, when compared with how the concept can be put to work after its formulation.

Following Brower, we can distinguish between two ways in which an extension (in this case, the extension of 'moral value') can be predetermined:

> "In one sense, an extension is predetermined when the facts that determine what is in the extension are completely distinct from any facts about how we would respond. In another sense, an extension is predetermined if we have beliefs about what is in the extension and we use those to guide our theory." (Brower 1993, p. 227)

It would be lethal to BIO$_{RD}$, and other ideal observer theories, if facts about value were distinct from facts about responses. But the second sense in which an extension can be predetermined is innocuous, at least for revisionary accounts. It is a given that we have a wide-range of pre-theoretical first order moral judgements and commitments. It is not objectionable to use those commitments to determine the broad scope of moral theory. That is to say, our judgements and commitments determine, broadly, the subject matter of morality. The revisionist must be careful here. After all, they do not much care about discovering the ordinary concept, and may be sceptical that such a concept exists. Yet, even the revisionist must start somewhere, and I am starting somewhere that is recognisable, by at least some members of a common conceptual community, as sufficiently similar to the concept used by others. For example, an ideal observer who

cared only about destroying red barns in the Ohio countryside, the Argentine Tango or the number of blades of grass in Mongolia, would be a terrible theoretical posit. This is because morality isn't about destroying red barns in the Ohio countryside, or the tango, or grass. Broadly speaking, it's about our conduct towards others, happiness, justice, virtue, etc. A good moral theory, even a revisionary one, should account for this broad shape of our discourse; that is, the fact that morality concerns itself with welfare, justice and the like and not with the quantity of Mongolian grass. If, at the other extreme, we posit an observer who has all and only the reactions we endorse in our first-order 'ordinary' discourse, this theory would be objectionably *ad-hoc*. Its claim to be revisionary would also be suspect. The reasonable position is to adopt an account which uses our pre-theoretical beliefs to determine the broad subject matter of morality but once, through our commitment to naturalism or by some other means, we settle on a response-dependent theory and are moved by the considerations above to include some final care on behalf of the appropriate subjects, the nature of the attitude should capture as much of those first-order beliefs as possible, but also discard the others in a principled way, whilst accounting, if possible, for their apparent force. This is what I take myself to be doing. Nothing about this procedure is *ad hoc* or objectionably circular.

A more direct route to this conclusion is the (perhaps cynical) observation that many moral theories that aren't response-dependent in nature engage in a similar kind of circular reasoning, often without being clear about it. For example, evolutionary

debunking arguments show that one can explain the moral judgments we have without reference to moral truth, and such arguments undermine the notion that we have a reliable ability to track a mind-independent moral reality (Street 2006). In the face of these arguments, some still insist that we have access to mind-independent moral truths and that these truths just so happen to be the very same truths evolution would predict that we believe (Enoch 2010). Thus, here is a clear case of us being strongly psychologically motivated for evolutionary reasons to endorse a particular moral position whilst claiming that the reasons we have for holding that position have nothing to do with our being motivated to endorse it. We do not directly perceive any mind-independent fact such as the wrongness of killing, but rather our contemplation or, if we are unlucky, observation of different killings motivates us to posit the existence of a mind-independent rule that forbids them. Is it wrong to use our feelings about particular cases of killing to guide us when considering the principle that 'killing is wrong'? I think not, but it is difficult to see why the response-dependence theorist should not be entitled to use first-order judgments when motivating their theory any more than a mind-independent theorist should.

3.3.4. Impartiality

The final component of welfare, as I define it, is impartiality. Impartiality denotes the fact that the benevolent observer does not deem the welfare of any one individual more

or less important than that of another, simply because one individual is that particular individual. The benevolent observer may well desire, on any given occasion, to bring about a state of positive welfare for some individual rather than another (if the welfare of the first would outweigh the second, for example), but they cannot desire this merely because the first individual is who they are. All welfare subjects must be given equal consideration.

Why must the benevolent observer's final care be impartial? Impartiality and morality have long been seen as intimately connected. Indeed, according to some (including the author), it is constitutive of the moral point of view that it is impartial. This is why impartiality is the second general theoretical aim specified in section 2.2. It's what distinguishes moral value from prudential value, for example. The claim that benevolence must be impartial is stipulative, in that it defines the subject-matter.

Arguments against the impartiality of moral value tend to focus on the importance of partial cares and concerns. Typical examples include the moral value in a parent's love for their child, or our concern for the welfare of our friends over that of strangers. These values are often wrapped in the language of duties: parents and friends, it is said, have moral duties towards their intimates that they lack towards strangers. The reader must consult chapter 5 for a full articulation of the relationship between BIO$_{RD}$ moral obligations, but suffice it to say, the response will be sufficiently similar to that given in the consequentialist tradition. BIO$_{RD}$ will, ultimately, suggest that the moral point of

view is impartial, but that the desires of an impartial agent are such as to recommend partial cares and concerns for particular agents. Despite misplaced accusations of self-defeatingness (Stocker 1976), it is a familiar and perfectly coherent consequentialist refrain that the best consequences are often not brought about by caring directly about bringing those consequences about (Railton 1984). Instead, the world will tend to be better, on the whole, if parents develop partial cares towards their own children. Although the ideal observer's final desire for the welfare of others is impartial, it may lead them to desire social arrangements in which individuals have final yet partial cares directed towards intimates.

3.4. Essential value and response-dependence

Before I argue that welfare is the only essential value, it is worth pausing to consider the concept of essential value and its relationship to response-dependent theories. What I am calling 'essential value' has often been discussed under the heading 'intrinsic value.' Moore famously gave the following characterization of intrinsic value:

> "In order to arrive at a correct decision [on what has intrinsic value], it is necessary to consider what things are such that, if they existed *by themselves*, in absolute isolation, we should yet judge their existence to be good." (*P*, Chapter VI, Sec. 112)

Another common characterization of intrinsic value is that it provides an end for chains of instrumental (or sometimes 'extrinsic') value.[29] Some things seem to matter because of how they relate to other things. My computer, for example, matters insofar as it helps me to be productive and provide entertainment. My productivity and entertainment in turn might matter because they improve or are components of my welfare. But my welfare, it seems, does not matter for the sake of something else in this way. My welfare matters *for its own sake*. This is sometimes expressed by saying that my welfare has intrinsic value, at other times that my welfare has final value. Kagan (1998) distinguishes between final value and intrinsic value arguing that some things might matter for their own sake yet not in virtue of their intrinsic properties. In the course of defending this distinction, one thing Kagan notices is that intrinsic properties and relations of objects can be contingent, so it is not clear why we should attach any special weight to 'intrinsic' value, *per se*.

> "Some, I suppose, might be tempted by the claim that value based on intrinsic properties alone is a kind of value that an object has necessarily. And necessary value would, I grant, be an interesting type of value to study [...] the tempting thought is mistaken: since intrinsic properties need not be had necessarily, value based on intrinsic properties alone need not be possessed necessarily." (Kagan 1998, p. 290)

[29] Korsgaard (1983) famously rejects the dichotomy between intrinsic and instrumental value.

Kagan has, I think, hit upon something important. Intrinsic features of objects can be had contingently. Thus, if we talk of the intrinsic value of some object, we must allow that some objects can have intrinsic value without having that value in all possible worlds. Yet, the thrust of Moore's thought experiment, and the importance of the notion of a final value is, I suggest, that of the value something always has, no matter what else may be the case. It is this sort of value which should be determined by the ideal observer's attitudes. Thus, I will speak of essential moral value instead of intrinsic moral value. Let us say, intuitively, that an object has essential value if and only if it is valuable in all possible worlds.[30]

I assume that objects of moral value either have essential moral value or their value depends upon some other object that does have essential moral value. This is another version of the familiar thought that all objects either have intrinsic value or instrumental/extrinsic value in which case their value depends upon something with intrinsic value.[31]

Now that we have a clearer picture of what essential value is, we must consider a potential problem for a response-dependent theorist relying on the notion. It is widely held that intrinsic value (and, we can suppose, essential value too) is non-relational.

[30] We could equally well call this necessary value.

[31] This may be disputed by those, such as Kagan, who think that those things which lack intrinsic value might nonetheless be of final value.

Therefore, the argument goes, essential value cannot depend on anyone's psychological states:

> "…the intrinsic value [read: essential value] of a thing is not dependent on its being the object of any psychological attitude. If a thing has intrinsic value, it has it independently of its being the object of any psychological attitude or its being conducive to or productive of any such attitude. If a thing has intrinsic value, it does not have that value because, or in virtue of, its being the object of anyone's psychological attitude or because it would be the object of such an attitude under some set of hypothetical conditions." (Lemos 2005, p. 19)

Recall that any response-dependent theory of value makes the following general claim:

RD: *x* is valuable if and only if and because *x* elicits *R* from *S*

where *x* is an object of value, *R* is a relevant reaction and *S* is an appropriate subject. The 'because' claim in RD would seem to suggest that there can be no essential value since all value is relational. Whatever gloss we give on the meaning of 'because' in RD, it is clear that it has the following implication:

(*r*) if *S* did not have *R* towards *x* then *x* would not be valuable.

I want to suggest that some essential value can be relational. Whether or not a

particular kind of value is essential and relational depends on our assessment of the counterfactual (*r*).

Consider a response-dependent theory, not of value, but of 'coolness'. Let us say that something is cool if and only if and because 20-somethings living in Brooklyn say that it's cool. In the actual world, 20-something Brooklynites say that portable record players are cool, and thus it is so. But a possible world in which they did not call portable record players cool is one in which portable record players simply are not cool. Consider, then:

> (r_{coolness}) if 20-something Brooklynites did not have the 'cool' response towards portable records players (in independent coffee shops) then portable record players would not be cool.[32]

Surely, there are plenty of nearby possible worlds in which 20-something Brooklynites lack the cool response towards portable record players. At these worlds, they are not cool. The counterfactual (r_{coolness}) is non-trivially true.

Now re-consider:

> BIO$_{\text{RD}}$: an object has some moral value if and only if and because it elicits a non-truth-oriented attitude from a properly informed, rational, benevolent and otherwise minimal spectator.

[32] The example is inspired by Pettit's (1991) discussion of 'U-ness'.

And consider also the counterfactual (r) with its ideal observer theory gloss:

(r_{BIORD}) if a properly informed, instrumentally rational, benevolent and otherwise minimal observer did not have the relevant non-truth-oriented attitudes towards welfare, then welfare would not be morally valuable.

There is an important difference between (r_{BIORD}) and (r_{coolness}). Whereas there are possible worlds in which the antecedent of (r_{coolness}) is true, there are no possible worlds in which the antecedent of (r_{BIORD}) is true. There are possible worlds in which Brooklynites don't have the 'cool' reaction towards portable record players but there is no possible world in which the benevolent ideal observer does not have the relevant response towards welfare. In the case of welfare, the relevant response is a final care. So, (r_{BIORD}) asks us to imagine a benevolent observer who did not have a final care directed towards welfare. But to be benevolent is to have a final care directed towards welfare. An agent cannot both have and lack a final care directed towards welfare, so there is no possible world at which the antecedent of (r_{BIORD}) is true.[33]

[33] There is disagreement about how to evaluative counterfactuals with necessarily false antecedents. According to Lewis (1973), all such counterfactuals are trivially true. According to Nolan (1997) and others, these counterfactuals form a distinct and interesting class of their own, counterpossibles, which can be substantively true or false when evaluated at impossible worlds.

When response-dependent theories are such that the conditional (*r)* has a necessarily false antecedent, we can say that the object in question has the relevant property the response-dependent theory is a theory of, and that it has that property essentially. This is because, in all such cases, the relevant agent(s) *S* will always have reaction *R* towards *x*. Thus, the object will have the property in all possible worlds, which is just what it is to have a property essentially.

It is often thought that intrinsic value could not be relational, as suggested by Moore's isolation test and explicitly stated by Lemos (2005). One may, therefore, initially think the same of essential value. But, the foregoing considerations show how welfare might be of essential moral value yet only have its value relationally. According to BIO$_{RD}$, it is true, in every possible world, that the benevolent observer would have a final care directed towards welfare. Also, according to BIO$_{RD}$, this is what makes welfare valuable. Thus, welfare is valuable in every possible world, which is just to say that welfare is of essential value, despite its value being relational.

3.4.1. Welfare as essentially valuable

In this section, I discuss why the ideal observer's only final care ought to be directed towards welfare, which is tantamount to defending the view that welfare is the only essential value. Sumner coined the term 'welfarism' to denote "…the view that nothing

but welfare matters, basically or ultimately, for ethics" (Sumner 1996, p. 184). BIO$_{RD}$ is an explicitly welfarist view. The difference between welfarism as it is normally presented and BIO$_{RD}$ is that BIO$_{RD}$ offers an explanation of why welfare matters: it's what an ideal observer would have a final care towards. The explanation for why the observer has this final care will be the same explanation the welfarist give for why welfare is the only essential (or intrinsic) value. In that sense, the explanation BIO$_{RD}$ gives makes no argumentative progress over welfarism as Sumner argues for it.

As Sumner writes, "[s]ince welfarism is a theory about the foundations of morality, it is difficult to know how to go about defending it." (1996, p. 193). This difficulty is, I think, unavoidable. In fact, I would go as far as to suggest that one cannot present anything like a satisfactory argument for why any particular candidate value is of essential moral value. This goes for common contenders such as freedom and the typical virtues as much as it does for welfare.

To illustrate this point: suppose that someone were to insist that positive welfare is, in fact, essentially bad. What might we say to persuade them that morality concerns itself with (at least in part) promoting positive welfare? We could begin by pointing out that positive welfare, as a matter of fact, does form a central part of our moral discourse. For instance, Coons, another welfarist, writes:

> "[…] the view is already deeply embedded in our moral thinking. Stereotypically wrong acts (e.g. killing, theft, rape, deception, disloyalty, etc.) seem to share one and only one obvious and unique characteristic: they each tend to be bad for their victims. Moreover, the relative seriousness of these sins tends be proportional to how bad these acts are for others." (Coons 2012, p. 207)

We could also give them a number of extreme test cases. We might invite them to imagine a world in which everyone has the worst life it is possible for them to have, but that our interlocutor can flip a switch and instantly make each one of these individuals lead a life with extraordinarily high levels of welfare. Would there be any moral pressure for an agent to perform this act? What if they were simply to insist that there would not be? I do not know what I or anyone else could say to convince such an individual. They do not seem to be making a mistake, but rather just talking about something else entirely. The best response would be, I think, to simply allow them to have their own discourse, call it 'schmorality', in which welfare has negative value, and allow myself not to care about 'schmorality' talk and resume discussion of morality, where positive welfare matters.[34]

[34] The conceptual ethicist must be careful. Of course such a person would be entitled to embark on the project of refining a 'promote pain' oriented morality, but such a project is simply wildly off the mark from how most people seem to concieve of moraliy. Again, the conceptual ethicist need not be totally idiosyncratic.

Sumner suggests that if we were to restrict ourselves to one essential value, then welfare is the only plausible candidate because it is *generic* and *abstract* enough to plausibly cover all cases. A trait such as loyalty or courage, though perhaps an excellent candidate for an essential value in a pluralistic theory, is not a serious candidate for the *only* value because it is too parochial. But here, I part ways with Sumner. There are other values generic and abstract enough that one might not implausibly consider them to be essential values, and one could not simply dismiss them as instances of a different discourse entirely, as one could with someone who insists that welfare is bad. One such view is that freedom or autonomy is the only essential moral value, and that the value of welfare (and everything else) depends upon its relationship to freedom.[35] But this view faces similar problems in certain extreme test cases. As per the previous example, imagine a possible world in which everyone has the same level of freedom but everyone suffers as much as possible. One could flip a switch and, without changing the amount of freedom any one individual has, improve their welfare. Is there any moral pressure to flip the switch? The freedom essentialist must say no.

We could also make the same argument with respect to a non-welfare-pluralist, that is, one who insists that freedom, plus additional notions such as justice or perfection, also have essential moral value, but not welfare. We can take each of these candidate

[35] One can imagine a political libertarian being attracted to such a view.

essential values and apply this same thought-experiment and see if we find any further reason to flip the switch.

We can generalise this test as follows:

> *Comparative essential value test:* For any two candidates for an essentially valuable property, *x* and *y*, imagine a world in which there is some fixed amount of *x* and some minimal amount of *y*. Some trivial action, *a*, can maximise the amount of *y*, whilst holding *x* constant. If there is no moral pressure[36] to perform action *a*, then *y* is not essentially valuable. If there is some moral pressure to perform action *a,* then *y* is essentially valuable. *x* may or may not be essentially valuable.[37]

This test relies on the assumption that if *y* were of essential moral value, then any extra amount of *y* one could bring about or instantiate in the world would be the source of some moral pressure to perform *a.* If *y* is a non-essential value, then the source of its

---

[36] I use the term 'moral pressure', as I did in chapter 1, as opposed to duty or obligation, since, as I will argue in chapter 5, I do not think that there are duties or obligations in anything like the standard sense. If, for example, *a* would make the world better, this is enough for there to be 'moral pressure' to perform *a*, even if one is not obliged to do so. One can think of 'moral pressure' as a disjunctive term designed to cover duties, reasons, obligations, making the world better, etc.

[37] Inspiration for this test is taken from Hutcheson's discussion of the cardinal virtues. Hutcheson alleges that:

> "[…] these four Qualitys, commonly call'd Cardinal Virtues, obtain that Name, because they are Dispositions universally necessary to promote publick Good, and denote Affections toward rational Agents; otherwise there would appear no Virtue in them." *Inquiry*, II.2.i

value depends upon some other valuable object; the presence of *y*, by itself, would be neither valuable nor dis-valuable, in the absence of the typical source of value and so there would be no moral pressure to perform *a.*

I can think of no better intuitive test for essential value than the comparative essential value test.[38] It isolates two candidate values, varies one whilst holding the other fixed, and asks us whether there is any impact on our judgements regarding moral pressure. If there is no impact on our judgement, then candidate *y* cannot be an essential value.

Above, I applied the test to the case of freedom and welfare, where welfare was the independent variable. In that example, one could bring about a state of affairs with the same amount of freedom but drastically[39] more welfare through flipping a switch. If freedom is the only essential value, then there can be no moral pressure to flip said switch since whatever value welfare had would be non-essential. Since, by stipulation, action *a* does not lead to greater freedom, the freedom monist must admit that there is

[38] Moore's isolation test is another example of an essential value test. It passes muster, since it is true that, in isolation, whatever is of essential value will be the only thing(s) valuable. The test does not, however, allow us to directly compare candidates for essential value. I suggest that the comparative essential value test is a better intuitive guide to what is and isn't of essential value.

