Prospects for Moral Success Theory

by

Emma Susan Wood

A thesis

submitted to the Victoria University of Wellington

in fulfilment of the requirements for the degree of

Doctor of Philosophy

Victoria University of Wellington

2014

Table of Contents:

Acknowledgments:

It would have been impossible to complete this thesis without the help and encouragement of so many dedicated friends and colleagues. I am very grateful to have been part of the academic community at Victoria University.

To the philosophy department at Victoria University, I give my thanks for providing an environment both challenging and nurturing. The members of department have shown great dedication to all the postgraduate students under their supervision, and I feel very fortunate to have benefitted from this.

Most of all I must thank my supervisor, Richard Joyce, for his extremely valuable input. Richard's positivity, creativity in problem solving, and sharp insight have made him the best mentor I could have hoped for.

Secondly, I thank my associate supervisor, Simon Keller, for his dedication and considerable time spent providing feedback alongside Richard's. I would also like to thank Stuart Brock and Sondra Bacharach for their pastoral care during my studies. Thanks must also go to the department's postgraduate administrator, Philippa Race, for her help in keeping me organised!

Thirdly, I am grateful to the Victoria University Chaplain, John Dennison, for his friendship, support, and wisdom. To all of my friends who share my passion for philosophy: Chiara Ferrario, Stephanie Lentz, Matthew Aroney, Jean-Paul Baumgartner, Sarah-Beth Magambo, Jemimah Roberts, Alexandra Geersen, and Byron Smith – thank you for the conversations!

Fourthly, I am grateful for my family's support through this process. Thank you to my mother, Sue, and my father, Tim, for encouraging my curiosity from a young age and for impressing on me the importance of thought in life. Thank you to Peter and Matthew, my brothers, for the refreshing down-time you've provided during this long process, and to my parents in law Richard and Suzy for your cheerful support.

Finally, a big thank you to my ever-patient husband, Adam. Thank you for your organisational, emotional, and spiritual support. I couldn't have done this without you – this work is as much yours as it is mine.

Abstract:

In this thesis, I will argue that the existence of moral facts does not rely on the existence of a reason for action, and that moral facts can be made sense of in other ways. My thesis is both a reply to a type of moral error theory that has been advanced by Richard Joyce and John Mackie, and an account of the truthmakers of moral judgments.

The argument for error theory that I respond to is roughly as follows: moral judgments are judgments about external practical reasons. But external reasons do not exist, and so no moral judgment is ever true. In the first part of my thesis, I will argue in favour of the latter premise of the error theorist's argument, but against the former: external reasons do not exist, but moral judgments are not committed to them.

In the second half of my thesis I build up a positive account of what moral judgments involve. If moral judgments are not judgments about reasons, then what are moral judgments about? I develop the widely supported idea that moral judgments are judgments that are based on welfarist considerations, and attempt to give this idea a more precise formulation than what has been previously offered. From this account, I go on to develop an account of the truthmakers of moral judgments. The account I end up with is an ideal observer theory that I believe makes sense of a broad range of intuitions about morality.

My hope is that this thesis will be of interest to others who feel the pull of moral error theory, but would prefer to see moral success theory vindicated.

1: Normativity, Morality, and Error Theory

1. Introduction

1.1 The scope of this project

The aim of this thesis is to outline a plausible metaethical framework. The inspiration for this project is a desire to respond to the moral error theory of Richard Joyce and John Mackie, and to offer a version of moral success theory not subject to the same sceptical attacks.[1] Joyce's and Mackie's error theory may be summarised as follows: morality is (or purports to be) normative in a kind of way that turns out to be problematic and which renders all moral judgments false.[2] What I intend to do is present an underexplored (but not all together new) way of understanding the normativity of morality, and, from this, to develop an account of the conditions under which moral judgments are true.

This thesis is an ambitious one given that the four central chapters (2-5) each address a different major question, each of which is pertinent to the aims of this project. Chapter 2 will address the question of which of the two major competing theories of practical rationality (internalism and externalism) is superior. Chapter 3 will contribute significantly

[1] This term 'success theory comes from Sayre McCord, 1986.

[2] The works from which these arguments come are Joyce's 2001 *Myth of Morality* and Mackie's 1977 *Ethics: Inventing Right and Wrong*. However, I will not discuss Mackie's arguments directly, as I take Joyce's presentation of his argument about the difficulty of accounting for the objective prescriptivity or categoricity of moral judgments (1977: 27-35) to be the most formidable presentation of the argument that has been made. I will, therefore restrict my discussion of error theory to Joyce's specific argument.

to the ongoing argument about what is involved in making a genuine moral judgment, and will thereby give an account of what sense in which morality and moral judgments are normative. Chapter 4 will discuss welfare and well-being, and the question of what it is for one to be well-off or for a life to go well. Given that (as I shall be arguing in Chapter 3) a concern with welfare or wellbeing is the defining characteristic of moral judgments, there will be some important metaethical issues at stake in any discussion of what wellbeing consists in. Chapter 5 will give a theory about what moral facts are, or, what the truthmakers of moral judgments are.

Although the four topics addressed in each of these four chapters are quite large topics in and of themselves, I shall connect them to each other in significant ways such that the arguments used in each chapter build on each other. So although this thesis is ambitious, it is not fatally so.

1.2 The need for this project

It is worth saying a little more about the need for projects like this one before proceeding. As I have said, the inspiration for this project came from a desire to respond to the version of moral error theory put forward by Mackie and Joyce. One may initially wonder: why should such a specific line of sceptical attack warrant such a lengthy response?

Firstly, although Mackie and Joyce have given the clearest articulation of the idea that the purported normativity of morality may spell its own undermining, many other philosophers, even if not error theorists themselves, have recognised the threat of error theory. Michael Smith worried that the 'objectivity' and the 'practicality' of morality 'pull in quite opposite directions from each other,' such that 'Nothing could be everything a moral judgment purports to be' (1994: 11). Smith acknowledges that this strangeness of moral judgments provides evidence in favour of a moral error theory,[3] and such a concern motivates him to provide a solution of his own to this 'moral problem'. Christine Korsgaard describes morality's 'normative question', as a question about what sense in which the claims of morality are justified or reason-giving for us. If the claims of morality turn out not to be

[3] In this context Smith uses the less technical term, 'moral nihilism' (ibid) but it amounts to the same thing: the combination of the acceptance of cognitivism: that moral judgments purport to describe something or express beliefs, but that there is nothing for them to describe, nothing to make such beliefs true. This accords with Miller's (2003:5-8) characterisation of metaethical views also.

justified or reason-giving in the way they purport to be, moral scepticism looms (1996: 9-17). And even Immanuel Kant felt the need to show that moral obligations turned out to have the normative force they purported to have, 'if duty is not to be everywhere an empty delusion and a chimerical concept.' (1997 [1785]: 15). Thus this thesis ought not to be seen as a response to the arguments raised by Mackie and Joyce alone, but a contribution to the same questions asked by several philosophers throughout history who have been concerned with the implications of the purported normativity of morality. The reason for a lengthy response to Joyce and Mackie's arguments is because questions about the nature of practical reasons, the nature of moral judgment, and the nature of moral facts must all be dealt with if such a response is to be comprehensive.

1.3 This chapter's discussion

In this first chapter, I will outline some different senses in which moral judgments can be said to be normative. One dominant way of understanding the normativity of moral judgments is as judgments about (or judgments which imply or entail judgments about) one's normative practical reasons. In fact, it is this very understanding of the normativity of moral judgments that gets the error theory of interest going, as we shall soon see. Another way of understanding the normativity of moral judgments is as understanding them as *standard-based* judgments, an account which I shall develop with the help of some vital ideas from David Copp (1995). In introducing this standard-based view of moral judgments, I will make a suggestion that will take the rest of the thesis to argue: that moral success theory is better served by this 'standard-based' understanding of the normativity of morality, rather than the popular 'reasons-based' view.

2. Normativity and Reasons

2.1 Moral judgments: the basics

Before exploring different ways in which moral judgments can be said to be normative, it is worth our briefly outlining the very basics of what moral judgments are. Moral judgments are mental states that can be expressed by certain claims. Examples of such claims include the following: 'one ought to give to charity', 'one ought not steal', 'it is right to keep your promises', 'it is wrong to murder' and so on. Of course, the word 'ought' can have many different meanings. There is a difference between the judgment that one morally ought to do something, and that one prudentially ought to do something, for instance. Similarly, terms like 'right' and 'wrong' stand for a range of different concepts. It is one thing to say that something is *morally* right or wrong, and quite another to say that something is right or wrong according to the rules of a game. So, to be precise, moral judgments are judgments about what *morally* ought to be done, what is *morally* right or wrong. Even though we may not make such a qualification explicit in uttering moral claims, we do, and we are, when we are making moral judgments, thinking about the *moral* rightness or wrongness of the thing we are evaluating: we are making a judgment about what ought to be done in the *moral* sense of the word 'ought'. The discussion of exactly what it is that differentiates moral judgments from other kinds of normative judgments, such as prudential judgments, or judgments or etiquette, or the rules of games, is a question to be dealt with in chapter 3. For now all that needs to be highlighted is the basic point that there are different kinds of normative judgments, and corresponding different usages of thin normative or evaluative terms 'ought', 'right' 'wrong' or 'good' or 'bad'.

Moral judgments are always evaluations of something. Most frequently, perhaps, moral judgments are about actions – the examples of moral claims given in the previous paragraph all express moral judgments about actions. But there are other things that moral judgments can be about besides actions. We often say of people that they are morally good or morally bad. We might say that we morally ought to emulate Jesus's character, and that we morally ought not to emulate Hitler's character. Plausibly, states of affairs can also be morally evaluated. We might say that South Africa under Apartheid was an unjust or morally undesirable state of affairs, for instance. There is a debate to be had about whether moral judgments about persons, character traits, or states of affairs reduce to judgments about actions. Perhaps our judgment that someone is a morally good person is reducible to the judgment that they perform lots of morally good actions, for instance. Perhaps our judgment that a state of affairs is unjust is reducible to the judgment that things are done which ought not to be done, and that widespread moral improvement in behaviour or actions is needed. Though I suspect it is the case that actions are the primary, basic elements of moral judgment, from which judgments about character and states of affairs are derivative, I will not insist on this view, and nothing I say in this thesis depends on such an argument. Though most of the discussion on moral judgments that takes place from this point on will take judgments about *actions* as the paradigm example of moral judgments, the reader should keep in mind that just about everything that is said can apply, mutatis mutandis, to judgments about character and states of affairs too. As a further point, just as moral judgments can be about actions, character traits, or states of affairs, so can non-moral normative judgments like prudential judgments or judgments about etiquette.

For the purposes of this thesis, I will also be assuming that cognitivism about moral judgments is true. Moral judgments are evaluations, or judgments about what one ought to do, and such judgments are truth-apt. Later on, I will give a brief overview of the problems that non-cognitivism faces. Those interested in the merits of non-cognitivism will find my brief treatment disappointing, but I must content myself with the knowledge that it simply lies outside the scope of this thesis to discuss non-cognitivism in detail. This thesis is primarily concerned with the arguments that two kinds of cognitivists – error theorists and success theorists – have with each other.

2.2 Normative judgments: the basics

What are the different senses in which moral judgments may be thought to be normative? We have already seen that normative or evaluative judgments come in many different varieties: moral, prudential, aesthetic, conventional. But there is a further distinction to be made between normative judgments: normative judgments as judgments about practical reasons, and normative judgments as standard-based judgments. The question of interest to this thesis is: which conception characterises *moral* judgments most adequately?

As a preliminary, it is worth a few things about normative judgments in general. A normative judgment is a judgment that can be expressed by a claim which involves certain normative or evaluative concepts, such as ‘ought’, ‘should’ ‘must’ ‘good’ ‘bad’ ‘right’, ‘wrong’ ‘obligatory’, ‘required’ ‘prohibited’ ‘permissible’ ‘admirable’ ‘praiseworthy’ and ‘blameworthy’ (and any other qualification that might serve to specify the type of normative judgment being made: for example, ‘*morally* praiseworthy’, ‘*prudentially* required’). Normative judgments can be expressed in two main forms. Firstly, normative claims can be express propositions about what we ought, should or must do: ‘You ought to give to charity', 'I should watch my step', 'one must not move one's rook diagonally' are all examples of normative claims in what I will call 'imperative' form.[4] The second form normative claims take is as propositions which ascribe evaluative properties such as 'good' or 'bad' to things. 'It was good of you to give to charity', 'It was wrong of me to hurt him', 'It is required of one to drive on the left in Australia' are all examples of normative judgments in what I will call 'property ascription form.'

Many normative concepts are interchangeable with, or imply, each other. For the most part, I believe, normative claims in imperative form are often equivalent to propositions in which an evaluative attribute is attributed. For instance, to say that one ought, in some sense (probably the moral sense), keep one's promises is to say that in that same sense (the moral sense) it is right or good to keep one's promises. To say that it is a (prudentially) bad idea to invest in Microsoft is to say that it one ought not (prudentially) invest in Microsoft. Even when things other than actions are being spoken of, the interchangeability of terms like 'good' and 'ought' can be seen. If I say something like 'It would be good if the world were more peaceful', I can be taken to be saying 'it ought to be the case that the world is more peaceful'.[5]

Normative claims can be distinguished from descriptive claims, which merely describe the world. There is a clear difference between a statement like 'the grass is green', and 'the grass ought to be cut,' or between a statement like 'Cheating on taxes is widespread in Australia' and 'Cheating on taxes is wrong'. The second pair of statements might initially raise a question as to whether there is really a distinction to be made, given that a property is being described. The first statement seems to literally describe something, (the widespreadness of tax cheating in Australia), and so does the other (the wrongness of it). Although both these statements are describing something through the attributing of a property, the fact that they attribute different *kinds* of properties is what makes one statement normative and the other not. The second statement attributes a clearly

[4] Of course, such expressions are not *literally* imperatives, as 'Give to charity,' or 'Watch your step,' are, nor am I making a non-cognitivist claim that such normative judgments are disguised imperatives. 'Imperative form' is simply my preferred terminology for a cognitivist normative judgment about what ought to be done.

[5] One might wonder what to make of this kind of judgment about the way things ought to be. On the one hand, we could interpret this judgment as a judgment that we have reason to collectively make the world more peaceful. On the other hand, I may simply be expressing a desire that the world be more peaceful. Either way, the judgment that it would be good if the world were more peaceful is consistent with the thesis that judgments in imperative form and judgments in property ascription form entail each other.

evaluative property to tax cheating, one that could be interchangeable with or entail a normative judgment in imperative form (that one ought not cheat on their taxes), while the other does not. So although, syntactically, two different statements may both look 'descriptive' given that they *describe* something, this does not mean that they are both descriptive in the non-normative sense of 'descriptive'.

There are some judgments which straddle this divide between descriptive and normative judgments: judgments which involve so-called 'thick' evaluative concepts. Concepts like 'rude', 'courageous', 'disgusting', 'ostentatious', or 'meek' are examples of such concepts. Judgments employing such concepts, such as my judgment that your act was courageous, appear to have both descriptive and normative or evaluative elements. The judgment that your act was courageous involves the descriptive judgment that your action was performed at considerable risk or danger to yourself. But my use of the term 'courageous' also indicates a note of commendation or positive evaluation in my judgment (if I had *not* commended your action, I might have described it as 'foolhardy' instead) (see Burton, 2006 [1992]: 511-2, from which this example comes). On the other hand, it is possible that so called thick evaluative concepts are purely descriptive, and that the element of commendation is an optional addition. Certainly, there are some concepts that seem to be like this. Take the concept of 'altruism'. To judge that an action is altruistic, is, plausibly, a merely descriptive judgment that the action in question is done with the wellbeing of another as the main motivation. And makers of this judgment may believe such actions to be commendable, or they may not (if, for instance, one is particularly cynical, one might believe that altruistic actions are foolish, and therefore not commend them at all). In this thesis, I will be assuming that this is the case for altruism, and for other related concepts often identified as moral virtues, such as 'generosity' and 'compassion'. I will be assuming that these concepts are merely descriptive in and of themselves (meaning that there is no conceptual confusion involved in using the concepts without a note of commendation), but that many users of the concept do happen to believe they are in fact commendable. With regards to concepts that are genuinely thick (concepts that necessarily involve a note of commendation on pain of conceptual confusion), most of my thesis will leave discussions of such topics untouched, given that most of the moral concepts I will be dealing with are thin concepts (such as the concept 'morally good'), or merely descriptive concepts (such as the concept ‘altruism’). However, there is one point at which brief discussion of thick moral concepts will be necessary in chapter three, in relation to questions about the ability of amoralists to make moral judgments.