[39] It is not theoretically important that the increase in welfare be drastic. If welfare, or any *y*, is of essential moral value, then any increase in it will present some moral pressure to perform *a.* The reason for supposing that *y* is as minimal as can be in the first instance, and as maximal as can be in the second, is because I consider it helpful to sharpening our reactions to the cases. The reader may disagree, because they believe that *x* and *y* may not be as independent as I am suggesting. I address this concern shortly.

no moral pressure to flip the switch. Yet, I would suggest that the increase in welfare does provide some moral pressure to flip the switch. Even those who are pluralists about essential value may share this reaction, if they accept welfare as one such value. It is only monists and pluralists who reject welfare as essentially valuable who will not share it. Is there anything that one can say to convince the non-welfare monists and the non-welfare pluralists (that is, pluralists who reject the essential value of welfare)[40] that I am right and they are wrong about this case? I do not think so.

Thus far, I have considered the case where freedom remains fixed and welfare is the independent variable. Now imagine the following: everyone lives a life with a certain amount of positive welfare, but with minimal freedom. One can flip a switch, and bring about greater freedom, with no change in welfare. Is there any moral pressure to flip the switch in this case? As a welfarist, I think there is no such pressure. The freedom-monist and non-welfare pluralist are now in a position analogous to the position I had previously found myself in, with little to say that could convince me otherwise. The welfarist must respond similarly for each case in which *x* is welfare and any other candidate value is the independent variable *y*. For the welfarist, nothing other than welfare can make any difference to whether there is moral pressure to perform action *a*.

[40] I consider the case of pluralists who insist on the essential value of welfare plus additional goods below.

The welfarist's response to the comparative essential value test is rather stark (as are any value-monists'). They insist that, where *y* is anything other than welfare, there can be no increased pressure to perform action *a.* Perhaps other people's responses are less stark. One might accept a pluralistic theory according to which both welfare and freedom (and perhaps some additional candidates) are all essential values. For instance, compare possible responses in the cases where *x* is welfare and *y* is freedom, and then vice-versa. One might think that, in both cases, there is some pressure to perform action *a*. If this is one's reaction, it appears that, once again, we have reached a stalemate. I can only report that I do note share this reaction, as I assume is the case with fellow welfarists.

It is perhaps worth pausing to discuss a metatheoretical problem that arises in light of this stalemate. Again, suppose that the welfarist always finds that additional welfare creates moral pressure in the comparative essential value test, and that the pluralist (one who accepts welfare as essential valuable) finds the same to be true with respect to other candidate values (in addition to welfare). Furthermore, suppose that both accept the comparative essential value test as a good guide to what is essentially valuable and have reflected on it sincerely and thoroughly yet still arrive at different conclusions. What then?

Consider the following remark from Lewis:

> "But when all is said and done, and all the tricky arguments and distinctions and counterexamples have been discovered, presumably we will still face the question of which prices are worth paying, which theories are on balance credible, which are the unacceptably counterintuitive consequences and which are the acceptable counterintuitive ones. On this question we may still differ. And if all is indeed said and done, there will be no hope of discovering still further arguments to settle our differences." (Lewis 1983, *x*)

If the welfarist and the pluralist really do disagree, then I have been suggesting that we are at the hopeless stage that Lewis describes. But, a third party might insist, in that case why believe either theory? Why not suspend judgement between welfarism and pluralism in light of irreconcilable disagreement? This third party does seem to have a point in the case where that third party themselves has no clear reaction when reflecting on the comparative essential value test. Their lack of a clear reaction means that the test fails to constitute evidence for them either way, and so they must rely on the differing testimony of what we can suppose are 'epistemic peers' (Christensen 2009).

But should the fact of persistent peer disagreement move the welfarist or the pluralist, both of whom we can assume *do* have clear reactions to the comparative essential value test? Following Elga (2007), let us call the thesis that both parties should move closer towards their interlocutors position the *conciliatory view* and the rival position, that both parties should remain confident in their initial judgement, the

steadfast view. The literature on peer-disagreement is vast and, although my sympathies lie with the steadfast view, I cannot fully argue for that position here.[41] However, one can note that, in the case of the comparative essential value test, the steadfast view appears to be the more intuitive option. The steadfast view is most appealing in cases where both parties have thoroughly examined all available evidence, perhaps several times, and come to same conclusion each and every time. There is nothing that the pluralist can say to the welfarist about the relevant facts to persuade them to change their mind – they have already considered those facts! It is similar to a case in which I am convinced that the solution to an equation is x and you are convinced that the solution is y (and $x \neq y$). I have checked my working out several times, using multiple calculators and other arithmetical methods, and every time I have arrived at x. It does not seem plausible to suggest that I should be any less confident in my answer upon having discovered that that your answer is y.

I think this leaves both the pluralist and the welfarist in the Lewisian stalemate described above. This is part of what I meant when, in section 1.3, I echoed Wittgenstein's thought that the only people who would understand (or in my case, agree) with his work were those who had already had the thoughts it contained. The

[41] For an overview of this literature, see Christensen (2009) and Frances and Matheson (2018). For a defence of the steadfast view based on the notioin of agent-centered evidence that applies neatly in the case of the comparative essetial value test, see Huemer (2011).

project I am engaged in is an attempt to articulate what I and others find persuasive about a certain meta-and normative ethical position. It may fail to persuade others, but this was not ever its aim, as much as one might perhaps wish it to do so. A third party that lacked clear intuitions either way is not compelled to accept welfarism or BIO$_{RD.}$

However, the welfarist may, on some occasions, be able to find cracks in the pluralist's armour. Suppose one finds when conducting the comparative essential value test that there is *more* pressure in one case than in the other. That is, one may think that when freedom is fixed but one can improve welfare, there is a greater moral pressure to perform action *a* than there is when welfare remains constant and freedom is increased, though, in the latter case, there is still *some* pressure to perform action *a.*

Even though these may be our actual judgements, it is difficult to provide a sound theoretical justification for them. Essential value does not come in degrees. Anything that has essential value either has it or doesn't have it. The amount of moral pressure there is to perform certain actions can vary with the amount of a particular essential value one brings about in performing that action, but if x and y are both of essential value, one would need to provide an argument showing how x and y can both be of essential value, yet one be a greater source of moral pressure, *simpliciter*. That is to say, although the prospect of more freedom or welfare may lead to greater moral pressure to perform an action, one would need to argue that freedom or welfare provided more moral pressure, all else being equal. The problem is that it is not clear what it means to

hold all else equal in this case. Presumably, it would mean that if the same amount of welfare and freedom were capable of being brought about, then there would be greater moral pressure to bring about one as opposed to other. But it is not clear what, if anything, it means to say that one could bring about 'the same amount' of welfare and freedom. The relative quantity of rival candidates for essential value does lend itself to easy conceptual analysis. In the absence of such an analysis, the intuition that there may be more pressure to perform *a* in one instance of the comparative essential value test than another does not have a sound theoretical basis.

On the other hand, the welfarist can explain why we might in fact have such theoretically unsound intuitions. The explanation is an instance of what Sumner calls 'co-option'. Since, the welfarist argues, freedom typically leads to an increase in welfare, we may come to value freedom for its own sake, mistakenly identifying its frequent causal connection to welfare as essential value. If we believe that there would be some moral pressure to perform action *a* where doing so would keep the overall level of welfare the same, this may be because we have mistakenly come to value freedom as if it were essentially valuable because of its actual (and so contingent) connection to welfare. We thus judge that there is some moral pressure to perform *a* when there is not.

Co-option presents a small opportunity for the welfarist to convince an undecided party that welfare is the only essential value. If one conducts the comparative essential value test with welfare and every other plausible candidate for an essential value and

finds that, in every case, one judges that there is a greater moral pressure to perform *a* when *a* would lead to an increase in welfare than when it would increase the amount of any other candidate, then one judges that welfare is the source of greater moral pressure than every other rival moral value. Given that it is likely theoretically untenable to suggest that each one of these candidates is an essential value but every one that leads to a lesser amount of moral pressure than welfare, one must correct for the initial intuition by either (i) revising one's judgements so that there is equal pressure to perform *a* in every case, or (ii) revising one's judgements so that there is no pressure to perform *a* except for when *a* increases welfare, and perhaps accepting Sumner's 'co-option' story as an explanation for why one previously thought that there was some pressure to perform *a* when it lead to something other than welfare. The opportunity is small because one may accept option (i) without contradiction, in which case the dialectic is, once again, at a stalemate.

The state of play is as follows: the welfarist insists that welfare is the only essential value. When conducting the comparative essential value test with welfare as the independent variable *y*, they find, in every case, an increased moral pressure to perform action *a.* On the other hand, with welfare as *x* and other candidate values the independent variable *y*, they find no increased pressure to perform action *a*. They acknowledge that others may not share this response, but invite them to consider that they have been led to judge other rival values as essentially valuable only because of

their relation to welfare. If this is not persuasive, we are at a stalemate. A third party looks on unimpressed.

Before considering objections to the comparative essential value test, it is worth pausing to discuss the role of intuitions. As made clear in section 2.3, BIO$_{RD}$ sits squarely within the metaethical tradition of moral sentimentalism. Sentimentalism comes in many forms, but the kind endorsed here says that moral judgments refer to sentiments, in particular, the responses of an observer borne out of benevolence. The sentimentalist and what we might crudely call 'the rationalist' understand thought-experiments like the comparative essential value test in distinct ways. Whereas the rationalist views intuitions as a guide to a mind-independent realm of moral facts, the sentimentalist is content to say that our (or some idealised version of our) emotions determine value. When conducting the comparative essential value test, the important question for the sentimentalist is not 'what is your intuition about this case?' but 'what would we come to value in that world?'. This is an altogether less mysterious process than is commonly supposed. All we need to consider is our psychology at distant worlds, a difficult task for sure, but not at all impossible.

Bearing this in mind, there is, however, an apparent problem with one central assumption of the comparative essential value test: one may not always be able to vary x and y independently of one another. My ambition so far has been to remain neutral over which particular theory of welfare is true. Suppose, however, that one accepts an

objective list theory and freedom is on this list. After all, it is not implausible to suggest that our welfare increases the more autonomous we are (up to a certain point at least) and this certainly should not be ruled out without further argument. But then it is not obviously possible to separate freedom and welfare independently in the way the comparative essential value test requires. If the overall level of welfare increased but freedom remained the same, welfare could increase due to improvement in other areas on the objective list, yet one would not have shown that freedom was of no essential value since it would remain a component of welfare.

This raises an interesting question with respect to objective list theories and welfarism. According to the most plausible interpretation of these theories, is welfare itself essentially valuable or are the components of the list essentially valuable?[42] If objective list theories demand us to say that the items on the list are essentially valuable, then welfarism cannot be true in the traditional sense and the ideal observer must have a final care directed towards each item on the list.

[42] Of course, one may have a theory of welfare without claiming that welfare, or any of its components, have essential value. I am here just considering how someone who did wish to make such a claim would interpret the objective list theory.

To answer this question, consider the simplest objective list theory: hedonism. According to hedonism, there is one item on the objective list: pleasure.[43] The hedonist welfarist holds that welfare is the only essential value and that welfare is identical to pleasure. Given the identity of welfare and pleasure, the hedonist welfarist holds that pleasure is the only essential moral value. A more complicated objective list theory might have two items on it: pleasure and freedom. According to this two-item objective list theory, are pleasure and freedom identical to welfare? It was, after all, the identity claim of the single-item objective list theory which allowed us to deduce that, according to the hedonist, pleasure is the only essential value. On a two-item theory, neither item is, by itself, identical to welfare. Thus, we cannot infer from the fact that welfare is essentially valuable, and that welfare is identical to freedom and pleasure, that freedom is essentially valuable or that pleasure is essentially valuable. All we can infer is that the conjunction [freedom & pleasure] is essentially valuable.[44] On this view then, objective list theories say that the conjunction of the items on the list is essentially valuable, rather than the items themselves.

---

[43] Depending upon how one concieves of hedonism, there may be an additional item on the list: avoiding pain. We can also understand pain and pleasure as opposite ends on a single spectrum such that increasing pleasure precludes increasing pain. The details do not matter for the purposes of my argument.

[44] Another possibility is that, according to the objectivst list theoy, both freedom and pleasure are, by themselves, sufficient for weflare. In this case, we must still say that the disjunction [freedom ∨ pleasure] is essentially valuble rather than the disjuncts themselves (one cannot infer necessarily p or necessarily q from necessarily (p ∨ q)).

However, there is nothing to say that the objective list theory must be understood this way. Perhaps the best welfarist interpretation of the objective list theory is one according to which all the items on that list are of essential value. However, there are reasons to doubt this theory. The objective list, for one, is a list of things that comprise welfare *for an individual*. The items on that list are good for a particular person, and not intended to be good more generally. Knowledge might be a component of my welfare and good for me, without it being good or bad, morally, that I have it. Throughout, the kind of essential value I have been concerned with is moral value, not prudential. It is possible, therefore, to insist that welfare, promoted impartially, is morally valuable, whereas the components of welfare are only prudentially valuable for the individual in question.[45] This possibility dulls the objection that the comparative essential value test treats as independent things which are not. In the case of freedom, anyone applying the comparative essential value test must be treating freedom as a potential essential moral value. Freedom may be essentially good for me, without being an essential moral value, even as a component of welfare. The welfarist can, therefore, insist that though freedom may be a component of my welfare, the comparative value test can be rightly applied to freedom as a candidate essential *moral* value. As a component of welfare, the closest

[45] And note that only one component of the objective list need be an implausible candidate for an essential moral value, since there would be no reason to insist that some items on the list were essentially valuable and other not.

freedom would get to being an essential value is as one conjunct of a longer conjunction which is of essential value. The same could be said of every other item on an objective list theory which also seemed a plausible candidate for an essential moral good.

With that said, if freedom is on the objective list, in a case where we try to keep welfare constant while freedom is increased, there must be some impact on the overall level of welfare. Therefore, the comparative essential value test cannot apply since welfare must vary in quality if not in quantity with any increase in freedom. Nonetheless, the welfarist may respond by saying that an increase in freedom is only better insofar as it would contribute to our welfare. The comparative value test may not be perfectly applicable, since an increase in freedom necessitates some change in our welfare, but it may still be the case that it's this increase in welfare which is the source of moral pressure to perform a, and not the increase in freedom itself. It is not clear how to isolate this intuition any further.

However, the welfarist can resist even this much. Why think that freedom is a component of our welfare? We can imagine welfare subjects for whom too much freedom makes their lives worse. Children and animals may be an example of this. Freedom appears to be a background non-essential requirement of welfare in typical conditions for adult humans and nothing more. Though my ambition has been to remain as neutral as possible between competing conceptions of welfare, totally impartiality may not be possible. The possible psychological variety of welfare subjects

is so vast that even broad human concerns like freedom may be too parochial after all. Whatever welfare is, it should be the common denominator between all welfare subjects (Lin 2018). A plausible candidate for this state is pleasure. Hedonism is often dismissed as a theory of welfare, but it has far more going for it than its detractors often acknowledge. Nonetheless, this dissertation is not a defence of hedonism and cannot examine every rival conception of welfare and see which ones best fit BIO$_{RD}$. What's clear is that pleasure, as a variable y, is capable of being separated from any other plausible candidate x. It is quite easy for me to imagine that freedom, loyalty and courage, for example, all cause certain welfare subjects to be in serious pain and that, through some trivial action, they might begin to cause immense pleasure. If BIO$_{RD}$ can only be made to fit with a hedonistic theory of welfare, then I will not consider this a problem. However, the category of theories with which the comparative essential value test can be applied most easily is broader than hedonism. Any theory according to which welfare is something essentially mental, consisting, perhaps, of a broad spectrum of psychological states with pleasure being the most basic, will most easily fit the test, since our mental states can be stipulated to vary wildly with any external facts. Perhaps, then, the comparative value test and BIO$_{RD}$ are best suited to 'internalist' theories of welfare. If so, it would remain an attractive ethical position even at the cost of abandoning total neutrality with respect to theories of welfare.

One final concern with the comparative essential value test is that it asks us to imagine something rather bizarre – that a trivial action could have large consequences for the instantiation of certain properties. Of course, certain trivial actions may have important consequences – there are a handful of people in the world who, at a press of a button, could wipe out most if not all life on Earth. But the comparative value test is particularly strange in that it asks us to vary candidate values that are typically co-extensive, and thus it is difficult to imagine a concrete causal mechanism by which flipping a switch, or some other trivial action, could initiate such change. 'How exactly,' one might wonder, 'is flipping a switch supposed to increase freedom whilst keeping welfare constant?'.

The fact that these cases are difficult to imagine should not detract from their significance. All that matters is the outcome. Indeed, the very point of the comparative essential value test is to isolate those elements which are of most concern, and vary them to an extreme degree so that our intuitions are clearer. However, some may find that the comparative essential value test has the opposite effect. The lack of detail prevents them from having any clear intuition about the case, and the test becomes moot.

It is difficult to know how to respond to this worry. As long as the comparative essential value test asks us to imagine what happens at some *possible* world, its opponent is unable to accuse it of being useless as a result of conceptual incoherence. The primary difficulty in attempting to describe these possible worlds is that human beings, as we

happen to be constituted, are such that our welfare is tied to a wide variety of other non-essential goods. We cannot imagine what it would be like for us to have our welfare and our freedom separated, some may insist, because we would no longer be ourselves. But for those of us who believe that the most fundamental facts about value are necessary truths, the fact that we happen to be constituted in a particular way which makes it difficult to separate distinct properties is no reason to think that these properties are theoretically inseparable, or it cannot be that just one of these properties is the only one that matters.

3.5. Minimalism

Finally, the benevolent observer is psychologically minimal. This is to say that aside from the psychological features required to be relevantly informed, rational, and benevolent, the observer has no other psychological traits. This is not to say that the observer will be a psychologically simple agent. The psychology required to exhibit these traits may be extremely complex. This is no violation of the minimalism requirement.

The justification for the minimalism requirement is rather simple: if the observer's final care for our welfare determines moral value, then the observer should not be

infected with other 'junk' attitudes that are not part of this final care and thus have no bearing on moral value.[46]

An interesting question that arises in light of minimalism concerns the number of possible benevolent observers. It is sometimes unclear whether proponents of ideal observer theories are envisaging one ideal agent or several. One might worry that minimalism settles this question in favour or a single observer, but in fact minimalism is compatible with there being multiple observers. Let us say that two ideal observers are distinct if and only if their reaction(s) would be different in at least one circumstance. There is nothing in the view that precludes there being two or more possible observers. However, I believe that we should suspend judgement about how many distinct benevolent, minimal observers there are. I do not believe we can determine an answer with any certainty until we have a better theory of mind and computation.[47] My suspicion is that there will only be one such ideal observer, but this is simply a hunch. What, after all, could cause the difference between the two observer's reactions? Both

[46] Of course, if these additional attitudes had no bearing on the relevant reactions of the ideal observer, then it would not matter whether the observer was minimal. But then these additional traits would do no theoretical work. They are simply best left out.