2.3 Normative judgments as judgments about reasons

Take the claim ‘one ought to ϕ’ as our paradigm formulation of a normative claim directed toward the action of a person. One thing we often do when expressing judgments like this is expressing the judgment that a practical reason in favour of ϕing is present: when one says that one ought to ϕ, one is saying that one has a reason to ϕ. Joyce brings out the intuitive appeal of the claim that normative claims are reasons claims in a passage from *Myth of Morality*:

> Mackie thinks that part of what we *mean* by “ought” is “has a reason”… Mackie’s platitude certainly has appeal. Imagine telling someone: “You really ought to have done that, but I accept that there was no reason for you to.” It sounds very odd, to say the least. Whenever we tell someone that she ought to do something, the question “Why?” is perfectly legitimate. But we want to say something more than “Well, you simply must, and that’s all there is to it!” We will want to provide her with a reason, and if we cannot, then she may feel justified in ignoring our imperative. 2001: 38-9

Joyce intends to show (and the above passage is part of his argument) that moral judgments are, as a matter of conceptual necessity, normative judgments which are judgments about practical reasons (ibid: 30-52). Whether he is right about this will not be settled here. What I simply want to draw attention to is the plausibility of the claim that, when we make normative judgments, these are very often judgments about what we have normative practical reason to do. Of course, not *all* instances of making normative judgments are like this. To borrow an example from Joyce (ibid: 35-6), I may judge that it is wrong according to etiquette (or that I ought not to, according to etiquette) eat with my mouth open at the table. But this may not involve the judgement that I have *practical reason* not to annoy someone – I may not consider the considerations of etiquette to have any reason-giving weight. So although our normative judgments do not *always* involve judgments about what there is reason to do, it is clear that they very often do. This is perhaps clearest when we consider prudential judgments about what we ought to do. When we judge that we really ought to exercise more, or take better care of our health, or study harder lest we fail our exam, or judge that a friendship with this or that person is really worth cultivating in order to contribute to a fulfilling life, these judgments are exactly the kinds of judgments about what we have reason to do: we have reason to look after our health, reason to study harder, reason to cultivate those friendships.

But it is important to be precise about what is meant by the term ‘practical reason’, and thus what a judgment about a practical reason is a judgment about. There are different ways in which the term ‘practical reason’ can be used, which each need untangling. A ‘normative practical reason’ is a type of reason that *counts in favour of* an action. This can be contrasted with an epistemic reason, which counts in favour of a belief.[6] So practical reasons which count in favour of actions are known as normative practical reasons. If I am considering whether I ought to do ϕ, then having a normative practical reason to do ϕ contributes to the justification of my ϕing – this we can call a pro tanto normative practical reason. Let’s say I had a reason to ϕ that ‘trumped’ all other reasons in favour of not ϕing and which ‘decided the case’ so to speak, in favour of ϕing. We can refer to a reason like this as an overriding normative practical reason. Suppose no one reason in favour of ϕing was decisive on its own, but suppose I had enough pro tanto reasons, taken together, to justify my ϕing. In this instance, I have *all-things-considered* reason to ϕ. Pro tanto and overriding normative reasons differ in strength, we might say, but there is a distinct sense in which they are similar concepts. They are both reasons which rationally justify (or contribute toward rationally justifying) actions.

But practical reasons are often spoken of in another way, as things that motivate an individual, or explain their action. When we say things like ‘The reason he went to the party was because Kate was there’, we are speaking of a ‘reason’ in this motivating sense. Accordingly, we call these reasons ‘motivating reasons’. Roughly, normative reasons are reasons that indicate what we should do or ought to do (at least to some degree, even if the reason is pro tanto and is overridden by a conflicting reason)[7] while motivating reasons are

[6] One might want to argue that there is not a clear distinction between epistemic and practical reasons, and that epistemic reasons are a type of practical reason. Whether this is the case or not is not a debate I have the space to go into here. Even if it is true, however, that epistemic reasons are a type of practical reason, most of the reasons I will discuss in this thesis are non-epistemic practical reasons, so nothing much hangs on such a debate.

[7] This reveals how I shall be understanding the relationship between judgments between what we ought to do and judgments about what we have reason to do. To say that one has a normative practical reason to ϕ is always to imply that we *to some extent* ought to something. Some might be hesitant, and may want to claim only that judgments about what we have *overriding* or *all-things-considered* reason to do are equivalent to

reasons we are motivated to act *for*. This distinction between normative and motivating reasons is a distinction that is widely drawn (Parfit 2010: 37, Darwall 1983: 29-33, Nagel 1970: 15, Smith 1994: 14).

Our motivating reasons and our normative reasons often coincide: it is often the case that we are motivated to do what we also have normative reason to do. But it should also be obvious that they often do not – the fact that I am motivated to do something does not necessarily mean in any sense that I *should*.[8] Say I am at a rowdy pub and somebody hurls an insult at me. Filled with anger, and the conviction that I have a reason to defend my honour, I prepare to go and punch the person who insulted me in the face. Despite the fact that I am convinced that I have a reason to punch this person, I think we can easily see that the facts about what I have reason to do may well diverge from my opinions on the matter. Suppose, for instance, that the person was in fact not yelling an insult at me, but rather, was affectionately having a dig at one of his friends who happened to be standing beyond me. Suppose also that, if I were to punch the person in the face, far from defending my honour, it would only make me look stupid and uncivilised in the eyes of everyone else there. In other words, the reason that I think I have, the reason I am thinking of acting *for*, is no reason at all, in the normative sense.

So far I have discussed what normative practical reasons are, and what motivating practical reasons are. But there is another, more colloquial sense in which the term 'reason' is used. A 'reason' is what we often call anything that follows in a statement that begins with words like 'I did that because…' or, 'the reason you should do this is that…'. Take the statement 'the reason one ought to give to charity is that it helps people. Colloquially, 'it helps people' might be cited as a reason for giving to charity – it is the proposition that follows a 'because'

judgments about what we ought to do. Though I agree that only overriding or all-things-considered reasons can fully justify or conclude our rational deliberation about what we ought to do, I see no need to deny that pro tanto reasons indicate (even to a small extent) what we ought to do, even if they don't indicate decisively what we ought to do. Each pro tanto reason, it might be said, adds more 'ought-ness' to the action it counts in favour of, until we have enough pro-tanto reasons to conclude that we have, all things considered, a case for the action. So I like to say that judgments about both pro tanto and overriding reasons entail judgments about what we ought to do. The difference between pro-tanto reasons judgments and overriding reasons judgments is not that the former does not entail judgments about what we ought to do, while the latter does, but rather, that they both entail judgments about what we ought to do which differ only in strength or degree.

[8] Of course, Bernard Williams' (1979) internalism holds that there is a very close relationship between our motives and practical reasons. But, as we shall see in chapter 2, this relationship is not so close so as to collapse the distinction between normative and practical reasons.

clause or a 'that' clause. But it is important to understand that this colloquial use of the term 'a reason' is *not* the same thing as a normative reason in the sense we have been discussing. 'It helps people' expresses what might be called a 'consideration' – a fact about the action that can be taken into account when considering whether or not to perform the action. But considerations are not the same things as normative practical reasons, because it is an open question as to whether any considerations actually *count in favour* of actions in any given instance. If the consideration 'that it helps people' does in fact constitute a normative practical reason to give to charity, this will be because it meets certain conditions. Different theories of practical reasons give different accounts of what any given consideration must meet in order to count as a reason (we will get to these theories shortly). But 'it helps people' does not count as a normative practical reason for giving to charity simply because someone can utter the sentence 'the reason one ought to give to charity is because it helps people'. For people can make false statements about what we have reason to do: some considerations may actually count *against* actions, rather than in favour of them. Suppose someone says to me: 'you ought to come to the pub tonight because there will be country music playing.' If I loathe country music, the consideration that there will be country music playing plausibly counts *against* the action that it is intended by the speaker to count in favour of. So, even though the term 'reason' is often used to designate any fact about a given action ('that it helps people' being a fact about giving to charity, or 'that there will be country music playing' being a fact about the party) that a speaker might draw attention to in order to support a claim that there is reason to perform the given action, we must not confuse this colloquial use of the term 'reason' with the concept of a normative practical reason. Facts about actions, as I have said, can be termed 'considerations' to avoid confusion, and each consideration potentially counts as a normative practical reason, and potentially does not.[9] It should be noted also that even

[9] Someone else who has identified this difference between considerations and reasons is Simon Keller, when he writes 'A reason to perform an act is a consideration that counts in favour of the act, in one way or another. A reason to act, that is to say, is a consideration that *really* counts in favour of the act, not one that merely seems to count in its favour or that someone merely thinks to count in its favour. (On this way of talking, all reasons are good reasons; a "bad reason" is no reason at all.)' 2013:3. Considerations, as I have been using the term, can be cited as either 'good reasons' or 'bad reasons'. But, following Keller, it is more technically correct to say that the considerations that are correctly said to be 'good reasons' are the ones that constitute normative practical reasons (and which meet whatever conditions are demanded by the correct theory of practical reason), while considerations that are correctly said to be 'bad reasons' fail to constitute normative practical reasons.

motivating reasons are distinct from considerations. For instance, I could report to someone what I was told (perhaps insincerely, or in a sarcastic tone): ‘The reason I have to go to the party tonight is because there will be country music playing’. But the fact that I referred to the fact that there will be country music playing as a ‘reason’ does not indicate that it is a motivating reason for me (since I hate country music) any more than it indicates that it is a normative practical reason.[10] Thus, we must keep three ways in which the term ‘reason’ can be used distinct: there are normative practical reasons, which count in favour of actions, motivating practical reasons, which are the reasons we act for, and considerations, which are facts about actions which may or may not constitute normative or motivating reasons in any given instance.

Now that we have these distinctions in mind, it will be helpful to briefly bring these distinctions back to the point at hand. When we make normative or evaluative judgments, what we are often doing is making judgments about *normative practical reasons*, (rather than motivating reasons), which we believe are present. Suppose I say to you, ‘you ought to save your money rather than spending it on the pokies’ (or, alternatively, if we want to put the expressed judgment in property ascription form: ‘it would be good if you saved your money rather than spending it on the pokies’). When I make a statement like this, I am not claiming that you are in fact motivated because *you* think you have a reason. I am saying that you have a *normative* practical reason to save your money and not spend it on the pokies, whether you agree or not. So, to be clear, when I said at the outset of this section that some of our normative judgments equate to judgments about reasons, I had in mind *normative practical reasons*.

There is just one last feature of this kind of normative judgment that I would like to draw attention to before I go on to outline the differences between two major theories of practical reason. The very terminology of normative practical reasons (reasons in favour of actions) and motivating practical reasons (considerations that we think are reasons, reasons we act *for*) presupposes that there is an important relationship between judgments one makes about one’s own normative practical reasons, and one’s motivations: namely, that *if*

[10] Darwall, 1983: 31 conflates motivating reasons and considerations, what he calls ‘*dicta*’. For the reasons I have given, however, I think we can see that it makes more sense to see motivating reasons and considerations as separate concepts. ‘Dicta’ or considerations *can* constitute someone’s motivating reasons, but also may not.

one judges that one has a normative practical reason to ϕ, one will be somewhat motivated to ϕ. Such a motivation may be defeasible, and may not be overriding (other motives may conflict with it, and weakness of will might interfere), but some degree of motivation will be present. Call this the 'motivational link' of judgments about reasons. Note that this doctrine should *not* be confused with other well-known metaethical doctrines. It should not, for instance, be confused with 'the internalist requirement' (Smith, 1995: 111-112 Korsgaard, 1986: 11), which claims that reasons must be capable of motivating persons insofar as they are rational. Nor should this claim be confused with the doctrine of moral motivational internalism (Smith: 1994: 12, 61-92 Brink, 1989: 45-50, Shafer Landau 2003: 142-3, Schroeder 2010: 9-10) which is the doctrine that one's genuine moral judgments will somewhat motivate one. The discussion of both of these doctrines will take place later on. For now, all we are assuming is that making a judgment about one's normative practical reasons will motivate one somewhat.

This doctrine that I call the 'motivational link' is, from what I can see, fairly widely assumed. Indeed, as I have said, the very terminology of 'normative practical reasons' and 'motivating reasons' assumes its truth. Furthermore, the 'motivational link' is used as a premise to support other internalist doctrines like motivational internalism about moral judgments: moral judgments, it is often assumed, are necessarily motivating, and the best explanation of this fact is that moral judgments constitute or entail judgments about our normative practical reasons, which, of course, motivate (Smith 1994: 10-14, 184-5). Such well known arguments gain whatever plausibility they have on the widely shared assumption that first personal judgments about one's normative reasons motivate. But despite the plausibility of this doctrine, one may initially want to doubt it: scenarios like the following come to mind. I am watching a TV show I like and I realise my building is burning down. I may, an objection might go, judge that I have a tiny reason to stay and watch my TV show without experiencing any motivation, given the danger I am in, to flee from the building. If this is the case, then there may not be a conceptual connection between making judgments about reasons, and being motivated, after all. But I don't think this objection (or others like it) succeeds. It does not seem unreasonable to claim that I would have a tiny motivation (tiny as the reason I judge myself to have) to stay and watch my TV show, even if such a motivation was evinced merely by a thought like the following: 'damn! My building is burning down, and what makes it even worse is the fact that I won't get to finish my TV show!' If one rejects this way of answering the objection and thinks it just too strange that I would have even a tiny motivation to watch my TV show, then it is plausible to conclude that it is equally strange that I would make the judgment (at least in that moment) that I have a reason to watch my TV show. For surely the reason one would think I have zero motivation to watch my TV show was because the danger to me presented by the fire was so great that my mind could not be occupied by anything but the motivation to flee. But in this case, neither could my mind be occupied by the judgment that I have reason to stay and watch my TV show. Either way we answer the objection, I believe the thesis stays intact: that if one judges one has a reason, one will be motivated accordingly.[11]

To summarise this section: often, when we make normative or evaluative judgments, these consist of or imply judgments about normative practical reasons. These judgments, if we make them first personally (in other words, with reference to what *we* ought to do or have reason to do), they necessarily motivate us, even if the motivation is defeasible. It remains up for debate whether *moral* judgments, as many have claimed, are normative judgments of the reasons-implying kind. Before we even consider this question, we must first introduce another important distinction between normative practical reasons themselves: internal and external reasons.

2.4 Theories of practical rationality

We know that normative reasons are reasons that count in favour of actions. But in virtue of what does a given reason count in favour of an action? What conditions does a consideration have to meet in order to constitute a normative practical reason? There are, broadly speaking, two different theories which give different answers to these questions: reasons internalism, and reasons externalism.

The distinction between internal and external reasons was developed by Bernard Williams in his famous 1979 paper. Not only did Williams establish the distinction, but he took a firm stand in favour of the thesis that only internal practical reasons exist, a thesis for which I will reserve the label 'reasons internalism'. Briefly, reasons internalism is the view that, in order

[11] I discuss this example and objection in Wood (forthcoming)

to have a reason for ϕing, it must be the case that ϕing satisfies what Williams referred to as our *subjective motivational set*. The subjective motivational set, for Williams, includes everything from one's desires, one's values, one's patterns of emotional reaction, or loyalties, plans and goals (1979: 20). So one has reason to ϕ, according to Williams's view, if ϕing would fulfil a desire, be in accordance with one's values, or would generate desired emotional reactions, or would contribute to one's various goals or commitments. There are theories of rationality that are slightly narrower than reasons internalism, which make only one such element of the motivational set the determining factor in what we have reason to do. The is the Humean theory of rationality, which says that reasons are constituted only by desire satisfactions (a view eloquently defended by Schroeder, 2007), and the instrumental theory of rationality, which says that we have reason to pursue whatever ends or goals we have (see for instance Joyce 2001: 53-79, Harman 1975). For the most part of this thesis, I will not discuss such theories, as I take reasons internalism to be a more plausible theory than both of them due to its breadth. The focus on just one element of the subjective motivational set is something I regard as largely unnecessary, though there is not room in this thesis to defend this claim. Given that I myself will argue in favour of reasons internalism (in chapter 2), most of the discussion of theories of rationality will focus on reasons internalism, and its traditional rival, reasons externalism.

If reasons internalism is the view that, in order to be reason-giving, that action must satisfy an element in one's subjective motivational set, then reasons externalism is a denial of that thesis. Reasons externalism holds that there are substantive rational requirements on us that are independent of whether acting in accordance with them satisfies any element of our motivational set (Parfit, 1997: 100-101). Suppose you believe I have an external reason to be kind to my siblings. What this would mean is that you believe that *being kind to my siblings* is in and of itself a normative reason, regardless of whether being kind to my siblings satisfies an element of my motivational set. I have deliberately chosen a moral-sounding normative judgment as a candidate example of a judgment about external reasons. For it leads me to a vital premise in Richard Joyce's argument for moral error theory that I will look at in section 4: that moral judgments, properly understood, are judgments which consist in or imply external reasons.[12]

Before I move onto the discussion of a second kind of normative statement, there are two points of clarification worth a mention. I have outlined the difference between reasons internalism and reasons externalism very roughly here, but a far more detailed exegesis of Williams' 1979 paper, and a discussion of the ways in which the two theories of practical rationality have been (mis)understood, will take up much of chapter 2. My rough outline of reasons internalism and externalism is adequate for my purposes in this chapter, but a full discussion of the merits of each theory will require much more detail, so more will be said in chapter 2.

The second point worth making is that this internalism/externalism distinction I have discussed is another one of many internalism/externalism distinctions in metaethics which, to avoid confusion, must be kept distinct. The internalist/externalist labels have been applied to doctrines about the link between the making of moral judgments, and motivation. They have also been applied to doctrines about the presence of moral obligations and normative practical reasons (Darwall, 1983), and, again, have been applied to the difference between the two theories of practical rationality. So that confusion is avoided from the outset, I will give below a list of doctrines that I will have need to discuss throughout this thesis, and give in bold the labels that I am attaching to each doctrine. Whenever I use the following labels throughout this thesis, the reader should refer back to this list to avoid any confusion over what is being discussed:

> **The motivational link:** If one makes a judgment that one has a normative practical reason to ϕ, one will be somewhat motivated to ϕ.
>
> **The internalist requirement:** Normative practical reasons must be capable of motivating one insofar as one is rational.
>
> **Reasons internalism:** All normative practical reasons are internal reasons.[13]

[12] In Joyce's main argument for error theory (2001: 30-53) the internal/external distinction is not the terminology he uses. But, as we shall see, (and as he has frequently confirmed in conversation with me) the idea that moral judgments imply judgments about external reasons is a fair interpretation of his argument, his terminological differences notwithstanding.