[47] It is important, however, to note that this complication would not destroy the objectivist credentials of BIO$_{RD}$. A worry such as this was presented to Firth by Postow (1978) and dispelled in Firth (1978). See also section 5.1 of this dissertation.

have a final care directed towards welfare and there appears to be nothing to separate one agent's care of this kind from another.

Minimalism coheres well with the aim of simplicity discussed in section 2.2. Yet, there are reasons against endorsing minimalism. Let us return to Firth's 'otherwise human' condition. One reason to include the 'otherwise human' constraint is to keep the observer as similar as possible to ourselves. We might want similarity in order to feel connected to the ideal observer, and motivated by their reactions. Admittedly, if, when deliberating about whether to φ, one learns that a more informed, recognizable version of *oneself* would be motivated to φ, one may well be more motivated to φ than if one had learned that a benevolent, minimal observer would endorse φ-ing, where such an observer's psychology appears alien to one's own. But, as I stated in section 2.2, BIO$_{RD}$ is not an internalist theory.[48] I simply deny that it is a problem if someone did not care about what the benevolent observer's reactions would be, even if they were reactions pertaining to their own conduct. Although I think that such an individual wouldn't

[48] At least, not on any typical understanding of internalism. BIO$_{RD}$ does not claim that there is any necessary connection between my judging that an ideal observer would desire that I φ and my desiring to φ. However, if one defines internalism, not as the view that *my* judging that it is valuable to φ must motivate *me* to φ, but instead as the view that there is some necessary conceptual connection between value and motivation, then, arguably, BIO$_{RD}$ is an intenralist theory. Desires, and other motivational states, will fall within the range of relevent attitudes (see chapter 4). Given that value is, at least in part, determined by the motivational states of the ideal observer, BIO$_{RD}$ is internalist in this sense, which may be enough to quell the worries of some weak internalists.

threaten the theory, I do believe that if someone did not care *at all* about what a benevolent observer would desire or hope for them, then that individual would be at least bizarre, even if they were not completely irrational. Part of what makes moral sentimentalism an appealing position is that our internal moral sense is, if not universal, possessed by almost all of us. To the individual who appeared not to care about the desires of a benevolent being, one could point out that they too are benevolent in some of their desires, at least those regarding themselves and their intimates. One hopes that such an individual would also notice that the benevolent observer merely extends this benevolence to all others. Perhaps in a calm and cool hour, or behind a veil of ignorance, they would come to see the value in an impartial but loving point of view. Failure to do so would be troubling, but would not indicate a lack of rational agency.

There is, however, perhaps one way of rescuing the minimalism requirement and the otherwise human condition, keeping the baby and bathwater intact. The inspiration for such a view comes from Gert (1998, 2004). Gert suggested that "[s]howing that all moral agents would endorse adopting a moral system that required everyone to act morally with regard to, at least, all other moral agents… [would provide] a justification of morality" (Gert 2004, pp. 81 -2). All moral agents, according to Gert, are rational and a further constraint on the justification of morality is that "rational persons use only those beliefs that are shared by all rational persons." (Gert, 2004, p. 82). Let us then consider the following modification of BIO$_{RD}$: value is still determined by the reactions of

benevolent observers, but those observers are otherwise human (hence, if we continue to individuate observers on the basis of their reactions, there will be large number of possible observers). This preserves Firth's 'otherwise human' requirement. However, it is only when those observers use beliefs they all share to arrive at their judgements, and perhaps even only when those judgements agree, that their reactions determine what is valuable.[49] Thus, we preserve the minimalism requirement by quantifying over the shared mental states of multiple observers, rather than across the mental states of a single minimal observer (assuming, that is, there is only one such minimal observer).[50]

Unfortunately, this theory loses all the appeal of the 'otherwise human' constraint it aspires to keep.[51] When it comes to fixing those states that determine moral value, our Gert-inspired theory quantifies over only those states shared by all observers. This leaves out all those mental states included by the 'otherwise human' constraint, leaving only those that would remain in a minimal observer. No one could sensibly be more motivated to do what a benevolent observer, otherwise like them, would desire when

[49] Another, perhaps more plausible, suggestion is that when these observers all agree we are *obligated* to perform some action, and when they disagree when are *permitted* but not obligated to perform either one of these reactions. The reader will have to wait until chapter 5 for my discussion of BIO_{RD} in relation to obligation and related terms. Suffice it to say for the moment, I do not believe the theory strongly supports the use of those terms.

[50] See section 5.1 for a detailed discussion of how I think this is best done.

[51] This is not to say that Gert's theory fails for these resaons. The theory I am considering is merely inspired by him, but I am not attributing it to him.

the only desires we took into account were those that stemmed from benevolence, than by a benevolent observer who was minimal but otherwise not like them at all. The Gert-inspired theory may be functionally equivalent to BIO_{RD} in that all the relevant reactions and conclusions we could draw from those reactions would be identical, but the Gert-inspired theory seems to muddy the waters unnecessarily by introducing the otherwise human condition for the sake of making the observer more like us, only to then quickly abandon the appeal of this condition by quantifying over a limited set of mental states. Let us simply do away with the 'otherwise human' constraint and embrace minimalism.

Chapter 4: Reactions and the objects of value

Recall that the central claim of the response-dependent theory of value is given by the following schema:

RD**:** *x* is morally valuable if and only if and because *x* elicits *R* from *S.*

In the previous chapter, I argued that '*S*' is best thought of as a relevantly informed, rational, benevolent and otherwise minimal observer. In this chapter, I discuss *R*, the relevant reactions, and *x*, the relevant objects of value. My initial aim is to discuss the concept of a non-truth-oriented attitude and defend the claim that all and only such attitudes are relevant to determining moral value. I consider reasons why one might wish to restrict the set of relevant attitudes to some proper subset of the non-truth-oriented attitudes, but find such reasons wanting. This concludes my discussion of the reactions *R*, in the schema RD. Then, I move on to consider the objects of essential value, or *x,* in that same schema. In the previous chapter, I argued that welfare is the only essential value, but in section 4.3 I flesh this out, suggesting that it is really states of affairs involving welfare that bear essential value. I then consider a variety of objections to this view, including the repugnant conclusion. I end with a discussion of the objects of non-essential value in section 4.4.

4.1. Relevant responses

The previous chapter made the case for minimalism *via* theoretical simplicity. Simply put, one ought to avoid any psychological inputs that are unnecessary for determining moral value. To see the merits of this approach, suppose that we were to drop the minimalism requirement. It would then be compatible with BIO$_{RD}$ that the benevolent observer had some strong desire that people perform an arbitrary act whenever that act was not incompatible with benevolence. Such an observer might desire that everyone briefly raise their left arm at 4.23pm on a Tuesday. Does this mean that raising our arm at this time is morally valuable? That it would bring about a good state of affairs? Or even that there might be a duty, obligation, or reason to raise our arm? If the answer to these or to any similar question is 'no', as it surely must be, then we need some principled way of excluding mental states like this one from the set of relevant attitudes. We saw in section 3.5 that there were two ways of doing this. A Gert-style quantification over multiple observers, or a minimalism requirement on each observer. Although both are functionally equivalent, I opted for the minimalism requirement.

Given that what we are interested in are attitudes borne only out of the observer's benevolence, why include other attitudes besides those borne of benevolence? If the arguments of the previous chapter establish that welfare is the only essential value, then the relevant attitudes of the observer must be those that are present *because* the observer

is benevolent (that is, has a final care directed towards the welfare of conscious creatures). Given this fact, our default assumption should be that *all* their responses have some moral relevance. Being conservative in our attribution of character traits to the observer permits us to be liberal with respect to the attitudes we deem relevant.

Those who disagree must argue that some additional refinement required so that only a proper subset of those attitudes borne of benevolence will count as relevant. Are there any such restrictions?

I will argue that the answer is a qualified 'no'. The qualification is this: all and only non-truth-oriented attitudes are relevant. At a first pass, a truth-oriented attitude, as I will use the term, is any attitude that aims to represent the world as it is. For example, belief is a truth-oriented attitude because belief 'aims' at truth (Williams 1970); this is often taken to mean that a belief is correct if and only if it is true. Other examples include certain factive mental states, such as remembering. Knowledge may be another example, depending upon whether one thinks it is a mental state. These attitudes often have an in-built standard of correctness, such that it is correct to have it if and only if one has it towards a truth. A non-truth-oriented attitude is any attitude which is not truth-oriented. These include the typical passions such as desires, hopes, fears and loves.

The argument for excluding truth-oriented attitudes from the set of relevant responses is as follows: since all truth-oriented attitudes are representational in a 'mind-to-world' direction of fit, one could only derive moral truths from an observer's truth-

oriented attitudes if at least one of those attitudes had normative content. More precisely, for an observer's truth-oriented attitude to have any normative implications, there must be an attitude with the propositional content '*x* is *N*', where '*N*' is some normative property. But defining '*N*' in terms of what the observer deems '*N*' would be insufficient as an analysis of the property *N*. Of course, one could infer from an observer's belief that *'x* is *P*' and a principle such that 'all *P*'s are *N*' that *'x* is *N*', but since the observer was supposed to determine what is valuable, appeal to an independent principle of this kind is not possible. The ground for such a principle would itself be a representational attitude and thus it would fall prey to the same circularity objection.[1]

This also serves to distinguish BIO$_{RD}$ from some of its response-dependent cousins. The typical response-dependent theory of the concept 'red' is often presented as follows:

[1] One possible way of avoiding this kind of circularity is to derive a moral principle from an entirely empirical observation of the observer's attitudes (of course, this observation cannot be empirical, *de facto*, since there are no ideal observers, even if it is empirical in principle). For instance, if one noticed that the observer always desired that one avoid breaking promises, one might thereby infer that 'promise-breaking is wrong.' The problem with this simple suggestion is that principles based on empirical generalisations of this kind are likely to admit of exceptions, given the limited number of circumstances in which one would have noted the observer's reactions. At best, one could infer 'it is *prima facie* wrong to break promises.' However, the basic point made above still stands: what grounds the truth of such principles are the observer's non-truth-oriented attitudes and we make a generalisation on the basis of those, rather than deriving the truth of the principle from the attitude 'promise-breaking is wrong.' In chapter 5, I discuss how our moral language might best be shaped to accurately reflect the psychology of the ideal observer.

RD_{RED}: an object is red if and only if and because it appears red to normal observers under normal lighting conditions.

This theory is clearly circular, but not, it is thought, viciously so. RD_{RED} is not offered as an analysis of the concept 'red', but rather as an articulation of the proper application conditions for using the concept (Smith 1998). As such, one is entitled to use a representational attitude containing the concept being explicated in the explication. BIO_{RD}, however, does aim to provide an analysis of MORAL VALUE, and thus this form of circularity, common to other response-dependent theories, is precluded.

4.1.1. An aside on truth-oriented attitudes

One worry with this liberal proposal is that it is unclear what it is for an attitude to be truth or non-truth-oriented. A second is that some relevant responses do seem truth-oriented and thus my proposal excludes too many attitudes. I will take each of these concerns in turn. This section is largely independent of the broader argument. If the reader has no interest in my response to these objections and is comfortable with the idea of normative attitudes being truth or non-truth oriented, this section may be skipped.

To repeat, the first worry is that it is unclear what it is for an attitude to be truth or non-truth-oriented. To be truth-oriented is to have some normative property. More

precisely, an attitude is truth-oriented when it has (i) a standard of correctness such that it is correct whenever the propositional object of the attitude is true, and (ii) this is the only constitutive normative standard that attitude has.[2]

The view that belief has a normative 'aim' and that this aim is either truth or knowledge is commonly called *constitutivism* or, alternatively, *normativism* (McHugh and Whiting 2014). Constitutivism about belief seems to be widely accepted, but there is strong disagreement about whether knowledge or truth is the proper standard of correctness. It is not my purpose to argue for constitutivism, since this would take us too far afield. Nor is it my desire to arbitrate the dispute between those who think that knowledge is the aim of belief versus those who think truth is the aim. Since truth is the more common candidate, I speak of truth-oriented attitudes, not knowledge-oriented attitudes. I say more about knowledge-oriented attitudes at the close of this section. Finally, the debate between constitutivists and non-constitutivists seems to focus almost exclusively on belief. There are, however, some truth-oriented attitudes aside from beliefs. However, given the extensive literature on belief, I take it as my primary example.

[2] Note that truth does not itself need to have any normative properties in order for a truth-oriented attitude to be normative. It's the fact that there is a standard associated with a property that makes the attitude in normative. In this case, the property is truth, but it could have been falsity, or having it on Tuesday afternoon, etc..

What does it mean for an attitude to be normative or have an 'aim'? According to Wedgwood:

> "…certain concepts are normative because it is a constitutive feature of these concepts that they play a regulative role in certain practices. Suppose that a certain concept '*F*' is normative for a certain practice. Then it is a constitutive feature of the concept '*F*' that if one engages in this practice, and makes judgments about which moves within the practice are *F* and which are not, one is thereby committed to regulating one's moves within the practice by those judgments." (Wedgwood 2002, p. 268)

This in turn, according to Wedgwood, makes it *irrational* to engage in the practice and knowingly violate the standards set by *F*, because violating those standards entails that one has an "incoherent" set of metal states. When applied to belief, the view entails that when one engages in the practice of forming, maintaining and revising beliefs, and one makes judgements about which beliefs are in accordance with the evidence and which are not, one cannot form a belief that is not in accordance with the evidence on pain of irrational incoherence. Evidence, of course, is an indicator of truth. Notice that according to Wedgwood's account it is both the normative standard of correctness *and* the judgement that some move is prohibited by that standard which generates the irrationality. Wedgwood could allow that someone engaging in the practice of belief who doesn't make any judgements about which moves are permissible and

impermissible by the lights of the standard of correctness and then forms a belief in violation of their evidence is not irrational or incoherent. This is the right result. Since none of the agent's judgements conflict they cannot be irrational in virtue of their incoherence. Nonetheless, it also seems that an agent of this kind is still making a mistake. They are forming beliefs whilst failing to live up to belief's own standard of correctness. But Wedgwood is unable to unambiguously account for this fact. According to his view, 'correctness' requires normativity and the kind of normativity relevant for belief is the kind that regulates a practice in the way outlined above. If one does not form the relevant judgements about the practice, there is no irrationality, no commitment to regulating one's moves within the practice, no violation of normative standards, no violation of the norm of belief and thus no sense in which one is making a mistake. If we tie correctness to rationality in the way Wedgwood does, we cannot explain why a believing agent who forms no judgements about what they ought to believe is making a mistake when they fail to believe in accordance with the evidence.

The solution is to divorce irrationality from norm violations. I suggest that one makes a mistake whenever one violates a standard of correctness that one is committed to in virtue of engaging in some practice, but one can violate that normative standard without any failure of rationality. In order to make sense of how this is possible, we need a thinner notion of normativity than Wedgwood's.

This thinner notion is what Thomson (2008) has called 'external correctness' (or *e*-correctness). *e*-correctness is relative to a kind. A shape can be *e*-correct relative to the kind 'map of the United Kingdom' but *e*-incorrect relative to the kind 'map of China.' *e*-correctness is silent about the procedures conducive to good map-making or the processes that one would enact in order to make an *e*-correct map. *e*-correctness is contrasted with internal or *i*-correctness, which is concerned with how the agent realizes *e*-correctness. *i*-correctness regulates our practices in the way Wedgwood describes, but not *e*-correctness. I propose that the standard of correctness for belief and all other truth-oriented attitudes should be thought of as *e*-correctness. Truth-oriented attitudes are attitudes that are correct iff they true, where 'correctness' is e-correctness, as opposed to *i*-correctness.

Engel (2013) has pre-emptively raised doubts about this view. He worries that the standard of correctness for belief cannot be *e*-correctness since *e*-correctness is not normative at all:

> "'Correct' in this sense [*e*-correctness] is an attributive adjective like 'good': that X is correct *qua* K does not entail that it is correct, period. But is it clear that correctness is a normative property? The standard for a tune is fixed by a set of notes, the standard for a map is fixed by the similarity between the map and the territory represented, the correct spelling is fixed by a certain pronunciation of the word. *These are descriptive properties, not normative ones*." (Engel 2013, p. 200)

The suggestion that *e*-correctness is not normative is unwarranted. Consider the two images below:

Figure 1. *A good map of the United Kingdom*

Figure 2. *A bad map of the United Kingdom*

One could exhaustively describe all the properties of both figure 1 and figure 2. Such a description would include all the facts about which areas of the image are shaded black and which are shaded white. It could be stored in a multitude of ways; as binary code on a hard-drive, or as co-ordinates on a Cartesian plane written by hand on a piece of paper. In detailing all this information, have we thereby shown that figure 1 is a good map of the United Kingdom and figure 2 is a poor map? Certainly not. Engel would agree with this much, since, in addition to all the facts about figures 1 and 2, we also need to specify an external standard. In this case, the standard is 'map of the United Kingdom'. I take Engel's worry to be that, once the standard is specified, we *do* discover

all the normative facts by exhaustively cataloguing the relevant shade and spatial information and, therefore, *e*-correctness is a descriptive rather than a normative notion. But this inference is not warranted. The first thing to notice is that if we are committed naturalists, all the colour and spatial facts about figures 1 and 2 and the external standard *must* entail all the relevant normative properties. In that sense, by describing the non-normative facts we also describe the normative ones. But the more important point is this: in specifying a standard, we are specifying a normative property. That, put bluntly, is what standards do, even if it is only normative in a thin sense. Otherwise, simply drawing the map again would be enough to show that it is good. It is the relational properties the map bears to the relevant standard that endow it with a (thin) normativity. We can see this easily in virtue of the fact that standards entitle us to use normative language. For example, figures 1 and 2 can meet the relevant standard more or less *well*, depending on how accurate the image is, and will be *good* or *bad* depending on how well they meet that standard. Standards might not allow us to speak in deontic terms, but that is no stain or their normativity more generally. Figure 1 is a *good* map of the UK and figure 2 is a very *bad* map of the UK indeed. To be sure, *e*-correctness does not regulate our practices in the same way *i*-correctness does, but that is no blot on the normativity of *e*-correctness. In fact, this is the very kind of normativity we were looking for. I hope this is enough to dispel any confusion about what truth-oriented attitudes

are: attitudes that have a standard of correctness tied to truth, where 'correctness' is understood as *e*-correctness.