> **Reasons externalism:** Normative practical reasons can include both internal or external reasons.[14]
>
> **Moral motivational internalism:** If one judges that one morally ought to ϕ, then one will be somewhat motivated to ϕ.
>
> **Moral motivational externalism:** If one judges that one morally ought to ϕ, one need not have any accompanying motivation to ϕ.
>
> **The rationalist conceptual claim:** If one judges that one morally ought to ϕ, then one will be judging (as a matter of conceptual necessity) that one has a reason to ϕ.

3. Normativity and Standards

3.1 Standard-based judgments

There is a second way normative judgments can be understood. They can be understood as containing propositions about systems of *standards*. My definition of a standard will follow Copp's (1995: 19):

> **Standard:** Something to which actions or behaviour can conform or fail to conform to, or according to which be classified as 'correct' or 'incorrect'.

Let's apply this definition to a concrete example. Take the norm 'one ought not steal', and understand this claim as the expression of a standard. Take two different actions: Sam takes money out of his friend Sarah's wallet without her knowing, and fails to tell her. His

[13] Some have discussed the internalist requirement as if it were essentially equivalent to reasons internalism (Smith 1995). But, as we shall see in chapter 2, this is not the case. Though the internalist requirement may be used as a premise in the argument for reasons internalism, we shall see that it is perfectly compatible with either reasons internalism or reasons externalism as understood by Williams. The true difference between reasons internalism and reasons externalism amounts to much more than an acceptance or denial of the internalist requirement. Thus the internalist requirement, and reasons internalism, are quite different doctrines.

[14] Needless to say, it is possible to hold the view that all normative reasons are external and that none are internal, but I regard this view as so implausible such that framing externalism as the inclusive thesis that there are both internal *and* external reasons is far more charitable. Even if there are such things as external reasons, it is unlikely that *all* such reasons are external. Take my reason to listen to Charlie Parker rather than Dizzy Gillespie, for instance. Supposedly what gives me this reason is my *desire* to listen to Charlie Parker over Dizzy Gillespie. Among the reasons there are, such reasons of taste, at least, are surely internal.

action, according to the above standard, is ‘incorrect’. Take another example, an expression of a standard which evaluates behaviour over a longer period of time: ‘If one’s income is over the equivalent purchasing power of 50 K Australian, one ought to give ten percent of one’s money to reputable charity organisations each year.’ Take two patterns of behaviour of two Australians. John earns 70 K each year, and gives away 2 K to charity. Linda earns 70 K each year, and gives away 7 K to charity. According to the above standard, John’s pattern of behaviour is incorrect, and Linda’s pattern of behaviour is correct.

Another feature about standards that Copp identifies is that standards themselves cannot be true and false, and are not, strictly speaking, propositions (ibid: 19-20). Of course, there are truth apt propositions *about* standards, of which the following is an example: ‘the prohibition against stealing is a standard’. This sentence simply states that something is a standard, namely, the prohibition against stealing. Another example of a proposition about a standard is in the following sentence: ‘your performance does not meet our company’s standards’. This proposition is about a particular action’s relationship (in this case, lack of conformity) to a particular standard. Take another example: ‘If you want to win gold in high jump at the Olympics, your ability to jump will have to be at a very high standard’. In each of these sentences, the standards in question are not truth-apt, but there are nevertheless truth-apt propositions that can be made about them.

There are many different *kinds* of standards that we recognise. We recognise moral standards, aesthetic standards, standards of etiquette, the standards set in the form of questions on an exam, and the rules of games. An example of a moral standard is the prohibition against stealing: there are different potential actions that either conform or fail to conform to this standard. A standard of etiquette might be the expectation that you shake the hand of the person conducting your job interview when you meet them. Your shaking their hand will thus conform to the standard, your failure to extend your hand and shake will fail to conform. A standard of fashion or aesthetics might be that one ought not to wear two different garments with clashing patterns. My wearing a plain white top with my floral skirt will conform to this standard, but my husband’s habit of wearing his stripy polo shirt with checked shorts will not. In my maths exam, any given question can create a standard. The question: ‘what is the square root of 25?’ creates a standard in that the response ‘5’ will constitute conformity, but any other answer besides 5 will constitute lack of conformity. If I am playing chess, there are certain moves I can make that will either conform or fail to conform to the standards of chess which constitute its rules. My moving my rook in any straight line will conform, but my moving my bishop in a straight line will not.

So standards come in many types. There exist what might be called 'systems of standards'. Morality constitutes one system of standards. Different games constitute different systems of standards. As do fashion, etiquette, and prudence. The question now arises: what makes a cluster of standards a 'system' (as opposed to an aggregation picked out at random)? What do all moral standards have in common by virtue of which they are *moral* standards, and what do all aesthetic standards have in common by virtue of which they are *aesthetic* standards? What makes a given standard a standard of a certain kind, rather than another kind? What, for instance, makes the standard, 'you ought not steal?' a standard of morality rather than a standard of etiquette? This question will be given fuller treatment in chapter 3, but the main idea can be outlined simply enough now: different systems of standards are 'about' different things, and apply to different kinds of actions or situations. If systems of standards could be personified, we would say they have different 'concerns'. Standards of etiquette concern beauty: the way things look, sound, fit together, the sensory impression they give. Standards of prudence are concerned with and evaluate actions insofar as they affect one's self-interest. The question of what standards of morality involve is somewhat controversial, but a doctrine I will be assuming (and to some degree defending, in chapter 3) is *welfarism*: the idea that what makes a given norm or standard a *moral* standard is that it is centrally concerned with welfare, or with regulating the interactions between interest-bearers with a view to promoting or respecting welfare.[15]

To summarise this section, the second kind of standards-based judgment is simply a judgment that something is required or forbidden by a certain system of norms ('chewing with one's mouth closed is required according to western etiquette') or is a judgment that employs a given system of norms to evaluate an action (morally, stealing is wrong). Accordingly, one might ask a particular question at this point, as to whether judgments like these truly deserve to be called *normative*. The following words from Alan Gibbard come to mind:

[15] For a discussion of welfarism and the various ways it has been discussed, see Keller 2009

> We can characterise any system N of norms by a family of basic predicates “N-forbidden”, “N-optimal”, and “N-required”. Here “N-forbidden” simply means “forbidden by system of norms N”, and likewise for its siblings… These predicates are descriptive rather than normative: whether a thing say, is N-permitted will be a matter of fact. It might be N-permitted without being rational, for the system N might have little to recommend it. People who agree on the facts will agree on what is N-permitted and what is not, even if they disagree normatively – even if, for instance, one accepts N and the other does not. 1990: 87

Should standards-based judgments be better described as descriptive, rather than normative judgments? The reason I prefer to maintain that they are a type of normative judgment is quite simple, namely that my view about them differs slightly to Gibbard’s. What Gibbard wants to emphasise here is that an action can be judged to be ‘N-required’ or ‘N-forbidden without this either entailing a judgment that the action is rational, or as indicating that the speaker endorses the system N. What I want to emphasise, rather, is that an action’s being judged to be ‘N-required’ or ‘N-forbidden’ can leave it *open* as to whether the speaker deems the requirement rational and thus endorses the system N. It is possible to judge an action to be N-required and believe that, in virtue of this, that there is a reason to perform the action, and possible to judge an action to be N-required and not make the latter judgment. In light of this, I consider it simpler to refer to standards-based judgments as a type of normative judgment rather than a descriptive judgment.

But there is a second reason I prefer to refer to standards-based judgments as normative, rather than descriptive, and that is that is that such standards-based judgments have normative import for many people who endorse the given system of standards in question. As I will argue later on, standards-based judgments about morality have fundamentally to do with considerations about welfare. Now suppose we have two agents arguing about the moral rightness of a given government policy. The first agent believes that the government’s policy is morally right, or required by morality, and the second agent believes that the government’s policy is forbidden by morality. Suppose, further, that both agents in the scenario care about morality. Whatever morality recommends *de dicto*, they endorse. It seems jarring to me to describe the disagreement here as one over a mere descriptive detail, where the two agents normatively agree (given their endorsement of a system N), as Gibbard would seem to imply. Their disagreement seems to be very much a *normative* one, even though it is a disagreement over what is N-required, because the two disputants take the question over what is N-required, in this case, to have normative import (given that they take moral requirements to be reason giving, or, they ‘endorse’ morality). To summarise: I prefer to consider standards-based judgments as a type of normative judgment, given that they have the potential to have normative import for people (even though they do not, as a matter of conceptual necessity, *have to*). Ultimately, if the reader prefers to think of standards-based judgments as descriptive, I will not insist that they think otherwise (for I do not think this affects the plausibility of later arguments). But for the reasons I have just given, I myself prefer to think of them as a type of normative judgment.

3.2 Morality: standards or reasons?

It should be quite clear that all normative judgments express standards. The question at hand, then, is not the question of whether moral judgments contain beliefs about standards or not, but whether those standards also give *reasons*. It should be obvious already that a normative judgment’s representing a standard does not entail that a judgment about a reason is present. For instance, I can judge that etiquette demands that I shake a new acquaintance’s hand. In other words, I am judging that shaking another’s hand is ‘correct’ as according to etiquette. But my appreciation of this correctness according to etiquette does not necessarily entail that I judge that I have a normative practical reason to shake hands: for I may think (and I may be correct in thinking) that the standards of etiquette are pointless and serve no helpful purpose for me or anybody else. If I am correct, then this is one instance in which something’s being a standard and something’s being a reason do not coincide.

But morality, it has been thought, is importantly different to etiquette and other norms in this respect. If a normative judgment expresses a *moral* standard, then, as a matter of conceptual necessity, a judgment about a reason must be present. To use some terminology from Joyce, norms that consist merely of ‘standards’ can be termed ‘weak categorical imperatives’: though they, (as Foot, 1972: 308-9, also observed) may apply to agents unconditionally in that they give standards which can be conformed to or not conformed to, (regardless of the agent’s interests) they may nevertheless fail to have *reason-giving force*. But moral judgments, Joyce contests, consist of ‘strong categorical imperatives’. Not only are moral judgments expressions of standards which apply unconditionally, but such standards are *reason-bringing*. To do something that is 'incorrect' according to a moral standard is to do something that we have reason to not do, and to do something that is 'correct' according to a moral standard is to do something that we have reason to do. At least, this is, what Joyce thinks, what moral judgments presuppose. He makes the point about the difference between 'lesser' norms and the purported norms of morality in the following passages:

> Consider Celadus the Thracian, an unwilling gladiator: he's dragged off the street, buckled into armor, and thrust into the arena... let's imagine that there are various rules of gladiatorial combat: you ought not throw sand in your opponent's eyes, for instance. Celadus is a gladiator, subject to the rules, and so he ought not throw sand in his opponent's eyes. But these are not *his* rules – what are they to him? Imagine that things are looking bleak – his opponent is a sadistic professional fighter, and Celadus finds himself pinned down and swordless. His only hope is to throw some sand in his rival's eyes… The rules still say that Celadus shouldn't do it, but he doesn't care about the rules – he has no particular reason to follow them, and every reason to reject them…. I think we will all agree that Celadus ought to throw sand in his opponent's eyes. He ought to do what he ought not to do.
>
> This last statement is not in the least paradoxical, so long as we keep track of our "ought"s. The first "ought" is something like an all-things-considered "ought" – it presents the *thing to be done*. The second "ought" is an according-to-the-rules "ought". Saying that Celadus ought to do what he ought not to do is just a way of saying that he ought to break the rules. The fact that he should break them doesn't make them evaporate – they continue to *be* rules – and so there continues to be a sense in which he ought not throw sand in his opponent's eyes. If we denied this we could make no sense of the fact that Celadus was *breaking a rule* at all…
>
> When I said earlier that morality is not merely a set of *Dos* and *Don'ts* that we are willing to back up with force, I meant that morality is something more than a set of weak categorical imperatives. Reflect on how differently we treat "Celadus ought not throw sand" and "Gyges ought not kill people." We're content to admit that the former case is just a matter of there being a set of rules which someone from the outside is imposing on the gladiator, and these rules can be overridden by the gladiator's personal desires and interests. They need not *bind* him; they need not be *his* rules; they do not present *the thing to do*; he may legitimately ignore them. But imagine saying something analogoue of Plato's shepherd: "Of course by killing an innocent person the shepherd is breaking a rule of morality, and so *according-to-the-rules-of-morality* he ought not do it; nevertheless, if he stands to gain something important by killing, then that's what he ought (all things considered) to do." That's *not* how we think of morality. Someone who reasoned in such a way might be accused of fundamentally misunderstanding what we mean by "morally ought".

> The pressing question, then, is what "extra ingredient" a *strong* categorical imperative has. If the prohibitions against killing innocents are intended to have much more force than the prohibitions against sand-throwing, or speaking with one's mouth full, what is the source and nature of the force?
>
> Foot interprets Kant as holding that moral imperatives do not merely "apply" to persons regardless of their desires or interests, but imply that persons also have *reasons* to act regardless of their desires or interests – and in this way they (putatively) *bind* persons, in a way that etiquette does not bind us. But she doesn't think that such reason-bringing imperatives are philosophically defensible, and so she accuses Kant of attempting to imbue morality with a "magic force." The failure of strong categorical imperatives, however, does not lead Foot in the direction of a moral error theory, since she thinks that the belief in their validity is a mistake that only Kant and like thinkers have made – ordinary users of moral concepts do not generally make this blunder. My contention is that Foot is correct about the "magical force" of strong categorical imperatives, but wrong in thinking that they are expendable to morality. 2001: 34-7

This question about whether moral judgments imply judgments about reasons (a thesis that I will refer to as the 'rationalist conceptual claim') has divided many philosophers. On Joyce's side are Michael Smith (1994), John Mackie (1977), Terrence Cuneo (2007), Alan Gewirth (1978), and Thomas Nagel (1970). On Foot's (1972) side are David Brink (1989) Peter Railton (1986; 1992), David Copp (1995), and Russ Shafer-Landau (2005).[16] In this thesis, I shall be taking the side of the deniers of the rationalist conceptual claim. I will suggest, contra Joyce, that morality *is* like other normative judgments in the sense that,

[16] Shafer-Landau's view is admittedly a little harder to categorise. In 2005: 109-112, he appears to deny the rationalist conceptual claim as defended by Joyce, by allowing for the conceptual possibility that one may have reason to act against one's moral duty. In his 'Moral Realism: A Defence', however, one could be forgiven for thinking that he subscribes to the rationalist conceptual claim 'Imagine someone doing an act because she thinks it is right – she acts from the motive of duty, and, let us suppose, in this case she is on target about what duty requires… If she correctly cites an action's rightness as her reason for performing it, we don't ordinarily question the legitimacy or conceptual coherence of her doing so. But if the rightness of an act itself was no reason at all for performing it, then we would have to do just that.' 2003: 192. But a careful reading of Shafer-Landau (ibid) suggests that he is claiming only that the rightness of an act *gives reason* to an agent for performing an act, not that one's *judgment* about the rightness of an act need imply a *judgment* that the act gives on a reason. For on the previous page he says: 'moral rationalism [the view that the rightness of an act gives reason to perform it] is compatible either with moral realism, or with any number of antirealist doctrines… The moral of this familiar story is that realism can easily resist one arm of the queerness objection issued by Mackie… Mackie alleges, first, that realists are committed to the view that moral facts entail reasons for action… On this line, the falsity of moral rationalism entails the falsity of realism. That argument is flawed because, as I have said, realism is by itself free of any commitment to moral rationalism. So even if rationalism is false, realism would emerge unscathed.' (ibid: 191) In saying that realism is compatible with a rejection of moral rationalism, Shafer-Landau implies the following: that I can consistently, without conceptual confusion, make a judgment (as a moral realist) that ɸing is right, without thereby having to assert that I have a reason to ɸ. Accordingly, Shafer-Landau is best placed as a denier of the rationalist conceptual claim, given that the rationalist conceptual claim is a claim about the relationship between moral *judgments* and judgments about reasons, rather than a claim about the relationship between moral *facts* and the presence of reasons.

while standards may be referred to by moral judgments, judgments about reasons need not be entailed.

Of course, this does not mean that there is *nothing* special or different about the norms of morality to other norms. That moral norms are typically imbued with a significance that other kinds of norms lack is hard to deny. All I am claiming is that this special ‘difference’ between moral norms and other kinds of norms need not be understood in the way that supporters of the rationalist conceptual claim demand. There is indeed a particular relationship between morality and practical reason that sets it apart from other kinds of norms, and just what this relationship is will become clear as the arguments in this thesis progress. But I will deny that moral judgments must necessarily be understood as judgments implying ones about practical reasons.

3.3 Moral standards and truth

But now we come to an important question. If moral claims are expressions of standards, and standards do not have truth value, then doesn’t this mean that moral claims do not have truth value? If the answer to this question is ‘yes’, then I would regard this as a serious drawback for my view, for all the familiar reasons that confront well known non-cognitivist views in metaethics.