The second worry presented at the start of this section was that some attitudes which seem relevant are, by this standard, truth-oriented and are thus incorrectly deemed irrelevant. Consider fear. What the benevolent ideal observer fears (if they fear anything at all) is, I think, relevant. Isn't it only appropriate to fear something if what one fears is 'true' in that it something that will or has occurred? It is not clear, at least to me, that this is so. I might fear that I'm overbearing, but not be overbearing. Is my fear 'incorrect'? I'm inclined to think not. However, even if others disagree, we must bear in mind that truth-oriented attitudes are those whose *only* standard of correctness is tied to truth. If fear is tied to truth, then surely it cannot be only tied to truth? The object of fear must itself be worthy of fearing. Fearing harmless puppies cannot be appropriate because puppies are not worthy of fear. So, if fear is a normative attitude, it is not normative with respect to truth alone. Fear, thus, may be an example of a 'mixed' attitude, where truth is part of the standard of correctness, but not exhaustive of it. These attitudes, I claim, are not truth-oriented and so relevant. To provide a counter-example to the present theory, one must give an example of an attitude that ought to be relevant and which is normative only with respect to truth, or a mixed attitude or completely non-truth-oriented attitude which seems irrelevant. Such an examination

must be conducted on a case-by-case basis, and I will omit performing this labour-intensive task since no plausible examples come to mind.[3]

I'll end this section by returning to the difficulties raised if we adopt the knowledge norm instead of the truth-norm. Supposing that knowledge is a mental state, then it is not truth-oriented, because although it may have truth as a normative standard, truth will not be its only normative standard, but presumably also justification of some sort.[4] But knowledge is exactly the type of attitude that I wish to exclude from the set of relevant responses. The only way to solve this problem is, I think, to acknowledge that if knowledge is a mental state, then knowledge is likely fundamental to epistemology in the way that Williamson (2000) has described. In that case, the truth-oriented attitudes become knowledge-oriented attitudes. This coheres with what Williamson seems to think about attitudes like belief, which he describes as a kind of 'botched knowing'. Remembering would be a way of knowing and thus a knowledge-oriented attitude. More precisely, a knowledge-oriented attitude is one that is correct (in the sense of '*e*-correct') iff it is knowledge. The set of relevant responses of the benevolent observer would then be the non-knowledge-oriented mental states. If knowledge is not a mental

[3] One interesting class of examples conerns attitudes that seem partly truth-oriented but have additional non-epistemic components. This might include curosity or surprise. However, it seems that both these attitudes could only be had by an agent who was previously unaware of some fact and, of course, the ideal observer already knows all the relevant facts (section 3.1.1).

[4] For the view that knowledge is simply true belief, see Sartwell (1991, 1992)

state, then this modification is not required, since it does count as among the set of mental responses out of which the set of relevant responses is carved.

4.2. Excluding additional non-truth-oriented attitudes

To repeat, once truth-oriented attitudes are excluded, the argument for permitting all non-truth-oriented attitudes into the set of relevant responses is relatively simple. Since we have agreed, at least for the sake of argument, that an observer whose *only* final care concerns welfare determines moral value, the exclusion of any response from the set of relevant responses will leave out some attitude that arises due to concern for our welfare (since all non-truth-oriented attitudes arise out of this concern). But since concern for welfare is the very attitude that plays a critical role in determining moral value, we can have no principled reason to exclude it. Our default assumption should be that all non-truth-oriented attitudes are relevant.

However, one may wish to place additional restrictions on the set of relevant attitudes. There are at least two plausible reasons of this kind.

The first is the need define common moral terms such as rightness, duty, obligation, etc. If we accept the foregoing, then these terms must be defined in relation to the observer's attitudes. Some of these attitudes will be better suited to the task than others. For example, let us say that a morally right action is one that I ought to perform. If we accept some version of the 'ought implies can' principle, then right action will be

constrained by what is possible for the agent in question to do.[5] Thus, any attitude that is not similarly constrained cannot be used to define right action. For example, the observer may wish that I perform an action such that everyone in the world is benefitted by it, but if it is not possible for me to perform such an action, then right action cannot be what the observer would *wish* that I do. If a term like 'right action' is constrained in this way, perhaps the same goes for other common moral terms? By compiling an exhaustive list of all these terms and finding suitable counterparts in the attitudes of the benevolent observer, one would, the thought goes, create a list of the observer's relevant attitudes, which may or may not include all non-truth-oriented attitudes. This would provide a principled way of excluding certain mental states even within the domain of non-truth-oriented attitudes.

This approach is mistaken on two fronts. The first is that even if defining all common moral terms could be done without using all the observer's non-truth-oriented attitudes, this would not rule out the moral significance of mental states for which there was no suitable widely used moral term. Perhaps, for example, there is no common term that neatly captures what the observer would be anxious about. But the fact that our moral discourse happens to lack a term which corresponds to a certain type of non-truth-oriented attitude tells us more about the limitations of our language than it does moral

[5] For doubts about this principle, see Sinnott-Armstrong (1984) and Henne et al. (2016).

value. Given that we communicate exclusively with common moral terms, it may be difficult to appreciate the artificiality those terms can impose on our moral thinking. Our language may lack neat correlates for what an ideal observer fears, loves, wishes or is anxious about (especially if such attitudes are non-propositional in character), yet our language *could* have had such terms, and perhaps other languages do. If we spoke such a language, then nothing would have changed about moral value. All the same attitudes would still be relevant; only the way we think about those attitudes would have changed.

The second, more pressing concern, is that this approach just gets matters the wrong way around. Rather than start with particular moral terms and rule out attitudes as irrelevant if they fail to fit those terms, we should take the attitudes of the observer as primary, adapting our moral terminology to those attitudes. This further claim, that we should, as far as possible, alter our language so as to better suit the attitudes of the observer, is, in part, why BIO$_{RD}$ is a revisionary ethical theory. But this revisionary claim comes cheap once we have accepted that the benevolent observer is the determiner of moral value. As stated previously, common moral discourse provides artificial constraints on what we tend to think is morally valuable. Any non-truth-oriented attitude is relevant, because that attitude is not merely aimed at representing the world as it is and it is borne of a final care directed towards welfare. Therefore, our moral discourse ought, as far as possible, to reflect the wide-range of attitudes a benevolent

observer may have towards our conduct and character. Chapter 5 contains an extensive discussion of how to revise our moral discourse in light of the truth of BIO_{RD}.

The second way in which one might attempt to limit the set of relevant attitudes is by insisting that only propositional attitudes ought to count since only these attitudes will be useful when attempting to define any moral term. For instance, suppose we wanted to define 'right action' in terms of the observer's attitudes as follows: it is right for p to φ when the observer most desires that p φ's. It is only because the observer desires *that p* φ*'s* that it is right *that p* φ*'s* as opposed to it being right *that q* φ*'s* or *that p* ψ*'s*. Attitudes that lack propositional content cannot ground right action because they are not *about* a particular person or action.

This suggestion can be dealt with in the same way as the previous one. Rather than starting with common moral terms and trying to find correlates in the attitudes of the observer, we ought to take the psychology of the ideal observer as primary. We ought to do so precisely because the observer's psychology is what determines moral value, not our decision to prioritise certain parts of our language of others. Yet, non-propositional attitudes raise some interesting questions. There are two ways in which an attitude may be non-propositional: it may lack any object, or it may have as its object something other than a proposition. One might worry that the first kind of non-propositional attitude, those that lack any object, will never be able to tell us anything of moral interest, since they are not themselves about anything. This worry, however, is misplaced, since even

though one cannot infer anything from the content of a non-propositional attitude one can still note the state of affairs that caused the attitude in the first place. Suppose that a friend you know to be morally outstanding becomes overwhelmingly sad when they reflect on a horrifying event. Although they may also have many propositional attitudes directed towards the event, their sadness itself may have no object and so is a poor candidate for grounding duties or obligations. Still, it is clear that even though their sadness lacks an object, it was nonetheless due to said horrifying event and is, in that weaker, non-intentional, causal sense, *about* it. It is, therefore, relevant to our assessment of that event.

The second variety of non-propositional attitude, those that have objects other than propositions, raise more interesting questions. In particular, consider the possibility that care itself is one such attitude (recall section 3.3.1). That it is to say, when one cares what one is typically doing is caring about some individual rather than a state of affairs. When we care about someone's welfare, we do so only because we care about *them*. We do not literally and fetishistically care about their welfare for its own sake. The thought that this kind of non-propositional care is really what's important for morality suggests that BIO$_{RD}$ makes a fundamental error in providing the observer with a final care directed at welfare, as opposed to a final care directed towards the individuals whose welfare matters. Thus, the proper objects of value are individuals, rather than states of affairs, as BIO$_{RD}$ suggests. I devote the next section to this thought.

4.3. Individuals, states of affairs and care

In *Economics and Value*, Elizabeth Anderson explicitly defends the view that the value of states of affairs depends upon the value of individuals:

> "… states of affairs are generally only extrinsically valuable, because our intrinsic evaluative attitudes do not generally take them as their immediate objects. It makes sense for a person to value most states of affairs only because it makes sense for him to value people, animals, and other things. […] Reflection on a few examples should convince one of its truth. All states of affairs that consist in someone's welfare are only extrinsically valuable. If it doesn't make sense to value the person (in a particular way), then it doesn't make sense to care about promoting her welfare. […] Enemies, who hate each other, have no reason to promote each other's welfare. Mary may rationally feel self-contempt for betraying her profession as a journalist. (Perhaps she published a story she knew to be false, as a favor to a government official.) Under this condition of self-disvaluation, it doesn't make sense for her to seek her own advancement in it until she has made amends, for she regards her advancement as undeserved and, hence, unworthy of pursuit." (Anderson 1993, p. 26)

David Velleman, endorsing Darwall's (2002) view of welfare according to which someone's welfare is what it would be rational to want for their sake, makes a similar point.

> "In Darwall's analysis [...] things that were good for you would not actually merit concern unless you merited concern; and if you didn't, then despite their being good for you, they wouldn't ultimately be worth wanting, after all [...] what's good for a person is not a categorical value, any more than what's good for a purpose. What's good for a purpose is worth caring about only out of concern for the purpose, and hence, only insofar as the purpose is worth caring about. Similarly, what's good for a person is only worth caring about in so far as he is worth caring about. A person's good only has hypothetical or conditional value, which depends on the value of the person himself." (Velleman 1999, p. 611)

Why is this a problem for BIO$_{RD}$? If Anderson and Velleman are correct then there is some sense in which it is a mistake to value someone's welfare without valuing the individual whose welfare it is. Although the benevolent observer may value an individual, the reason why they value their welfare will not be because they value that person, since the benevolent observer values welfare *for its own sake* (this is what I've called a final care). And to value welfare for its own sake is to regard positively states of affairs in which there is positive welfare and regard negatively states of affairs in which there is negative welfare for their own sakes.

If we take these criticisms to heart, we arrive at a modified version of the benevolent ideal observer theory. The person who has done the most to articulate and defend this alternative approach is Christian Coons (2006, 2012). Coons defends what he calls the dependence thesis, or DT. DT is a general claim about the goodness of states of affairs,

and states that their goodness depends upon the value of some individual. He then defends a specific version of DT, DT*, according to which the goodness of states of affairs depends upon the existence of individuals who merit *concern* or *respect* (Coons 2006, p. 45). Later, Coons (2012) develops his view into what he calls the 'Ideal Carer theory'. The ideal carer is a fully informed, rational being who cares (and perhaps respects) individuals for their own sakes.[6] What's morally relevant is what such an observer would want or will.

In the course of their arguments, Anderson, Velleman and Coons provide four objections to a view such as BIO$_{RD}$. The first is that it is simply incoherent to value states for their own sakes without valuing individuals. The second is that valuing states for their own sake would be pointless, and thus violate the purported authority of morality. The third objection is that a view such as BIO$_{RD}$ gets the wrong result in certain important cases in population ethics. The fourth objection is that the benevolent observer is akin to a kind of civil servant, who may issue in correct judgements, but

[6] There are some additional important similarities and differences between Coons' ideal carer and the benevolent observer. Both Coons and I agree that such a being should have the least complex psychology required to realise the relevant states and thus we both reject Firth's 'otherwise human' constraint (thus, "the [ideal carer] has no desire to sing while listening to the radio – normal though it may be." (Coons 2012, p. 225). However, Coons rejects response-dependence in favour of a fitting-attitudes account of value (see section 2.4).

somehow misses the point of morality. I take each of these objections in turn.[7]

We can see the first objection clearly in the Anderson quotation above. The claim is that it does not "make sense" to value states of affairs without valuing some individual. There are two ways in which one can understand this. The first is that it doesn't make sense to value states of affairs without valuing some individual because it is simply conceptually incoherent to do so. This, however, is simply too strong. Certainly, I can coherently imagine valuing some states of affairs for the own sake, otherwise I wouldn't have bothered to write this dissertation. Indeed, Coons seems to agree.

> "At first glance the position that states always or sometimes have independent value seems compelling. After all, it is often said that things like pleasure, happiness, freedom, and knowledge are good in themselves. And of course, when we say such things, we don't mean to say that the properties denoted by these concepts are themselves good. Instead, we seem to mean that states of someone being pleased, happy, free, or having knowledge are good in a way that does not depend on the value of anything else." (Coons 2006, p. 2)

Anderson's thought must be the weaker one, that it does not make sense *morally* to value states of affairs without valuing some individual. But this is exactly what I am alleging is false and so begs the question.

---

[7] The first objection is made, in some form, by all three but the final three are Coons' and, indeed, he also uses some of them to correct flaws in Anderson and Velleman's own accounts.

The remaining three, more interesting objections, are all due to Coons. Although I aim to rebut them here, I think that all of these considerations are insightful. They map clearly the deep fault lines between much consequentialist and non-consequentialist thought. In reading these objections, the reader may come to simply disagree with my view. I suggest that this is a moment at which fundamentally incompatible and competing visions of morality come into clear focus, and aside from merely articulating the consequences of adopting each position and seeing where one's sympathy lies, it is difficult to know how to resolve such disputes.

Coons' first objection is that views like BIO$_{RD}$ are, in a fundamental sense, arbitrary and as such cannot provide morality with the authority we typically associate with it. Coons suggests that the good is normatively authoritative in that "[g]ood states direct and bind us, not physically, but by setting normative goals and limits for our actions." He continues:

> "On reflection, a good state that is not worth realizing for the sake of any individual(s) is incompatible the normative authority of the good. Realizing a state that's not worth realizing for anyone or anything's sake appears to be pointless. Therefore, holding that such states can be good may commit one to holding that agents can be appropriately required to do something pointless. But no demand to realize a pointless outcome would be justified. Therefore, it seems that states can't be good independently of being worth realizing for the sake of individuals. […]

> Those who resist this line of argument owe us an explanation of why we really could be properly required to realize states that are not worth realizing for any individual's sake, an explanation of why realizing them is not pointless. No explanation is forthcoming. These states are alleged to be good independent of the value of anything else. So there can be no further *evaluative* explanation about why we could be required to realize them. And it does not appear that any change in our non-evaluative beliefs could get us to see that it *does* make sense to realize outcomes that are not worth realizing for anyone or anything. So any explanation using nonevaluative claims will fail too." (Coons 2012, p. 61-2)

There are multiple ways to respond to this argument. The first is to deny that the good is authoritative in the sense Coons describes. Recall that, in chapter 2, I stated that one of the constraints on moral theory is that it be action-guiding. We should be able to turn to morality to help us resolve conflicts between different choices, at least in principle. However, it is not part of this picture that morality ever demands or requires that we perform an action. This somewhat radical position is defended at length in chapter 5, and so I won't discuss it further here. I mention it only to note that it is one possible line of response. Nonetheless, I think Coons' challenge can be met independently of this consideration.

The claim that it is pointless to realise a good state (or for the observer to want to realise such a state) is question-begging. To provide a point, one would need to cite

other normative facts, but the central tenet of BIO$_{RD}$ is that the final care towards welfare by the ideal observer is the fundamental good-making feature of the world. Coons is aware of this problem. His response is to note that we are in fact inclined to suspect that there must be some further point when told that a state is fundamentally good and that "it's strange to suppose we must simply make the world a certain way, without any possible further evaluative explanation for why we should make it that way" (Coons 2006, p. 61, fn. 27). However, when we are told that individuals capable of being welfare subjects merit concern (and/or respect) we are not inclined to demand a further explanation. The question, "what's the point?" is simply less compelling. Individuals just *deserve* care or respect. The same cannot be said for states of affairs.

It's important to note Coons' argumentative strategy. It is a clear case of an appeal to how an ordinary and competent user of a common but interesting philosophical concept would respond. I'm unsympathetic to this kind of conceptual analysis for the reasons discussed in chapter 2. It's not good enough to assert most of us would respond that way without actually providing some empirical evidence that this is the case. My own suspicion (and it's nothing more than a suspicion) is that the issue is so philosophically complex that most ordinary speakers wouldn't have a clear reaction one way or the other. Yet, Coons can (and I think should) simply adopt the conceptual ethicists' approach and insist that his concept better meets the relevant criteria and does so in a way that requires fewer revisions from his starting concept, whether it be 'ordinary' (if

there is such a concept) or not. In this case, we have reached theoretical bedrock. I simply don't feel the pull of the question 'what's the point?' when I think of the observer's desire to maximise welfare states.

I hope, however, to do slightly better than this and provide a story about why this question doesn't move me in the way it does Coons.[8] The observer's final care about welfare is not arbitrary in the same way it would be if the observer cared about realising states that maximized the number of barns in the Ohio countryside, or the number of practitioners of the Argentine tango. As biological creatures, we are not so disposed as to orient our lives and conduct around barns or the tango. Instead, we seem to be creatures whose interaction with the world is fundamentally governed by our welfare. This is what Bentham was getting at when he famously claimed that "Nature has placed mankind under the governance of two sovereign masters, pain and pleasure." (*PML*, Ch. 1) Typically, we cannot help but care about pain and pleasure, and, I would suggest, our welfare. And it's not just our own that cannot help caring about either. This was the starting point of the sentimentalist tradition discussed in section 2.3 and it's still, I

[8] The story is, however, subject to certain empirical psychological claims which we are at present unable to demonstrate with any certainty. However, these claims, such as 'we cannot help but care about our welfare' are intended to be generic and inoffensive enough to be fairly plausible without perfect evidence. Of course, there will be exceptions (some depressed people do not care about our welfare), but it would be odd to build a sentimentalist moral theory out of sentiemnts we have in atypical and often undesirable psychological states.

suggest, the most plausible way we can ground morality. Both Coons and I are welfarists and thus agree on the centrality of welfare. Our difference lies in how we take this to impact the moral landscape. Coons insists that valuing individuals grounds the goodness of states, whereas I insist that our care for individuals stems from our caring about welfare. If there is any arbitrariness is my position, it is in the decision to ground the value of persons in the value of welfare, and not vice-versa. However, the decision to focus on welfare in the first place is not arbitrary at all (or at least exactly as arbitrary Coons' decision to do the same).