When we engage each other in moral discussions, we often disagree about what is right or wrong. What we take ourselves to be doing when we have moral disagreements with opponents, is believing that we are right about something and that our opponents are wrong – that we know the facts of the matter and that they do not. At least, this is how things appear to us. Non-cognitivism cannot make sense of such appearances without some hard work.

The Frege-Geach problem also presents a challenge to the notion that moral claims are not truth apt, because moral claims can be imbedded in arguments that we take to be logically valid, such as:

1. Stealing is wrong
2. If stealing is wrong, then it is wrong to get your little brother to steal.
3. Therefore it is wrong to get your little brother to steal.

We take such arguments to be examples of arguments that are logically valid. But, they cannot be logically valid if the moral claims made are not truth-apt, as logical validity is a matter of preserving truth across premises to a conclusion. So here is another challenge non-cognitivism faces: it cannot at first glance seem to explain how moral arguments such as the above are logically valid. Non-cognitivists have responded to this challenge and have come up with creative solutions to this problem (see Blackburn 2006a, 2006b). But it is, arguably, an uphill battle, and the fact that cognitivist moral theories don't have such difficult explaining to do, and that they take moral language to be as it appears to be gives cognitivism a simple appeal over non-cognitivism. If moral propositions are expressions of standards, then, does this mean my view will face all the challenges that non-cognitivism faces?[17]

In short, my answer is 'no'. Although standards are not truth apt, my view about moral judgments is a cognitivist one (or at the very least, a partly cognitivist one), and I will explain why now. Take a typical moral judgment expressed by the following claim: 'one morally ought to keep one's promises.' What is meant by 'morally ought' in such a judgment, I contend, is 'according to morality'. Our moral judgment, then, is a truth-apt proposition that can be unpacked as follows: the requirement of keeping one's promises is a requirement that belongs to the system of norms, 'morality'. When one makes a moral judgment about what is or is not a moral requirement, what one is doing according to my standards-based view is judging that the act being morally evaluated conforms or fails to conform to a standard that is truly a *moral* standard.

What, then, makes such judgments *true*? A plausible hypothesis that we can go with is that, for each system of standards, there are particular truth conditions for any given judgments about those standards. I will give some examples, to make this idea clearer. Take etiquette, for example. (Or to be more precise, take Western etiquette, given that 'etiquette' describes not one but many different normative systems that vary from culture to culture). Based on what the supposed purpose of etiquette as a system of rules is, we can get a fairly good idea of what the truthmakers of etiquette judgments are. We know that etiquette, as a concept, is a system of norms concerned with governing social behaviour. In some situations, the rules of etiquette function to put people at ease with each other: the rule of

[17] For a summary of the problems that non-cognitivism faces, see Schroeder, 2010.

not chewing with one’s mouth open at the dinner table, or shaking the hand of a new acquaintance, are examples of this. On other occasions, the rules of etiquette serve to generate a certain emotion perceived to be fitting for the situation. For example, before the bride enters the church at a wedding, etiquette demands that all guests stand, and the action of standing supposedly signifies and reinforces the significance of the event, or creates heightened expectation to add more gravitas to the occasion. And at funerals, etiquette deems that black should be worn – again to create a solemn mood fitting for the event. So the purpose of etiquette could be said to be to generate certain moods on certain occasions or ease social relations. Now, take the judgment that one ought, according to etiquette, refrain from eating with one’s mouth open at a dinner table with new acquaintances. What makes this judgment true (if it is indeed true)? In other words, what makes it the case that the prohibition of eating with one’s mouth open really is a standard of western etiquette? The answer to this question would be that the conformity to this standard really does serve to ease social relations among new acquaintances. Given that this is what the conformity to the standard does, this is what makes the judgment about it true.

Now, this example about the way truth conditions function regarding etiquette judgments is not to be taken as a rule about how the truth conditions of *all* normative systems work. For example, the fact that moving one’s rook in a straight line is a standard of chess (and the fact that the judgment that this is the case is true) has nothing to do, necessarily, with the fact that moving the rook this way ‘does’ anything – it is simply because this is what, by convention, we have decided are the rules of chess. (Of course, the rules of chess ‘do’ something in one sense: they combine together to create a game that has a particular beauty and requires a particular combination of skill in geometric and probabilistic reasoning in order to win. But such a game could be realised by a different combination of rules. Therefore, I think it makes most sense to say that what makes any given rule of chess really a rule of chess is due to the fact that convention has established it as such). What I simply wanted to illustrate is that there can be such a thing as a fact about whether a standard is one of a certain type, (or one belonging to a certain system), or not. It is a fact that the rule that one must bowl with a straight arm really is a standard in cricket, just as it really is a fact that the prohibition against wanton murder really is a standard of morality.

There are facts about any given system of standards: what its supposed purpose is, where and in what situations it applies, which generate facts about which standards belong to the system or not.

If we want to resolve the question about whether a given standard is a standard of morality, we must ask the following question: what characteristic does a standard have to have in order to be a standard of morality? In chapters 3 and 5, I will make the argument that a standard must have a particular relationship to *wellbeing* in order to truly be a standard of morality. Thus any moral judgment claiming that something is a given moral standard will be true just in case the standard in question has this characteristic. Spelling out just what this characteristic is will be a major challenge. But at least we can see, at this point, that the standard-based view about moral judgments does not commit one to any rejection of cognitivism. For the view says that moral judgments are not mere expressions of standards (in which case the view would be a non-cognitivist one), but rather contain propositions about standards. Specifically, when one makes the judgment that one morally ought to ϕ, one is judging that ϕing is a requirement, that is, a standard of morality. And there are facts about what does and does not make something a standard of morality, which forms the truth-conditions of these judgments.

One qualification should be made. In saying that the standard-based view of moral judgments is a cognitivist view, this is not meant to preclude the possibility that non-cognitivist mental states may often come along for the ride when moral judgments are made. For people who believe morality to be very important, who give moral considerations a lot of weight, moral judgments will not only indicate beliefs about what the standards of morality are or are not, but will also most likely evince certain desires of theirs to act in accordance with such standards, or evince other desire-like states such as pro-attitudes or con-attitudes. All the standard-based view denies is that the making of moral judgments will consist *only* in these non-cognitivist states. In order for a moral judgment to be made, a belief must be held about what a standard of morality is. Any non-cognitive states in addition to this, though not a conceptually non-negotiable element of moral judgments, are most certainly possible, and maybe even typical.

To recap what has been said so far: there are two views about the way in which moral judgments may be said to be normative. There is the rationalist conceptual claim, which is the claim that moral judgments are judgments about (or entail judgments about) normative practical reasons as a matter of conceptual necessity. The standard-based view denies that judgments about reasons are essential to the making of moral judgments, and claims only that moral judgments need be judgments about whether a given standard is a true standard of morality. I will be arguing in chapter 3 that the standard-based view about moral judgments is perfectly sufficient, and that the rationalist conceptual claim is false.

4. The threat of moral error theory

4.1 Moral judgments as implying external reasons

So far we have surveyed two different ways in which judgments can be normative. Judgments can be normative merely in the sense that they contain propositions about standards. Or, normative judgments can sometimes be used to imply (or consist in) judgments about normative practical reasons. Accordingly, we have two alternative ways of characterising moral judgments. One way is to characterise moral judgments as containing propositions about standards: namely, that a given prohibition, requirement or recommendation is a *moral* standard, which I call the standard-based view of moral judgments. To be clear, the standard based view does not claim that judgments about reasons must be *absent* when moral judgments are made. Once again, people who take morality very seriously, who value morality and who *believe* that moral considerations are reason-giving will, when they judge that one morally ought to ϕ, be making the judgment that one has reason to ϕ. All the standard-based view claims is that judgments about reasons are not a *necessary* feature of the making of moral judgments. If one judges that one morally ought to ϕ, and does not judge (or denies that) this thereby brings a normative practical reason to ϕ, they are not making any conceptual mistake. Of course, the rationalist conceptual claim states the opposite. To make a judgment that one morally ought to ϕ, but to fail to make the judgment (or imply that) one has a normative practical reason to ϕ is to exhibit a failure to grasp moral concepts, or a failure to be making a *genuine* moral judgment. So we have two possible views at hand about moral judgments: the standard-based view, which sees the making of judgments about reasons a merely 'optional' component of moral judgments, or the rationalist conceptual claim, which sees it as essential.

Which view about moral judgments is correct? But firstly, why does it matter, and why in particular does it matter for the discussion of moral error theory and success theory? It matters because of a persuasive argument that Joyce makes in *The Myth of Morality:* namely, that if moral judgments imply a commitment to the presence of reasons, these must be *external* reasons. If, then, external reasons do not exist (if reasons internalism is the correct view), then no moral judgments are true. Let's survey Joyce's argument step by step.