Coons' second objection focuses on thought-experiments in population ethics. Consider the diagram below made famous by Parfit (1984):

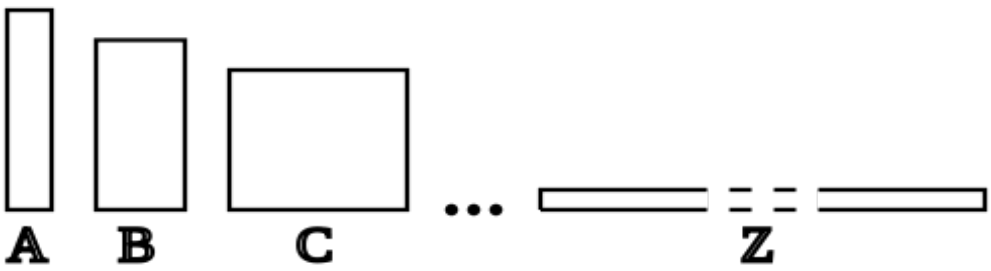

Figure 3. *Parfit's 'Repugnant' Conclusion*

Each bar represents welfare at a world. The area of the bar signifies the total amount of welfare, whilst the height represents the amount of welfare each individual possesses

at that world. The higher the bar, the higher the welfare. As we move from world *A* to world *Z*, the total amount of welfare increases whilst the amount of welfare possessed by any individual decreases. A theory according to which more overall welfare is always better must claim that $A < B < C < \ldots < Z$.[9] Parfit famously called this 'the Repugnant Conclusion.' It does not seem that a world where many beings have minimal amounts of welfare (but still lives that are, on the whole, worth living) must be better than a world where fewer people have far greater welfare. As such, BIO_{RD} (and the interpretation I offer) must embrace the Repugnant Conclusion, despite the intuitions of some, perhaps most, others.

Coons claims that his ideal carer theory can (rather ingeniously) avoid the repugnant conclusion. This is because the ideal carer only cares about welfare for the sake of actual individuals. Thus, there is no reason to move from the *A* world to the *B* world, despite the increase in overall welfare. The individuals at the *B* world are either (i) all different individuals to those at the *A* world, or (ii) the same individuals at the *A* world plus some additional individuals. If (i) is true, then there could be no reason to move from the *A* world to the *B* world since the *B* world contains individuals who do not yet exist, and the ideal carer's care is directed towards the individuals at the *A* world who would cease

[9] One popular way of avoiding the repugnant conclusion is to deny the transitivity of 'better-than'. See, for instance Tempkin (2012) and Rachels (2004). This literature is vast, not directly relevant to my main thesis, and, I think, ultimately unconvincing, so I will not discuss it, aside from this brief note.

to exist. You cannot want someone to prematurely cease to exist out of care. If (ii) is true, then the individuals from the *A* world for whose sake the ideal carer cares would have their welfare lowered, so there could be no reason to move from the *A* world to the *B* world (and the additional individuals at the *B* world couldn't tip the scales for the same reason that (i) would also fail to provide a reason to move from *A* to *B*). Even if we simply add more lives to the *A* world (stipulating that they are lives worth living) without lowering the welfare of any current *A* world individual, these additional individuals won't make any difference to Coons' ideal carer, since they only care about welfare for the sake of individuals who currently exist in the *A* world.

This, many agree, is the correct result; the point of ethics is to make people happy, not to make happy people (Narveson 1973). I, on the other hand, believe that this is incorrect. Parfit's Repugnant Conclusion is not, I think, repugnant at all. The argument that leads to the Repugnant Conclusion is simple and compelling. It's a better world in which the average welfare is lowered only very slightly due to the addition of some people who are almost as happy as the original individuals at that world. And given the transitivity of 'better-than', the Repugnant Conclusion follows.

More pressingly, getting the 'correct answer' for the Repugnant Conclusion leads to absurd results elsewhere. Consider the following case: at world *w*, there are one million happy individuals, all leading lives filled with joy and meaningful experiences. We could add to *w* one hundred million additional individuals living equally happy and

meaningful lives. Let us stipulate that this additional population live their lives in total epistemic and, for all practical purposes, causal isolation from the original one million *w*-world individuals. If we flip a switch, we can begin a causal process that would create these additional, welfare-filled lives. If we do nothing, then the original one million individuals will continue to lead excellent lives, but no additional individuals will come into being. Coons' ideal carer seems committed to total indifference between these two options. They only care about welfare for the sake of presently existing individuals. Since the extra one hundred million individuals do not presently exist, and, given their isolation, their existence could not impact the lives of any of the original one million individuals, it could not be *for the sake* of any of the original individuals that the ideal carer desires that there be additional happy people. The ideal carer theory gets the wrong result in this case. More controversially, I suggest that this is true even when the numbers are smaller. Even if we could add one further individual to *w* in the way described above, there is at least some reason, even if it is very weak, to prefer this world to that with only the original individuals in it. The ideal carer cannot capture this result.

There are also reasons to doubt the typical responses people have towards the repugnant conclusion. Huemer (2008) identifies four such reasons to doubt the typical responses of those who wish to reject the repugnant conclusion. The first is our egoistic bias. When contemplating Parfit's original case, we may imagine which world we would

prefer to live in, the *A* world or the *Z* world. Of course, in the case of any particular individual, that individual would prefer to live in the *A* world, since they would, at that world, lead a better life. But our egotistical preferences take no account of the number of other individuals living at the world and are thus irrelevant in assessing the repugnant conclusion. The second is an inability to properly conceive of large numbers.[10] Beyond a certain point, it becomes excessively difficult to imagine the differences between large numbers, even if those differences are extraordinarily large. It is difficult to imagine the difference between, say, five billion years and five trillion years, despite the latter being three orders of magnitude greater than the former. In the case of the repugnant conclusion, the number of individuals required at the *Z* world to exhibit the same aggregate level of welfare as the *A* world may be enormous. Our responses to this case are thus (at least in part) shaped by our inability to conceive of the significance of extremely large numbers. A third reason is our inability to compound small numbers accurately.[11] Huemer gives a common but dramatic illustration of this failure: the famous example of our inability to accurately intuit the thickness of paper when folded many times. A piece of paper, one thousandth of an inch thick, when folded in half fifty times will be approximately 18 million miles thick, far greater than what most people

[10] See also Broome (2004, pp. 57-9).

[11] Huemer cites one study in which people are more willing to use car seatbelts when the lifetime risk of failing to use one is presented with than, as compared with the risk per individual trip ((Fischhoff et al. 1978). cited in Huemer (2008, p. 468)).

would estimate. Similarly, many people have the intuition that one death would be far worse than any possible number of headaches. Yet, as Norcross (1997) points out, most would likely object to lowering the speed limit dramatically to prevent loss of life, even though this would at worse cause widespread small inconvenience. Returning to the repugnant conclusion, the suggestion is that we are simply bad at compounding the small amount of welfare experienced by individuals at the *Z* world, so much so that we underestimate the value of that world. The final reason provided by Huemer is our tendency to underrate the value of low-quality lives. Huemer asks us to consider lives that are "unrealistically simple" in that they contain no experiences besides that of a uniform, mild pleasure. This is the best analogue of lives at the *Z* world, yet even this may be very hard to imagine, since we tend to imagine how we ourselves would find such a life (thus, there are echoes of the egotistical bias).[12] You or I would likely experience such a life as boring or meaningless, but as soon as those thoughts occur, the life begins to lose its value in our assessment. As Huemer says, all this "may combine to give us a negative reaction to what we intended to be a slightly *positive* state" (2008, p. 910). For these reasons, those who buy-in to the repugnance of the Repugnant

[12] This is a similar point to Nagel's famous argument that we cannot adequately imagine what it is like to be a bat because, when we try, we are imagining what it would be like *for u*s to be a bat and not what it's like for the bat (Nagel 1974).

Conclusion ought to be sceptical of their initial reactions.[13]

Coons' final objection comes by way of a thought experiment:

> "Consider a civil servant who works at the department of Human Welfare, who is so invested in his work that he comes to deeply care about, and acts to promote, human welfare – non-instrumentally. Clearly we can stipulate, without contradiction, that the civil servant neither cares for anybody, nor views such concern as appropriate. Instead, the civil servant literally values human welfare for its own sake, and not for the sake of the humans to whom it accrues. So, valuing an individual and valuing her welfare are distinct. And so the value of an individual and the value of her welfare are not the same thing. […] It may be laudable that the civil servant works to improve human welfare, and it may be more efficient to value human welfare directly, or "for its own sake," yet he seems to miss the point. His position seems fetishistic. He cares for no one, and he thinks no one deserves care. Yet he thinks that the promotion of human welfare is good. I submit that he misses the reason why human welfare is non-instrumentally good: Our welfare non-instrumentally merits promotion because humans merit concern." (Coons 2006, p. 27)

This thought-experiment is simply a way of putting Coons' central point directly. The hypothetical civil servant is like the benevolent ideal observer in that they care about

[13] For additional reasons to accept the Repugnant Conclusion, see Tännsjö (2002), Mackie (1985), Hare (1988) and Ryberg (1996).

welfare for its own sake, rather than caring about welfare for the sake of welfare subjects. And don't we have a strong sense that such an observer simply misses the point? If there is something suspect about Coons' civil servant, won't there be something suspect about the benevolent ideal observer too?

There are a variety of ways to understand Coons' civil servant. The simplest interpretation is that they desire welfare for its own sake (and, we can suppose, nothing else for its own sake). As such, all their desires are proportioned to the aggregate amount of welfare in each possible outcome. What seems objectionable about such a being is that their desire is, in a sense, arbitrary. They are fetishistically desiring that some state of affairs be realised, but they have no understanding or engagement with the reality of being a creature with welfare states. Indeed, it is easy to imagine taking what Dennett (1987) has called 'the intentional stance' towards a machine with 'desires' whose outputs matched that of such an observer. One images the mechanical civil servant coldly making decisions on the basis of a spreadsheet, without any regard for the significance of the numbers he manipulates.

Still, this civil servant, for all their faults, is perhaps not quite as bad as Coons' describes. He writes that "clearly we can stipulate, without contradiction, that the civil servant neither cares for anybody, nor views such concern as appropriate" (2006, p. 27) Whether or not the civil servant cares for anybody depends upon the definition of care and upon the facts. The civil servant may well desire that individuals fare well, albeit

only instrumentally. But whether such a desire amounts to a care depends upon our definition of 'care.'

In section 3.3.1, I discussed the nature of care. I suggested that the benevolent observer's final care towards welfare might be more than a mere desire for welfare for its own sake, though that may be essential part of care. Care may be, in addition, an imaginative capacity. According to this view, in order to care for another individual, one must have had certain conscious experiences oneself and imagined what it is like for others to have had similar experiences. What is not required is a Firthian omnipercipience, in which the observer must be able to imagine what it is like for every agent to be the subject of every possible experience. This is because imagination is not required *in order for the observer to know all the relevant facts*. The observer can know these facts under any description, including an objective, physical one.

The point is relevant here because an observer with a final care directed towards welfare is distinct from a civil servant with a mere desire that welfare be maximised. Caring about welfare may (psychologically) entail possessing a wide variety of additional non-truth-oriented attitudes. A benevolent observer who cares may also feel regret, love, shame, disgust, admiration, etc. All these mental states will be relevant to our moral assessment of an event and it seems that all of these attitudes can be had by someone capable of caring about something such as welfare. It is doubtful that such states could be had by the civil servant Coons describes. This is an important difference

between the benevolent observer and the civil servant. One that, I suggest, ought to remove any sense that the benevolent observer 'misses the point.' In fact, I would suggest that the intuition Coons is expressing is part of the reason to accept a richer conception of benevolence as care, rather than as a simple desire.

There is at least one additional reason to prefer BIO$_{RD}$ to Coons' Ideal Carer theory. This reason requires one to accept a further controversial philosophical theory. I only mention it here because I find this theory persuasive. I will not make a serious attempt to convince the reader that this theory is true, though I will explain why I and others find it persuasive and how it relates to the choice between BIO$_{RD}$ and the Ideal Carer theory.

As is clear from the final chapter of his *Methods of Ethics*, Sidgwick believed that the most important problem in ethics was reconciling egotism with morality. He called it "the profoundest problem in ethics" (Sidgwick, *Methods*, Bk. 3, Ch. xiii, fn. 4). There appear to be two sets of reasons, those stemming from our self-interested point of view, and those stemming from a generalised perspective, or 'the point of view of the universe', as it were. Sidgwick thought it impossible to show why the moral reasons trumped the egotistical ones, and this troubled him greatly. Parfit's masterpiece, *Reasons and Persons* (1984), is an attempt to meet Sidgwick's challenge. His solution is ingenious. By considering a number of thought experiments, Parfit argues that what we care about, from a self-interested point of view, ought not to be that there is some person in the

future who is uniquely psychologically connected to us, but the existence of those psychological connections. Let us say that two global mental states are psychologically connected if and only if they share enough of the same individual mental states. Psychological continuity is the psychological connectedness relation closed under transitivity. We can then say that two individuals are numerically identical iff they are uniquely psychologically continuous with one another. If we have egotistical reasons to promote our own welfare, then those are reasons to promote the welfare of whoever is uniquely psychologically continuous with us.

Parfit famously considers 'branching cases' involving fusion and fission. In one example, I am transported to Mars *via* a machine which scans my body on Earth, destroys it, and recreates a prefect replica on Mars. If, like me, you believe that this is as good as survival and thus a way of transporting me to Mars, then it seems you must also think this is as good as survival when the transporter malfunctions, and you survive on Earth for an additional 15 minutes before your body is belatedly destroyed. Although it seems incredible to think that the fact that there is now a person on Mars with an identical psychology to my own (all but for the final 15 minutes) is as good as survival, it seems that this must be so. In another case, suppose that my brain is severed in two, as is the case in some medical procedures designed to cure certain forms of epilepsy. Now imagine what is only at present science-fiction: that these two hemispheres are placed in two separate bodies. Which one am I? I cannot be both, since personal identity over time

requires uniqueness. I cannot be one but not the other, for what reason could there be to identify with one of these hemispheres but not the other? But I also cannot be neither, since everything I seem now to care about self-interestedly survives into the future, albeit separated into two distinct bodies. Parfit concludes that there is simply no further fact about which person I am. Facts about individuals are simply not that deep, given that what we care about is the relationship of psychological connectedness, and this relationship can be spread across individuals who are not uniquely psychologically connected with one another.

Here is another modification on Parfit's own cases. The day I was born, I presumably had a complex and confusing bundle of basic mental states. By the age of 5, I had acquired a far more complex set of mental states, and had begun, for example, to develop a theory of mind. By now, at age 25, I have very few of the beliefs, desires or intentions I had 20 years earlier. If I survive another 50 years and live to age of 75, my mental states may be similar to those I had at 25, or they may be extremely weakly psychologically connected with my 25-year-old mental states. Imagine that we have cured aging and that people no longer die through any natural cause, continuing to live healthy lives conducive to welfare. I live to be 1000 years old. It is *highly* implausible that my 1000-year-old self has all but the most basic mental states in common with my new born, 5-year, 25-year or even 75-year-old self. Now let us suppose that in the year 2992 AD, when I am 1000 years old, there exists some 25-year-old individual who happens to

share more psychological connections with my 25-year-old self than my 1000-year-old self does. He has very similar goals, desires, likes and aversions, his passions are similar and so is his taste. Does my 25-year-old self in 2018 have more self-interested reason to promote the interests his 1000-year-old self in 2992, who is as different from him as a stranger is now, or does he have more self-interested reason to promote the interests of the 25-year-old at 2992? If, as seems clear, I now have more self-interested reason to promote the interests of this future 25-year-old than my future 1000-year-old self, then apparently self-interested reasons cannot really be uniquely tied to me at all.

If these (and other) cases are convincing, and Parfit's reasoning is sound, there is reason to believe that individuals are less important than one might initially think. Parfit suggested that his views were similar to that of Buddha, who, he claimed, believed that the self was an illusion and that one could be rid of it through extensive meditative practice. This insight allows us to ask some rather radical questions. For instance, what if there was a positive experience that no individual had? What if there were a world filled with such experiences? I believe that there would be reason to create such world or to improve the experiences in such a world in whatever way was possible. Coons' view, taking the individual as fundamental, is unable to explain this, admittedly strange, result. BIO$_{RD}$ can, because states of affairs are the objects of essential value, not

individuals.[14]

4.4. Non-essential moral value

Welfare is the only essential value, and thus the observer's only final care is directed towards welfare. Yet, the benevolent observer will have many non-final cares directed towards other objects and these, I suggest, will be non-essentially valuable.

The general approach is, once again, liberal. Any object to which the observer responds with a non-truth-oriented attitude or any object which is the propositional

---

[14] This dissertation ultimately defends a form of utilitarianism. Here is what Coons says about utilitarianism:

> "One suspects […] that utilitarians adopt their conception of the good because they care for, or think it is appropriate to care for, sentient creatures. Just look at the classical arguments that utilitarians provide for the value of happiness, pleasure, and personal welfare. Typically, they argue that these state types are good because we, or idealized and benevolent individuals, non-instrumentally desire, prefer, value or approve of such states. But how can the fact that we, or a benevolent and idealized agent, care about particular states indicate that the state is good unless it further supposed that we merited concern, benevolence or respect? So it seems that either these arguments don't work or they rest on tacit assumption that some individuals have value. One certainly feels inclined to attribute such an assumption to utilitarians. We would be at least perplexed if they denied that it is non-instrumentally appropriate to care for people or sentient creatures or if they revealed that caring for humans or sentient creatures had nothing to do with the grounds for maximizing states of welfare. Of course, utilitarians can claim that care is an appropriate attitude to bear towards persons if bearing that attitude towards people maximizes good states. But utilitarians can say that about any attitude, even *hate* for persons. We believe that persons are worthy objects of concern even before we examine the effects of having concern for persons. I presume this partly why utilitarians adopt well-being, happiness, or pleasure as constituents of their conception of the good."

I think Coons' is quite right to point out what attracts some people to utilitarianism and identifies what they are likely, on reflection, to believe is an error. Coons' reveals that utilitarianism is in fact a far more radical theory than it is usually taken to be.

object of such an attitude is relevant. Once again, we can guarantee this is so because of the minimalism condition defended in section 3.5. The fact that the observer has no final cares for anything of non-essential value, is rational and otherwise minimal guarantees that all their other cares will result from their single final care. This liberality entails some interesting results with respect to what it takes to live in accordance with morality, and does, I hope, go some way to assuaging the doubts of soft pluralists, who are alarmed at the welfarist's monism.

One way to understand the potential variety of the objects of non-essential value is to imagine asking the observer different questions. The triumvirate of popular normative ethical theories consists of deontology, virtue ethics, and consequentialism. Speaking broadly, without paying great attention to subtle distinctions one could make within each of these respective theories, each insists that a different kind of object is the primary bearer of value. The deontologist insists on actions or action-types, the virtue ethicist on character traits and the consequentialist on states of affairs. Thus, we might imaginatively consult the observer on any one, or combination of, the following three questions:

(Q_A) Which action-types do you most desire that I perform regularly?

(Q_C) Which character-traits do you most desire that I possess?

(Q_S) Which states-of-affairs do you most desire I bring about on this occasion?[15]

It is easy to imagine cases in which the answers to these questions recommend different courses of action. Consider a scenario in which I have promised to tell a friend if they become overweight. However, telling them they are overweight would cause them extreme sadness and lead to a depression in which their weight only increased. It turns out that by lying and telling my friend that they are not overweight, they will become happy at having achieved their goal and feel less need to eat, eventually losing more weight than they otherwise would have. Although lying would lead to the best consequences, it would also be a direct violation of a promise. As such, the kindest thing to do, we can stipulate, would be to say nothing or avoid the question.

Suppose that in response to (Q_A), one of the action-types the observer most desires that I perform is promise-keeping (and they also desire that I refrain from promise-breaking). In response to (Q_C), the character trait they most desire I have is kindness. And, in response to (Q_S), they most desire that I do whatever would lead to the most overall welfare. Although I will argue in chapter 5 that BIO_{RD} is a purely evaluative theory, the evaluations in each case are different. In order to act in accordance with the observer's action-type evaluation, we must keep our promise and make our friend

[15] All these questions concern desires, but of course any appropriate non-truth-oriented attitude can be substituted-in.

miserable and worse-off in the long run. To act in accordance with the character-trait evaluation, we must be silent, failing to keep our promise, but not making our friend as worse off as they could be. Finally, to act in accordance with the state-of-affairs-evaluation, we must be dishonest and fail to be as kind as we could have been. What to do?

It may seem that the only fact to which one could appeal in order to resolve this dispute is the psychology of the observer. Thus, we might ask them 'what do you most desire that I act in accordance with: your desires about my action-types, character-traits or the state of affairs I bring about?' Whichever of the three the observer prefers settles the question, does it not? Matters are not so simple. Asking the observer this higher-order question begs the question against the deontological and virtue ethical approach. The higher-order question concerns what the observer would most desire that I do on a particular occasion. Yet, this is simply another way of asking (Q_S); which state of affairs do you most desire I bring about? And, on any given occasion, the observer will most desire that I do that which maximises aggregate welfare. Determining the relative importance of each question, (Q_A), (Q_C) and (Q_S) cannot be done, therefore, by consulting the observer.

As such, I want to suggest that there is no further fact about whether it is more important to consult the observer with respect to their non-truth-oriented attitudes regarding action-types, character traits, or states of affairs. We may consult the observer

about their desires regarding each of these different objects of value and respond accordingly. Thus, despite being monistic at the level of essential value, BIO$_{RD}$ is deeply pluralistic at the level of non-essential value. This result may seem especially surprising given the emphasis on states of affairs discussed at length earlier in this chapter. I claimed that the benevolent observer's final care directed towards welfare was most accurately characterised as a final care directed towards states of affairs containing welfare. But settling the question of the objects of essential value has little bearing on non-essential value. We cannot move from the fact that states of affairs involving welfare are the bearers of essential value to the claim that we always ought to prioritise states of affairs when imaginatively consulting the benevolent observer about our conduct.

Let us return to my overweight friend. What is to be done? All we can say is as follows: in light of my commitment to performing certain moral action-types, I should tell the truth.[16] In light of my commitment to being a moral person, I should refrain from speaking on the subject. Finally, in light of my commitment to perform the best action on this occasion, I should lie. Which of these takes precedence? There is no fact that decides

[16] In chapter 5 I argue that, at the most fundamental level, moral value can be characterised in purely comparative language, lacking in all deontological import. Yet, I allow that it may be practically indispensable, when more than mere accuracy is taken into account, to preserve deontic notions like 'ought', 'should', or 'right.'

this question.[17]

Living a moral life is a difficult and messy affair.

4.5. Moving on from BIO$_{RD}$

In this chapter, I continued the task of filling in the response-dependent schema:

RD**:** *x* is morally valuable if and only if and because *x* elicits *R* from *S*

where *S* is the benevolent observer described in chapter 3, *R* are the relevant reactions and *x* are the objects of value. I suggested that the relevant reactions are any non-truth-oriented attitude. This is because, given psychological minimalism, all such attitudes will be borne from a final care directed towards welfare, which I have argued is the only essential moral value. Therefore, there can be no reason to exclude any such attitude.

I then discussed the objects of essential value, which I claimed were states of affairs involving welfare. I compared this view with Coons' Ideal Care theory, according to which moral value is determined by the reactions of an observer who cared about

[17] It may seem inconsistent to claim that 'I ought to *x*, *y* and *z*' where *x*, *y* and *z* are incompatible. Yet, what is being said is more subtle. It is that 'based on the observer's desires about what action ought to perform I perform, I ought to *x*' and 'based on the observer's desires about what kind of person I am, I ought to *y*.' This is similar to the claim that different beliefs might be warranted on the basis of different sets of evidence, which could not be inconsistent. The analgous further claim I make is that there is no fact of the matter about what set of evidence one ought to pay attention to. Once again, chapter 5 considers 'oughts' and BIO$_{RD}$ in more detail.

welfare for the sake of the individuals whose welfare it is. I responded to three of Coons' objections, including arguing that we ought to embrace the Repugnant Conclusion. Finally, I considered the objects of non-essential value, once again recommending a liberal approach, despite its revisionary consequences.

Thus, I have completed the task I set for myself at the start of this dissertation to defend the following theory:

> BIO$_{RD}$**:** an object, *x*, is morally valuable if and only if and because it elicits a non-truth-oriented attitude, *R*, from a properly informed, rational, benevolent and otherwise minimal observer, *S*.

In chapter 2, I defended the response-dependent schema, RD, of which BIO$_{RD}$ is an instance. In chapter 3, I made the case that *S* is best conceived of as a properly informed, rational, benevolent and otherwise minimal observer. In this chapter, I discussed the reactions, *R*, and objects of value, *x*. Thus, the form and elements of the schema have been defended.

I might, therefore, stop here. But in the next chapter, I begin a discussion of how accepting BIO$_{RD}$ recommends revising our moral discourse in a somewhat radical direction. However, this should not be seen as part of BIO$_{RD}$ but an interpretation of it. I believe it is the most natural interpretation, but it is by no means essential.

Chapter 5: Moral discourse and the ideal observer

Thus far, I have defended the following thesis:

BIO$_{RD}$**:** an object, *x*, is morally valuable if and only if and because it elicits a non-truth-oriented attitude, *R*, from a properly informed, rational, benevolent and otherwise minimal observer, *S*.

I have suggested that welfare is the only essential moral value and, therefore, that an observer who determines moral value must have a final care (a care for its own sake) directed towards welfare. BIO$_{RD}$ is intended to answer what, in chapter 1, I called t*he central question*: what, if anything, is moral value? Any answer to the central question will have far reaching consequences for the rest of moral theory. In this chapter, I consider one such consequence: what, if any, is the effect of adopting BIO$_{RD}$ on our moral discourse?

5.1. Eliminativism about moral discourse

I have already briefly discussed the relationship between BIO$_{RD}$ and moral discourse in chapter 4. There, I defended the view that all and only non-truth-oriented attitudes are relevant for determining moral value. It was suggested that one might limit the set of relevant non-truth-oriented attitudes by prioritising certain terms in our moral discourse. For instance, we might decide, in advance, what the important moral terms

are and then look for suitable correlates of these terms in the psychology of the ideal observer. This approach was rejected in favour of a more liberal policy. *All* non-truth-oriented attitudes are relevant, and if our moral discourse lacks a particular term that corresponds to some aspect of the observer's psychological state, then so much the worse for our moral discourse. The guiding principle behind this stance is that the psychology of the ideal observer should be primary. After all, it's the observer's psychology that determines what is valuable.

This strategy is largely possible thanks to the fact that BIO$_{RD}$ is an exercise in conceptual ethics, rather than conceptual analysis. BIO$_{RD}$ is a card-carrying revisionary ethical theory. As such, there is a degree of theoretical freedom absent in most ethical theorising. For example, it is commonplace to start by assuming that there must be some moral fact that corresponds to the notion of 'rightness', for example, and then attempt to provide a metaphysical justification for our use of that term. But I have made no such assumption. My starting point was the concept MORAL VALUE**.** This was intended to be as ecumenical and generic as possible. Yet, others may insist that this neutrality is merely illusory. Naturally, for any cognitivist approach that avoids claiming moral discourse is in systematic error, there must be some fundamental normative starting point, or else one would have violated Hume's 'ought-is' dictum. Why not start with rightness and analyse moral value in those terms?

The debate regarding the priority of the good (or the right) stretches back at least to

Ross (1930) and arguably back to Kant. Like much in this area of moral theory, I suspect that there is no fact of the matter that either side can appeal to in order to settle the dispute. As I suggested in section 2.3, moral theory is not a matter of describing an external world of moral facts. Instead, it is about squaring our cares and commitments with our nature as conscious creatures, capable of experiencing great emotional highs and lows, and with the natural world both as it appears to us in conscious experience, and as it is revealed by the methods of the natural sciences. Moral theory is a matter of describing a normative universe one can, on reflection, accept in light of these constraints. If one wishes to insist on the priority of the right over the good, or the evaluative over the deontic,[1] then such a theory must be examined on own its own merits, in comparison to others, and the careful and reflective reader may decide which to accept. Little else can be done besides this.

Thus, my position has been to begin with the traditionally evaluative notion of moral value. And since value is determined by psychology, the place to start questioning how BIO_{RD} might impact our discourse is by examining the psychology of the ideal observer, in particular, their non-truth-oriented attitudes. The paradigmatic example of such an attitude is desire. Desires, though they represent a certain state of affairs, do not aim to represent the world as it is. They are non-truth-oriented (that is, they have no

[1] See, for example, Wedgwood (2009).

constitutive standard of correctness aimed at truth). How do we best translate between descriptive language regarding the psychology of an observer and the normative language required to guide our action? That is, how are we to understand, in moral terminology, the fact that the observer *desires* that I treat you kindly, for example?

One obvious answer is that this entails that it is *right* that I treat you kindly. But this cannot be so. The observer may also have some desire that I treat you indifferently. But it cannot be right that I treat you indifferently and right that I treat you kindly (or at least, we can imagine some case where it isn't right for me to do both, yet the observer desires both of these options). The failure of this simple suggestion is because it neglects two basic facts: that desires come in varying degrees of strength and that desires can conflict without there being anything wrong or incoherent about one's mental states. Suppose that we are concerned with the observer's desires about how I treat you on a particular occasion. The three possible options are {treat you kindly, treat you indifferently, treat you nastily}. And let us further suppose that the observer desires that I treat you kindly most of all, indifferently less than kindly but more than nastily, and nastily least of all. In this case, the best normative language will be comparative. That is, of the three options {treating you kindly, treating you indifferently, treating you nastily}, treating you kindly is best, indifferently worse than kindly but better than nastily, which is worst of all.

'Better-than' and its correlates ('worse-than' and 'the same as') are, I suggest, the

fundamental normative terminology that we ought to adopt after accepting BIO$_{RD}$. I will call this 'comparative language' for short. In fact, I wish to go further and suggest that comparative language is the *only* tool necessary for fully describing the realm of moral value. This is a radical (and revisionary) proposal, since it does-away with all deontic terminology, such as 'rightness', 'duty' and 'obligation.' Nonetheless, I want to suggest that comparative talk is sufficient, and thus that considerations of simplicity suggest that we need not adopt any other normative language.

Though this view is radical, it is not without allies. Alastair Norcross has argued for a normative ethical theory he calls 'scalar utilitarianism' (Norcross 2006a, 2006b). According to scalar utilitarianism, morality makes no demands on us. It does not tell us what we ought to do but instead simply ranks outcomes. Morality, according to Norcross, is purely *evaluative.* It tells us which actions and outcomes would be better relative to other actions and outcomes, but no more. As a utilitarian, Norcross suggests that the correct rankings are made on the basis of utility. The state(s) of affairs with the greatest overall utility is best of all, those with the second greatest utility one ranking below that, and so on. One can construct a similar ranking, not with utility, but with the strength of the observer's attitudes. To repeat, the observer may desire that I treat you kindly more than they desire that I treat you nastily, with indifference in-between. A set of desires such as this allows us to make certain claims about the relative moral value of these actions. Out of the set {treating you kindly, treating you indifferently, treating you

nastily}, kindness is best, indifference worse than kindness but better the nastiness, which is worst of all. We are entitled to make these claims because the observer's desires are guaranteed to have stemmed from their benevolence (since they are otherwise minimal) and because benevolence is directed towards the sole essential value, welfare.

There are, however, complications. BIO$_{RD}$ does not (necessarily) deliver one unique ranking, but multiple rankings across at least two dimensions: the number of attitude-types and the number of observers.

Since all the observer's non-truth-oriented responses are relevant, we must consider the relative strength each of these attitudes, where they exist.[2] In the previous example, I provided a ranking of the options {treating you kindly, treating you indifferently, treating you nastily} as determined by the desires of the ideal observer. Equally, the observer may have wishes with respect to each of these actions, and the relative strength of these wishes are likely the same as the desires. But there may be cases in which rankings differ. Out of the set {A, B, C} the observer's desire-ranking may be [A > B > C], but their wish-ranking may be [C > B > A]. Admittedly, it is difficult to conceive of scenarios where desires and wishes do not align, but the point being made is only that the theory should allow for such discrepancies. If a relevantly informed and rational

[2] We do not require a ranking for every attitude in every context. Although desires, wishes, fears, loves etc. may all be relevant, the observer may only have wishes and desires with respect to one set of propositions, but lack any fears. There is no need to create an artificial 'fear' ranking for the sake of completeness.

observer had such a set of preferences, it would not constitute an objection to BIO$_{RD}$.

The second dimension along which complications may arise is the possibility of disagreement across multiple observers. In section 3.5, I expressed doubts about whether there could be qualitatively distinct benevolent and minimal ideal observers. But I also stated that we cannot rule out the possibility. If there are multiple observers, then they would disagree.[3] Their disagreement would, as above, take the form of differently weighted rankings. Two observers may also have different rankings within a single attitude-type, such as desire, which no single observer can do, trivially.

It may be helpful to envisage the spectrum of complexity as follows:

Most simple	**More complex**[4]	**Most complex**
One observer	Multiple observers	Multiple observers
All attitude-type rankings agree	Attitude-type rankings may agree or disagree	All attitude-type rankings disagree
All cross-observer rankings agree (trivially)	Some cross-observer rankings disagree	All cross-observer rankings disagree

So long as the agent described in chapter 3 is a possible agent, there will be a fact of the matter about which one of these scenarios is true. But, to repeat: aside from my

[3] Recall that I am individuating observers by their responses, rather than by the mechanism which produces those responses. Thus, two distinct observers disagree with respect to their responses by definition.

[4] Another 'more complex' possibility is that there is one observer, but their attitude-type rankings disagree.

earlier remarks expressing sympathy with the single observer view, I take no stance on which one is true. Answering this question is important but it is also extremely difficult and has no overall impact on the plausibility of BIO$_{RD}$, at least conceived of as a revisionary theory.[5] Notice that it is also an empirical question. We could, in theory, construct a benevolent observer (or observers) and note each of their attitudinal rankings. In general, I think philosophers should avoid empirical speculation, hence my reticence to come down firmly on one side of this question.

BIO$_{RD}$, therefore, ranks states of affairs on the basis of the benevolent observer's attitudes. There can be multiple rankings for distinct attitudes, which may or may not align, and there may be multiple observers with distinct responses. These rankings entitle us, as Norcross has suggested, to use comparative normative language. If there is a case where one state of affairs is more highly ranked by one attitude than it is by another attitude, then there is no fact of the matter about which one is better or worse than the other. Notice that this does not mean that there is no fact of the matter about other states of affairs that the observer takes an attitude towards. If, in a ranking of one hundred states of affairs, two states of affairs are swapped such that their rank fifty-one desire is their rank fifty-two wish, and their rank fifty-two desire is their rank-fifty-one wish, but all other rankings align, then we can still infer that rank one (shared across

[5] If BIO$_{RD}$ was not a revisionary theory, this would certainly be a problem, since in ordinary discourse we appear to know plenty of claims about the relative value of different states of affairs.

both attitudes) is better than rank two (also shared across attitudes), for example. However, there will be no fact of the matter about whether fifty-one or fifty-two is better or worse than the other.

If we can describe the observer's (relevant) psychology by reference to the relative strength of particular mental states, and descriptions of this kind entitle us to use comparative normative language, then we can accurately characterise the landscape of moral value using only comparative normative language.[6] All value is determined by mental states that come in degrees, and comparative language is capable of minimally reflecting these degrees in all cases, even when the rankings differ across a single observer's mental state-types.

Because I suggest revising our moral discourse so as to only include comparative language, the theory I'm offering is eliminativist in kind. Most famously, in the philosophy of mind, the eliminativist takes the view that folk psychological terms fail to accurately characterise the actual workings of our brains as described by (an eventually completed, or at least sufficiently accurate) neuroscience. Thus, strictly speaking, there

[6] This can only be done on the assumption that all non-truth-oriented attitudes have degrees of strength. Such a claim is not susceptible to any *a priori* proof, but must be shown through a case-by-case analysis of each and every attitude. If it turns out that there are non-truth-oriented attitudes that do not admit of degrees (I cannot think of one, myself) then this would require special treatment. It is difficult to know what to say about such a case in the absence of a particular example, but I do not believe it would constitute a threat to $BIO_{RD,}$ though it may require altering some of what is said above regarding the appropriateness of normative terminology.

are no beliefs or desires, since, it is argued, there are no clear neurological kinds that correspond to these folk psychological natural kind terms (Churchland 1981). Eliminativism more generally is best understood as consisting of both a descriptive and prescriptive claim. Descriptively, eliminativists say that some area of discourse fails to square with the facts it purports to represent. Prescriptively, it makes the claim that we should replace the original discourse with a more accurate one in light of this failure.

BIO$_{RD}$ is no exception. Descriptively, it claims that moral discourse as it actually is fails to square with the fact that value is determined by the reactions of a benevolent and minimal observer since these reactions provide no sound basis for common deontic terms. Prescriptively, it claims that we should replace ordinary moral discourse with a language where comparative language replaces our current discourse. However, one must be careful about this 'should'. Perhaps other eliminativists take their prescriptive should's to be categorical. Mine is not. We only ought to replace our moral discourse with comparative language, *if* we care about our discourse being accurate, where accuracy is measured by fidelity to the observer's psychological profile, *and nothing else*. Although it may sound obvious that we ought to aim for accuracy, it is not. Indeed, it is false. There are many excellent practical reasons to have a less accurate common moral discourse. BIO$_{RD}$ is not so strange that it prevents any mapping of deontic language onto the psychology of the observer. It's just that these mappings are artificial and, in an important sense, arbitrary. They fail to accurately reflect the complexities of moral value

by requiring us to set hard and fast thresholds, when there are no thresholds in the observer's psychology.

For example, consider rightness. How are we to understand rightness according to BIO$_{RD}$? The most natural answer is that an act is right iff the observer desires it more than any other alternative. And I have no qualms with such a definition. Furthermore, it seems that the concept of 'rightness' will be extremely useful in the deliberative practices of creatures like us, perhaps to the point of indispensability. Yet, rightness, for all its uses, fails to accurately describe the psychology of the observer in the fullest detail. Knowing that an act is right tells us the observer desires it more than any other. But it leaves open the rest of their psychology. Most crucially rightness is ill-equipped to handle degrees, given that it is typically thought not to admit of degrees.[7] To see why, suppose that we only permitted the terms 'rightness' and 'wrongness' in our discourse. Consider a scenario in which I can choose to save a person from a burning building. Let us suppose, too, that the observer's two least desired options are for me to try to do nothing and, below that, to scream and run away (we can suppose that this is the least desired option since it causes distress to bystanders). If we insist on using only 'rightness' to capture the mental states of the observer, then we must describe both these actions as 'right', even though they differ in their degrees of rightness. This seems like

[7] For a dissenting view, see Peterson (2013).

an unnecessarily drastic revision, given that comparative language is far better placed to perform the task. This is also the reason why any definition of 'rightness' is arbitrary. Since 'rightness' is ill-equipped to capture the total psychology of the observer, its definition will be stipulative.

But this arbitrariness need not be at all damaging or worrisome. As already mentioned, there is a strong practical need to provide definitions of common moral terms and some definitions can still be better than others, even without there being a fact of the matter as to which definition is 'correct'. Insofar as we find ourselves in a world that demands action and often requires us to co-ordinate our actions with others to improve our welfare, the notion of an action that is *required* by members of a society seems vital. Although deontic terms, in general, do not sit well with **BIO**RD, some concepts can be better suited to this task than others. One would be making no moral mistake if one defined right action as 'what the observer would hate for us to do on Tuesday evenings', but this definition would be seriously deficient in its needless specificity and ill-suitedness to the task of improving our welfare. There are simply better candidates. But the 'better' here is not a moral one. It is a pragmatic 'better' which means 'better suited to our aims' which are broader than mere accuracy. In addition, there are independent non-moral standards about what constitutes a 'good' meaning for a particular term; meanings should be stable across relevantly similar contexts, for example. Perhaps, in the case of moral terms, a meaning should have some emotional

valence such that others are encouraged to act in accordance with its recommendations. A definition that fails to meet standards of this kind is one we can reasonably reject as practically deficient, without needing to say that it is incorrect.

Thus, although I have advocated for eliminativism, it is eliminativism in a fairly limited context; one in which the only thing we care about is a total isomorphism between our normative language and the psychology of the ideal observer. This may be a context only of interest to moral philosophers, but given its emphasis on accuracy and our typical desire to be accurate, it should, I suggest, occupy an important position in our minds.

5.2. Reasons

There is one area of our ordinary discourse that this eliminativism leaves relatively untouched: reasons. It is fairly harmless to speak of reasons since one can translate between standard comparative language and reasons talk without any inflation, provided one has a comparative understanding of reasons.

What is it to understand reasons as comparative? Some philosophers have endorsed a *contrastive* theory of reasons (Sinnott-Armstrong 2006; Snedegar 2017). According to these philosophers, reasons are always contrastive, such that a reason to φ is always a reason to φ as opposed to ψ. More precisely, a reason in favour of an action is always relative to a contrast class consisting of a set of alternative actions. Out of the contrast

class {φ, ψ}, *x* may be a reason to φ, but out of the contrast class {φ, ψ, *α*}, *x* may be reason to *α*. One way to understand BIO$_{RD}$ is that the contrast class out of which reasons ought to be given is fixed by the set of options over which the benevolent observer has non-truth-oriented attitudes. As such, one can always infer from their non-truth-oriented attitudes that there is some reason for or against a particular action. If the observer desires that I treat you indifferently, then there is some reason to treat you indifferently. We do not know how strong this reason is. There may also be some reason to do this because the observer weakly desires that I treat you indifferently. Their strongest desire may be that I treat you kindly, in which case I have most reason to treat you kindly. A genuine reason to φ is always a reason out of the contrast class {all actions considered by the benevolent observer}.[8]

According to other theories, the contrast class is context-dependent.[9] Similarly, we could have a context sensitive theory of reasons constructed using a more limited set of options considered by the observer. In my discussion of eliminativism in this chapter, I argued that 'better-than' and 'worse-than' were appropriate primitives if what one cared about was accurately reflecting the psychological states of the benevolent observer. I admitted that there may be practical reasons not to care about accuracy, and thus to

[8] In the language of section 3.1.1, this is the set of all facts {F} out of which the set {F$_R$} is constructed.

[9] According to Sinnott-Armstrong (2006), we ought to suspend judgement about what the correct contrast class is.

continue to use deontic terms like 'right-action'. One can take a similar stand with respect to contrast-classes. Although the contrast class {all actions considered by the benevolent observer} is the correct contrast class if one wishes to accurately account for all the reasons that there are, it may be prudent to pick a more limited set in certain circumstances and speak as if there were most reason to perform some action when, in fact, there would have been most reason to perform an alternative action.

If, for example, I am considering whether to join the French resistance or care for my sick mother (Sartre 1948), I might consider what a benevolent observer would most desire that I do with respect to these two options. These two options may constitute the entire contrast class. Furthermore, it may be an appropriate contrast class because considering the third option which would in fact be best (joining the Nazi army and sabotaging it from within, for example) would require too much mental effort. In fact, no reasonable person in my position would even consider this option. In which case, if the observer would desire that I care for my mother more than I join the resistance, then we may wish to speak as if there is 'most reason' to care for my sick mother, when in fact there is most reason to join the Nazi's as a saboteur. I take no stand on what makes certain contrast classes relevant in certain complex contexts. It may be determined by what any rational person in my circumstances could reasonably be expected to consider, or it may be something else entirely. The point is simply that there may be practical reasons not to aim for total accuracy in providing reasons out of the contrast class {all

actions considered by the observer}, even if this contrast class gives an accurate account of the reasons that there are.

The easy translation between comparative talk and reasons is particularly pleasing given that it is now an expectation of any satisfactory theory of value that it accounts for talk of moral reasons. Some meta-ethicists set their sights even higher and offer theories of reasons in general, omitting qualifiers like moral or prudential altogether. The views most seriously committed to the centrality of reasons to moral theory often go under the heading 'reasons-first.' Though some prominent 'reasons-firsters' tend to gravitate towards certain forms of rationalistic non-naturalism, the reasons-first approach, in general, makes no overt commitments of this kind. It's a broad church, capable of sheltering naturalists, absolutists, relativists, objectivists, subjectivists, and even, perhaps, certain kinds of expressivists. The only qualifying criterion is that a theory makes a commitment to reasons being, in some sense of the term that will require further elaboration, *fundamental*.

Though reasons might be natural facts about the world or our own mental states, certain kinds of relations between the same, or peculiar non-natural properties, so long as one insists that other normative facts, properties, or language be analysed in terms of reasons, with reasons themselves admitting of no further normative analysis, one will qualify as a 'reasons-firster.' One consequence of adopting a reasons-first approach is an inability to go on and make any further interesting claims about reasons. They are, after

all, fundamental. Both Parfit and Scanlon suggest, in similar language, that all that can be said about reasons is that they count in favour of something, and if asked what it is for one thing to count in favour of another, the best one can do is to respond with 'by being a reason, of course' (Scanlon 1998; Parfit 2011a, 2011b). If we are to retain the Humean division between facts on the one hand and values on the other, then this is not necessarily a bad or surprising outcome. The suggestion that certain moral concepts or properties admit of no further analysis is old. Moore expressed it most famously, with his favoured unanalysable term being 'the good'. If normativity bottoms out somewhere, why not with reasons? The answer is simply because illuminating things *can* be said about reasons, beyond what the reasons-first approach offers. BIO$_{RD,}$ and response-dependent theories more generally, have something to tell us about reasons.

As a response-dependent theory of value, BIO$_{RD}$ is an instance of the following general schema:

RD: *x* is morally valuable if and only if and because *x* elicits *R* from *S.*

The step from RD to a theory of reasons is relatively simple: moral reasons are just the set of facts which are relevant to determining *S*'s reaction *R*.[10] A brief argument for

[10] There are good arguments for conceiving of reasons as propositions, rather than facts (see, for example, Singh (*forthcoming*)). Although I speak of facts, I do not mean to commit myself to any particular position in this argument. If, for independent considerations, reasons are best thought of as propositions and not as

this claim is as follows: subject *S* bases her reaction *R* on all the reasons if and only if she bases her reaction *R* on all the relevant information. All the relevant information just is the information about all the relevant facts. So, *S* bases *R* on all the reasons if and only if *S* bases *R* on all the relevant facts. The suggestion is that this is because the relevant facts just are the reasons. And the relevant information can constitute the relevant facts regardless of whether we accept a response-dependent theory or not. However, if we combine this conclusion with the response-dependent schema, RD, we arrive at the following:

(θ) the reasons are the facts upon which subject *S* bases her reaction *R*.

Notice that (θ) is perfectly general in that it can apply to *any* theory that fits the response-dependent schema, not just BIO$_{RD}$.

We might begin to untangle (θ) by focusing on what it is for *S* to *base* her attitude on some fact. Considerable work has been done on the basing relation in epistemology. These philosophers ask: what it is for a belief to be based on a (normative) reason? We can modify the epistemologist's question as follows: what is it for some attitude to be based on a fact? The question is rather similar, since we're understanding reasons as facts and belief, of course, is but one attitude which may or may not constitute a relevant

---

facts, then the reader is free to translate what I say about facts into sentences about propositions corresponding to those facts.

response depending on the particular version of the response-dependent theory one endorses. In the case of BIO_{RD}, of course, belief is not relevant, but one hopes that what is said about belief may be said about desires and other non-truth-oriented attitudes. Unfortunately, there is no universally agreed upon account of the basing relation and this is not the place to arbitrate the dispute. Advocates of a particular account should feel free to substitute it into (θ).

It is worth exploring, however, the *counterfactual interpretation* of (θ). According to this interpretation, some fact *f* is a reason if and only if, if *S* had known *f,* then the relevant reaction *R* would have changed. Unfortunately, this simple proposal cannot be correct. Consider again the example from section 3.1.1; if a husband buys his wife flowers, it will make her happy and make her laugh, since the flowers remind her of some inside joke. The fact that the flowers will make her happy is a reason to buy them for her and so is the fact that they will make her laugh. Now suppose that *S* first comes to know that flowers will make my wife happy. Afterwards, she learns that flowers will also make her laugh. In that case, the relevant response *R*, may not change. Supposing, for the sake of argument, that desires are the relevant sort of response. It is plausible that *S* will desire that the husband gives his wife flowers after she learns that they will make her happy, and the same desire will persist after *S* learns that the flowers will also make her laugh. Thus, learning that flowers will make her laugh does not change *S*'s reaction even though it is, I submit, a reason.

The natural response on behalf of the counterfactual interpretation is that the wife still would have laughed if the husband had given her the flowers, even if they wouldn't have also made her happy. To repeat: the problem with this suggestion is that this counterfactual might be false. The closest world in which the husband gives his wife flowers and she laughs but this action doesn't make her happy might be a world in which giving her flowers makes her miserable by reminding her of a traumatic event (her laugh might occur moments before she recalls the troubling events). More abstractly, in cases where f_1 and f_2 occur and are, in the actual world, individually sufficient for *S* to produce one response, the closest possible world in which f_2 obtains but f_1 fails to obtain might be one in which *S*'s attitude does not change because some other, far stronger attitude outweighs any change that might have otherwise obtained if that stronger attitude (in this case misery) hadn't obtained.

For the reasons given in section 3.1.1, the modified Brandt proposal solves this problem. As well as providing an account of relevant knowledge, the proposal also provides an account of basing sufficiently detailed for our present purposes. Combining the response-dependent insight about reasons with the modified Brandt proposal delivers the following view:

Some fact, *f*, is a reason if and only if, when

(i) after the permutations, p_1, p_2., …, p_n., of the power set, $\wp\{F\}$[11] are added one-by-one to subject *S*'s belief set, and

(ii) after each permutation, p_n., is added, *S*'s response *R* is noted before beliefs are removed one-by-one in some fixed order,

(iii) if *S*'s response *R* changes after a belief is removed, then the fact that was the content of the removed belief is a reason.

We can now modify (θ) to fit BIO_{RD}, understanding the basing relation as described by the modified Brandt proposal:

(θ_{BIO}) the reasons are the facts upon which a benevolent, relevantly informed, rational and otherwise minimal observer bases their non-truth-oriented attitudes.[12]

---

[11] $\wp\{F\}$, the reader will recall, is the powerset of the set of every fact, {F}. According to the modified Brandt proposal, the ideal observer considers every possible permutation of the powerset $\wp\{F\}$ so that every fact is considered in every possible order. Any fact that makes a difference to the observer's non-truth-oriented attitudes is relevant. See section 3.1.1 for the details.

[12] One worry with this proposal is that it includes too many facts. For instance, if small microscopic events make a difference, aren't they reasons? Yet, there is no principled way of excluding these facts. This is due to what Sinnott-Armstrong has called 'General Substitutability' (Sinnott-Armstrong 1992). Let us say, with Sinnott-Armstrong, that "doing Y enables an agent to do X if and only if Y is part of a larger course of action that is sufficient for the agent to do X, and the agent can do the other acts that make up what is sufficient for X." Then general substitutability says that "If there is a reason for A to do X, and if A cannot do X without doing Y, and if doing Y will enable A to do X, then there is a reason for A to do Y." Suppose that I have a

Following the modified Brandt proposal, we know that when a fact makes a difference in the sense revealed by that proposal, this fact is a reason. Yet reasons are more complicated than this in that they have structure. A reason can be a reason *for* or *against* some action and a reason can be a reason for some agent but not for another agent. For instance, the fact that my family lives in London is a reason *for me* in *favour* of travelling to London. It is not a reason for anyone else to perform or refrain from performing that action (at least not under that description).

The ability of reasons to be for or against some action can be accounted for by the response-dependent theory as follows: whether or not a reason is a reason for or against some action φ can be determined in two ways. The first is the valence of the attitude. Once again, the point is most easily illustrated if we restrict the set of relevant attitudes to desires and aversions. If *S* has a positively-valanced desire that some agent(s), *A* φ's, then there is a reason in favour of φ-ing. If *S* has a negatively-valanced desire (aversion)

---

reason to drive you to the hospital. Opening the car door enables me to do so and is causally necessary. Then, according to general substitutability, I have a reason to open the car door. But the argument goes all the way down. If I have a reason to open the car door, then I have a reason to move certain air molecules out the way with my hand as I reach for the door, and so on, so long as this enables and is necessary for me to open the car door. The upshot is that reasons can be as fine-grained as fine-grained facts can be. **BIO**$_{RD}$ explains General Substitubaility since any observer who desired that you opened the care door, must also desire that you move your hand in a certain way, and so on.

that some agent(s) *A*, φ's, then there is a reason against φ-ing. The second way is *via* negation. If *S* has a positively-valanced desire that *A*(s) not-φ, then there is a reason not to φ. Similarly, if *S* has a negatively-valanced desire that *A*(s) not-φ, then there is a reason to φ.[13]

The other important structural feature of reasons, that they can be reasons *for* an individual or a group of people, can be explained in much the same way as valence. Whether or not some fact *f* is a reason for one agent, a collection of agents, or all agents, is determined by the content of the propositional attitude that is *S*'s response *R*. Taking desires as our main example, if *S* desires that *A* φ in response to fact *f*, then *f* is a reason for A (to φ). Similarly, if *S* desires that every member of some group *G*, φ in response to fact *f*, then *f* is a reason for everyone in *G* (to φ), etc.

In this way, **BIO**$_{\text{RD}}$ is capable of being expressed in reasons-talk with relative ease. Although the theory is contrastive, it is capable of explaining why we might frequently talk as if reasons were not contrastive: we are implicitly assuming that the contrast class is the widest possible one out of which the observer considers all possible actions. In this context, the observer's strong desire that I be kind to someone gives me a reason to be kind. This is not an explicitly contrastive statement, but this is merely because the

[13] A critical point to make clear is that the fact that *S* has or would have response *R* is not a reason of any kind. Facts 'out there' in the world are reasons and *S*'s reaction determines which of those facts are reasons.

background contrast class is assumed to be the widest one. If we wish to limit it for other reasons, we are free to do so.

Chapter 6: Concluding remarks

The reader now has before them the entirety of BIO$_{RD}$ and has seen one important consequence of adopting that theory. Still, there are a number of important questions that remain: what is the impact of BIO$_{RD}$ on our moral epistemology? Given our impoverished epistemic state compared to the observer's, can we ever know what is morally valuable? How does BIO$_{RD}$ relate to other kinds of value, like prudential or aesthetic value? I'm hopeful that illuminating answers can be offered to these questions, but that will have to wait for another time.

As I stated in section 1.3, my hope is that BIO$_{RD}$ articulates something that those inclined towards utilitarianism have suspected all along, but either had not thought of themselves or had wrongly assumed was already established. To those not so-inclined, I hope that I have illustrated what their interlocutors find persuasive about utilitarian theories and, more importantly, mapped more clearly the areas of contention so that we can better disagree with one another.

Ultimately, I hope to have shown that utilitarianism, often viewed as a cold and simplistic picture of the moral landscape, is actually complex and best thought of as flowing from benevolence, an attitude which is, in Hutcheson's sense, a kind of love. Far from rejecting our humanity, utilitarianism embraces its best parts.

Bibliography

Anderson, Elizabeth. 1993. *Value in Ethics and Economics*. Harvard University Press.

Arpaly, Nomy, and Timothy Schroeder. 2013. *In Praise of Desire*. Oxford University Press.

Ayer, A J. 1936. *Language, Truth and Logic*. V. Gollancz.

Blackburn, Simon. 1984. *Spreading the Word: Groundings in the Philosophy of Language*. Oxford University Press.

———. 1985. "Errors and the Phenomology of Value." In *Morality and the Good Life*, edited by Thomas L Carson and Paul K Moser, 324–37. Oxford University Press.

———. 1993. "Circles, Finks, Smells and Biconditionals." *Philosophical Perspectives*.

———. 1998. *Ruling Passions: A Theory of Practical Reasoning*. Oxford University Press.

Brandt, Richard B. 1954a. "Some Comments on Professor Firth's Reply." *Philosophy and Phenomenological Research* 15 (3): 422–23.

———. 1954b. "The Definition of an 'Ideal Observer' Theory in Ethics." *Philosophy and Phenomenological Research* 15 (3): 407–13.

———. 1955. *Hopi Ethics*. University of Chicago Press.

———. 1959. *Ethical Theory*. Englewood Cliffs.

———. 1979. *A Theory of the Good and the Right*. Prometheus Books.

Brink, David O. 1986. "Externalist Moral Realism." *Southern Journal of Philosophy* 24 (S1): 23–41.

———. 1989. *Moral Realism and the Foundations of Ethics*. Cambridge University Press.

Broad, C D. 1930. *Five Types of Ethical Theory*. Kegan Paul, Trench, Trubner & Co Ltd.

———. 1944. "Some Reflections on Moral-Sense Theories in Ethics." In *Proceedings of the*

Aristotelian Society, 45:131–66. Wiley-Blackwell.

Broome, John. 2004. *Weighing Lives*. Oxford University Press.

Brower, Bruce W. 1993. "Dispositional Ethical Realism." *Ethics* 103 (2): 221–49.

Burgess, Alexis, and David Plunkett. 2013a. "Conceptual Ethics I." *Philosophy Compass* 8 (12): 1091–1101.

———. 2013b. "Conceptual Ethics II." *Philosophy Compass* 8 (12): 1102–10.

Cappelen, Herman. 2018. *Fixing Language: An Essay on Conceptual Engineering*. Oxford: Oxford University Press.

Chisholm, Roderick M. 1986. *Brentano and Intrinsic Value*. Cambridge University Press.

Christensen, David. 2009. "Disagreement as Evidence: The Epistemology of Controversy." *Philosophy Compass* 4 (5): 756–67.

Churchland, Paul M. 1981. "Eliminative Materialism and the Propositional Attitudes." *Journal of Philosophy* 78 (February): 67–90.

Cohen, G. A. 1996. "Reason, Humanity and the Moral Law." In *The Sources of Normativity*, 167–88. Cambridge University Press.

Coons, Christian. 2006. *The Value of Individuals and the Value of States of Affairs*. Dissertation.

———. 2012. "The Best Expression of Welfarism." In *Oxford Studies in Normative Ethics*, edited by Mark C Timmons, 206–37. Oxford University Press.

Copp, David. 1997. "Belief, Reason, and Motivation: Michael Smith's 'the Moral Problem.'" *Ethics* 108 (1): 33–54.

D'Arms, Justin, and Daniel Jacobson. 2000. "Sentiment and Value." *Ethics* 110 (4): 722–48.

Darwall, Stephen. 2002. *Welfare and Rational Care*. Princeton University Press.

Davidson, Donald. 1970. “Mental Events.” In *Essays on Actions and Events*, edited by L Foster and J W Swanson, 207–24. Clarendon Press.

Dennett, Daniel C. 1987. *The Intentional Stance*. MIT Press.

Dworkin, Ronald. 1986. *Law’s Empire*. Harvard University Press.

Elga, Adam. 2007. “How to Disagree About How to Disagree.” In *Disagreement*, edited by Ted Warfield and Richard Feldman, 175–86. Oxford, UK: Oxford University Press.

Engel, Pascal. 2013. “Doxastic Correctness.” *Aristotelian Society Supplementary Volume* 87 (1): 199–216.

Enoch, David. 2005. “Why Idealize?” *Ethics* 115 (4): 759–87.

———. 2010. “The Epistemological Challenge to Metanormative Realism: How Best to Understand It, and How to Cope with It.” *Philosophical Studies* 148 (3): 413–38.

Finlay, Stephen, and Mark Schroeder. 2017. “Reasons for Action: Internal vs. External.” Edited by Edward N Zalta. *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

Firth, Roderick. 1951. “Ethical Absolutism and the Ideal Observer.” *Philosophy and Phenomenological Research* 12 (3): 317–45.

———. 1954. “Reply to Professor Brandt.” *Philosophy and Phenomenological Research* 15 (3): 414–21.

———. 1978. “Comments on Professor Postow’s Paper.” *Philosophy and Phenomenological Research* 39 (1): 122–23.

Fischhoff, Baruch, Paul Slovic, Sarah Lichtenstein, Stephen Read, and Barbara Combs. 1978. “How Safe Is Safe Enough? A Psychometric Study of Attitudes towards Technological Risks and Benefits.” *Policy Sciences* 9 (2): 127–52.

Foot, Philippa. 1972. “Morality as a System of Hypothetical Imperatives.” *Philosophical Review* 81 (3): 305–16.

Frances, Bryan, and Jonathan Matheson. 2018. “Disagreement.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N Zalta, Spring 201. Metaphysics Research Lab, Stanford University.

Gallie, W B. 1955. “Essentially Contested Concepts.” *Proceedings of the Aristotelian Society* 56 (1): 167–98.

Gert, Bernard. 1998. *Morality: Its Nature and Justification*. Oxford University Press.

———. 2004. *Common Morality: Deciding What to Do*. Oxford University Press.

Gettier, Edmund. 1963. “Is Justified True Belief Knowledge?” *Analysis* 23 (6): 121–23.

Gibbard, Allan. 1990. *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. Harvard University Press.

Greaves, Hilary. 2017. “A Reconsideration of the Harsanyi--Sen--Weymark Debate on Utilitarianism.” *Utilitas* 29 (2): 175–213.

Hare, R. M. 1952. *The Language of Morals*. Oxford Clarendon Press.

———. 1972. “Rules of War and Moral Reasoning.” *Philosophy and Public Affairs* 1 (2): 166–81.

———. 1981. *Moral Thinking: Its Levels, Method, and Point*. Oxford: Oxford University Press.

———. 1988. “Possible People.” *Bioethics* 2 (4): 279–93.

Harrison, Jonathan. 1956. “Some Comments on Professor Firth’s Ideal Observer Theory.” *Philosophy and Phenomenological Research* 17 (2): 256–62.

———. 1971. *Our Knowledge of Right and Wrong*. New York: Humanities Press.

Harsanyi, John. 1977. “Morality and the Theory of Rational Behavior.” *Social Research* 44 (4): 623–56.

———. 1986. *Rational Behaviour and Bargaining Equilibrium in Games and Social Situations*. Cambridge University Press.

Hart, H L A. 1961. *The Concept of Law*. Oxford University Press.

Held, Virginia. 2005. *The Ethics of Care: Personal, Political, and Global*. Oxford University Press.

Henne, Paul, Vladimir Chituc, Felipe De Brigard, and Walter Sinnott-Armstrong. 2016. "An Empirical Refutation of 'Ought' Implies 'Can.'" *Analysis* 76 (3): 283–90.

Henson, Richard G. 1956. "On Being Ideal." *Philosophical Review* 65 (3): 389–400.

Hospers, John. 1972. *Human Conduct; Problems of Ethics*. Harcourt Brace Jovanovich.

Hruschka, Joachim. 1991. "The Greatest Happiness Principle and Other Early German Anticipations of Utilitarian Theory." *Utilitas* 3 (2): 165–77.

Huemer, Michael. 2008. "In Defence of Repugnance." *Mind* 117 (468): 899–933.

———. 2011. "Epistemological Egoism and Agent-Centered Norms." In *Evidentialism and Its Discontents*, edited by Trent Dougherty, 17. Oxford University Press.

Jackson, Frank. 1998. *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford University Press.

———. 2001. "Précis of From Metaphysics to Ethics." *Philosophy and Phenomenological Research* 62 (3): 617–24.

Jackson, Frank, and Philip Pettit. 2002. "Response-Dependence Without Tears." *Noûs* 36 (1): 97–117.

Kagan, Shelly. 1998. "Rethinking Intrinsic Value." *The Journal of Ethics* 2 (4): 277–97.

Kauppinen, Antti. 2017. "Moral Sentimentalism." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N Zalta, Spring 201. Metaphysics Research Lab, Stanford University.

Kitcher, Philip. 2011. *The Ethical Project*. Harvard University Press.

Kneale, William. 1950. "Objectivity in Morals." *Philosophy* 25 (93): 149–66.

Korsgaard, Christine M. 1983. "Two Distinctions in Goodness." *Philosophical Review* 92 (2): 169–95.

———. 1986. "Skepticism About Practical Reason." *Journal of Philosophy* 83 (1): 5–25.

———. 1996. *The Sources of Normativity*. Cambridge University Press.

Kripke, Saul. 1980. *Naming and Necessity*. Harvard University Press.

LeBar, Mark. 2013. *The Value of Living Well*. Oxford University Press.

Lemos, Noah M. 2005. "The Bearers of Intrinsic Value." In *Recent Work on Intrinsic Value*, edited by Toni Ronnow-Rasmussen and Michael J Zimmerman, 181–90. Springer.

Lewis, David. 1973. *Counterfactuals*. Blackwell.

———. 1983. *Philosophical Papers*. Oxford University Press.

———. 1986. *On the Plurality of Worlds*. Wiley-Blackwell.

Lin, Eden. 2018. "Welfare Invariabilism." *Ethics* 128 (2): 320–45.

Loeb, Don. 1995. "Full-Information Theories of Individual Good." *Social Theory and Practice* 21 (1): 1–30.

Mackie, J L. 1985. *Persons and Values*. Clarendon Press.

Mackie, John L. 1977. *Ethics: Inventing Right and Wrong*. Penguin Books.

McDowell, John. 1985. "Values and Secondary Qualities." In *Morality and Objectivity*, edited by Ted Honderich, 110–29. London: Routledge.

McHugh, Conor, and Jonathan Way. 2016. "Fittingness First." *Ethics* 126 (3): 575–606.

McHugh, Conor, and Daniel Whiting. 2014. "The Normativity of Belief." *Analysis* 74 (4): 698–713.

Nagel, Thomas. 1970. *The Possibility of Altruism*. Oxford Clarendon Press.

———. 1974. “What Is It Like to Be a Bat?” *Philosophical Review* 83 (October): 435–50.

Narveson, Jan. 1973. “Moral Problems of Population.” *The Monist* 57 (1): 62–86.

Nesbitt, Winston. 1992. “Utilitarianism and Benevolence.” *Cogito* 6 (3): 170–72.

Noddings, Nel. 2013. *Caring: A Relational Approach to Ethics and Moral Education*. University of California Press.

Nolan, Daniel. 1997. “Impossible Worlds: A Modest Approach.” *Notre Dame Journal of Formal Logic* 38 (4): 535–72.

Norcross, Alastair. 1997. “Good and Bad Actions.” *Philosophical Review* 106 (1): 1–34.

———. 2006a. “Reasons Without Demands: Rethinking Rightness.” In *Contemporary Debates in Moral Theory*, edited by James Lawrence Dreier, 6–38. Blackwell.

———. 2006b. “The Scalar Approach to Utilitarianism.” In *The Blackwell Guide to Mill’s Utilitarianism*, edited by Henry West, 217–32. Wiley-Blackwell.

Olson, Jonas. 2014. *Moral Error Theory: History, Critique, Defence*. Oxford University Press.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford University Press.

———. 2011a. *On What Matters: Volume One*. Oxford University Press.

———. 2011b. *On What Matters: Volume Two*. Oxford University Press.

Peterson, Martin. 2013. *The Dimensions of Consequentialism: Ethics, Equality and Risk*. Cambridge University Press.

Pettit, P. 1991. “Realism and Response-Dependence.” *Mind* 100 (4): 587–626.

Postow, B C. 1978. “Ethical Relativism and the Ideal Observer.” *Philosophy and Phenomenological Research* 39 (1): 120–21.

Price, Huw. 2013. *Expressivism, Pragmatism and Representationalism*. Cambridge University Press.

Rabinowicz, Wlodek, and Jan Österberg. 1996. "Value Based on Preferences." *Economics and Philosophy* 12 (1): 1–27.

Rabinowicz, Wlodek, and Toni Ronnow-Rasmussen. 2004. "The Strike of the Demon: On Fitting Pro-attitudes and Value." *Ethics* 114 (3): 391–423.

Rachels, James. 1986. *The Elements of Moral Philosophy*. McGraw-Hill.

Rachels, Stuart. 2004. "Repugnance or Intransitivity: A Repugnant But Forced Choice." In *The Repugnant Conclusion: Essays on Population Ethics*, edited by Jesper Ryberg Torbjorn Tannsjo. Kluwer Academic Publishers.

Radcliffe, Elizabeth S. 2004. "Love and Benevolence in Hutcheson's and Hume's Theories of the Passions." *British Journal for the History of Philosophy* 12 (4): 631–53.

Railton, Peter. 1984. "Alienation, Consequentialism, and the Demands of Morality." *Philosophy and Public Affairs* 13 (2): 134–71.

———. 1986. "Moral Realism." *Philosophical Review* 95 (2): 163–207.

———. 1989. "Naturalism and Prescriptivity." *Social Philosophy and Policy* 7 (1): 151–71.

———. 1998. "Red, Bitter, Good." In *European Review of Philosophy, Volume 3: Response-Dependence*. Stanford: CSLI Publications.

Rawls, John. 1971. *A Theory of Justice*. Belknap Press of Harvard University Press.

Rosati, Connie S. 1995. "Persons, Perspectives, and Full Information Accounts of the Good." *Ethics* 105 (2): 296–325.

Ross, W D. 1930. *The Right and the Good*. Clarendon Press.

Ryberg, Jesper. 1996. "Is the Repugnant Conclusion Repugnant?" *Philosophical Papers* 25 (3): 161–77.

Sainsbury, R M. 1980. "Benevolence and Evil." *Australasian Journal of Philosophy* 58 (2): 128–34.

Sartre, Jean-Paul. 1948. “Existentialism Is a Humanism.” *Philosophy: Key Texts*, 115.

Sartwell, Crispin. 1991. “Knowledge Is Merely True Belief.” *American Philosophical Quarterly* 28 (2): 157–65.

———. 1992. “Why Knowledge Is Merely True Belief.” *Journal of Philosophy* 89 (4): 167–80.

Sayre-McCord, Geoffrey. 1994. “On Why Hume’s General Point of View Isn’t Ideal - and Shouldn’t Be.” *Social Philosophy and Policy* 11 (01): 202–28.

Scanlon, Thomas. 1998. *What We Owe to Each Other*. Belknap Press.

Shope, Robert K. 1983. *The Analysis of Knowing: A Decade of Research*. Princeton University Press.

Singh, Keshav. n.d. “Acting and Believing Under the Guise of Normative Reasons.” *Philosophy and Phenomenological Research*.

Sinnott-Armstrong, Walter. 1984. “`Ought’ Conversationally Implies `Can’.” *Philosophical Review* 93 (2): 249–61.

———. 1992. “An Argument for Consequentialism.” *Philosophical Perspectives* 6: 399–421.

———. 2006. *Moral Skepticisms*. Oxford University Press.

Smart, J. J. C. 1968. “Quine’s Philosophy of Science.” *Synthese* 19 (1–2): 3–13.

———. 1977. “Benevolence as an Over-Riding Attitude.” *Australasian Journal of Philosophy* 55 (2): 127–35.

———. 1980. “Utilitarianism and Generalized Benevolence.” *Pacific Philosophical Quarterly* 61 (1–2): 115–21.

Smart, J. J. C., and Bernard Williams. 1973. *Utilitarianism: For and Against*. Cambridge University Press.

Smith, Michael. 1994. *The Moral Problem*. Blackwell.

———. 1998. “Response-Dependence Without Reduction.” *European Review of Philosophy* 3: 85–108.

Smith, Michael, David Lewis, and Mark Johnston. 1989. “Dispositional Theories of Value.” *Proceedings of the Aristotelian Society* 63: 89–174.

Snedegar, Justin. 2017. *Contrastive Reasons*. Oxford University Press.

Sobel, David. 1994. “Full Information Accounts of Well-Being.” *Ethics* 104 (4): 784–810.

Stevenson, Charles Leslie. 1944. *Ethics and Language*. Oxford University Press.

Stocker, Michael. 1976. “The Schizophrenia of Modern Ethical Theories.” *Journal of Philosophy* 73 (14): 453–66.

Street, Sharon. 2006. “A Darwinian Dilemma for Realist Theories of Value.” *Philosophical Studies* 127 (1): 109–66.

———. 2009. “In Defense of Future Tuesday Indifference: Ideally Coherent Eccentrics and the Contingency of What Matters.” *Philosophical Issues* 19 (1): 273–98.

Sumner, L. W. 1996. *Welfare, Happiness, and Ethics*. Oxford University Press.

Svavarsdottir, Sigrun. 1999. “Moral Cognitivism and Motivation.” *Philosophical Review* 108 (2): 161–219.

Tannsjo, Torbjorn. 2002. “Why We Ought to Accept the Repugnant Conclusion.” *Utilitas* 14 (3): 339.

Temkin, Larry S. 2012. *Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning*. Oxford University Press.

Thomson, Judith Jarvis. 2008. *Normativity*. Open Court.

Velleman, J David. 1988. “Brandt’s Definition of ‘Good.’” *Philosophical Review* 97 (3): 353–71.

———. 1999. “A Right of Self Termination?” *Ethics* 109 (3): 606–28.

Wedgwood, Ralph. 1997. "The Essence of Response-Dependence." *European Review of Philosophy* 3: 31–54.

———. 2002. "The Aim of Belief." *Philosophical Perspectives* 36 (s16): 267–97.

———. 2009. "The 'Good' and the 'Right' Revisited." *Philosophical Perspectives* 23 (1): 499–519.

Westermarck, Edward. 1932. *Ethical Relativity*. Greenwood Press.

Wiggins, David. 1987. "A Sensible Subjectivism?" In *Needs, Value and Truth*, 185–214. Blackwell.

Wiland, Eric. 2017. "Moral Testimony: Going on the Offensive." *Oxford Studies in Metaethics* 12: 51–75.

Williams, B. 1970. "Deciding to Believe." In *Problems of the Self*, edited by Bernard Williams, 136–51. Cambridge University Press.

Williams, Bernard. 1979. "Internal and External Reasons." In *Rational Action*, edited by Ross Harrison, 101–13. Cambridge University Press.

Williamson, Timothy. 2000. *Knowledge and Its Limits*. Oxford University Press.
———. 2007. *The Philosophy of Philosophy*. Wiley-Blackwell.

Wright, Crispin. 1988. "The Inaugural Address: Moral Values, Projection and Secondary Qualities." *Aristotelian Society Supplementary Volume* 62: 1–26.

Biography

Michael John Patrick Campbell was born on 14th October 1992 in Isleworth, a suburb of London in the United Kingdom. He received his bachelor's degree in philosophy from Gonville and Caius College, University of Cambridge in July 2014 where he was a scholar. He subsequently received his masters from the same institution in March 2018. He is a Fellow of the Royal Society of Arts.