Why does Joyce say that moral judgments must be judgments about *external* reasons? This stage of his argument involves two steps. Firstly, Joyce claims that there is a certain degree of 'inescapability' that attaches to moral judgments (an inescapability which applies to any standard or weak categorical imperative). If I make a moral judgment about you, I make it on the assumption that it 'applies' to you regardless of your interests or desires. If I judged that a criminal morally ought not to have killed an innocent person, the criminal's protestation that killing the innocent person served an end of his would hardly cause me to retract my moral judgment (Joyce, 2001: 32). Now, recall Joyce's other contention, that moral judgments are imbued with the kind of 'force' that only something like the rationalist conceptual claim could make sense of. We then have the following argument:

> 1) If x morally ought to ϕ, then x ought to ϕ regardless of whether he cares to, regardless of whether ϕing satisfies any of his desires or furthers his interests.
> 2) If x morally ought to ϕ, then x has a reason for ϕing.
> 3) Therefore, if x morally ought to ϕ, then x has a reason for ϕing regardless of whether ϕing serves his desires or furthers his interests. (ibid: 42)

A way of rephrasing 3) would be to say that if an agent ought to ϕ, then an agent has a reason for ϕing regardless of whether ϕing satisfies an element in his subjective motivational set. Of course, Joyce only mentions two possible elements of the motivational set (desires and interests), but I think it is fairly clear that 3) could apply to any element of the motivational set. Take our criminal again. My judgment that the criminal morally ought not to have killed his victim would not be retracted if he claimed that killing served a desire or interest. But surely, neither would the judgment be retracted if he claimed that killing innocent people was in accordance with his values, personal commitments or 'patterns of emotional reaction' (for example, because it makes him happy). Thus we can interpret Joyce's argument to be saying that moral judgments imply judgments about reasons that are *external*.[18]

Let's now continue with Joyce's argument, following on from 3) above:

> 4) But there is no sense to be made of such reasons.
> 5) Therefore, x is never under a moral obligation. (ibid)

We see now what is at stake in the argument over the rationalist conceptual claim. For the reasons Joyce has given, the rationalist conceptual claim is only plausible as a claim that moral judgments consist in or imply *external* reasons. I will give a further defence of this claim in chapter 3, given that certain relativists (Harman, 1975, 1996) would dispute this. But if external reasons prove to be an indefensible notion, then the truth of the rationalist conceptual claim would spell victory for error theory.

4.2 This thesis chapter by chapter

We are now in a position to fully appreciate why this thesis – a treatment on the relationship between moral judgments, practical reason, and moral facts – is needed. In chapter 2, I will argue that, indeed, reasons internalism is the correct theory of practical reason. Thus, if moral success theory is to prevail, the rationalist conceptual claim had better be false.

In chapter 3, I will argue that it is indeed false. I will argue that moral judgments are primarily defined by their *content*, and that content is what differentiates systems of standards from each other. The moral is distinguished from the prudential, the aesthetic, and the conventional, on the basis of its particular concern for welfare. Chapter 3 will outline just what it means to morally evaluate an action: namely that this involves judging that an action has a particular relationship to the well-being of another. Such an

[18] As I have said, Joyce has confirmed that this is an accurate interpretation of his argument, in case there is still any doubt.

understanding of moral judgments fits nicely into the standard-based view, but renders the rationalist conceptual claim superfluous, or so I shall argue.

In chapters 4 and 5, I attempt to provide what might be called a basis of an in-principle way to ascertain the truth of moral judgments. Whether or not we have the epistemological faculties to have much success in such an endeavour is another matter, and I will not comment extensively on this. What I simply want to provide is an account of the truth-makers of moral judgments, so that the case for success theory can be fully made. In chapter 4, I will give an account of well-being, and in chapter 5, I will give an account of the relationship an action must have to the well-being of another. The relationship will be spelled out in terms of an account of a certain kind of ideal observer's responses.

Finally, in chapter 6, I will make a few concluding remarks about the relationship between morality and practical reason that will, by this stage, have become fairly obvious. I will reach the conclusion that moral actions, among other things are (at least most of the time) reason-giving as a contingent matter of fact.

4.3 Motivations for this project, and further constraints

The possibility that any form of moral scepticism is true, I think we can agree, does violence to some of our most deeply held intuitions. Moral error theory asks us to accept that we are not 'technically' correct when we claim that the holocaust was morally abhorrent, that mother Theresa's actions were morally admirable, that looking after one's children is a moral duty, or that stealing from a stranger is (at least in most circumstances) morally wrong. But this sort of thing is something we are, instinctively, very reluctant to accept. Of course, moral error theory may be consistent with the view that there is still some rationale in continuing to use moral thought and talk (this is Joyce's own view, 2001), but to many of us, this might be cold comfort.

Needless to say, the fact that we hold to some of our moral judgments so strongly does not necessarily constitute a justification for thinking that such judgments are true. But it should, nevertheless, show why many would find this project interesting (and perhaps want it to succeed!). If we possibly *can* vindicate our most deeply held moral judgments, this would make many of us sympathetic to success theory glad.

So my aim to propose an understanding of moral judgments, practical reason, and moral facts that is immune from the powerful kind of attack that I have surveyed in this chapter stems from a desire to vindicate deeply held moral judgments. This same motivation also generates a desire to avoid certain kinds of relativism in the success theory I will generate. The kinds of intuitions that drive us to say that the holocaust was morally abhorrent also drive us to reject any notion that the genocide is 'wrong for us' but was 'right for Hitler'. Note what is being said here: the brand of relativism that I wish to avoid is the brand of relativism that claims that moral judgments have relativistic truth conditions. This is Harman's (1975, 1996) view. I do not mean to deny that some actions can be right in some circumstances but wrong in others, a view which the so-called relativism of David Wong (1991) is much closer to. I would reserve the label 'contextualism' or 'particularism' for such a view, and apply the term 'relativism' only to the view that moral judgments have relative truth conditions. Throughout this thesis, relativism as defined here will be something I shall seek, for the most part, to avoid.

I will not take any stance on whether my view could properly be called 'realist' or not. This too is a term that has a variety of different meanings. If 'realism' denotes the view that moral facts are independent of any given opinion that actual agents might have, then my view is realist. If 'realism' is used to denote the view that moral facts are all together mind independent, then my view probably fails to count as a 'realist' view.[19] Either way, I am not particularly concerned. What matters to me is that the views I will offer will be able to vindicate our most deeply held objectivist moral intuitions, show that they do not rest on a mistake, and give a non-relativistic account of what makes the judgments inspired by them true.

[19] There is an extensive literature on how moral realism is to be defined, and whether mind-independence is essential for moral realism. See Dancy & Hookway 1986, Pettit 1991, and Wright 1988, for discussions on this.

2:
Reasons Internalism

1. Understanding Reasons Internalism

1.1 Introduction

In this chapter, I will be defending reasons internalism. As defined in chapter 1, reasons internalism is the view that all practical reasons are internal reasons, and that there are no external reasons. Such a view was first made explicit by Bernard Williams in his paper 'Internal and External Reasons' (1979).

Evaluating reasons internalism is a difficult task, partly because there is some misunderstanding over what the debate between reasons internalism and reasons externalism is really about (Finlay 2009, discusses the varieties of objections that have been levelled against Williams which are based on misunderstandings of his views). What I will attempt to do in this chapter is give a clear summary of Williams' argument and various objections to it. I will give a defence of reasons internalism and will hope to show that Williams's view, properly understood, is very plausible. In sections 2 and 3 I will consider arguments in favour of reasons internalism and externalism, and conclude that internalism is the superior theory.

1.2 Williams' Arguments

Williams' reasons internalism can be summarised by the following three, interconnected claims:

1) Reasons statements, such as ‘A has a reason to ϕ’ are true in case an element of their subjective motivational set can be satisfied by their ϕing.
2) <R is a motive that a explains the action ϕ of A, where A deliberates correctly, exercises full imaginative capacity, and has knowledge of all relevant facts> gives a conceptual analysis of <R is a normative reason for A>
3) There are no substantive requirements of rationality (only procedural requirements).

For the remainder of this subsection I will explain these claims, and cite where they are found in Williams’ writing, and explain how these claims are connected.

William’s paper ‘Internal and External Reasons’ begins with his claim that there are two possible ways of interpreting reasons statements:

> Sentences of the forms ‘A has a reason to ϕ’ or ‘There is a reason for A to ϕ’… seem on the face of it to have two different sorts of interpretation. On the first, the truth of the sentence implies, very roughly, that A has some motive which will be served or furthered by his ϕ-ing, and if this turns out not to be so the sentence is false: there is a condition relating to the agent’s aims, and if this is not satisfied it is not true to say, on this interpretation, that he has a reason to ϕ. On the second interpretation, there is no such condition, and the reason-sentence will not be falsified by the absence of an appropriate motive. I shall call the first the ‘internal’, the second the ‘external’, interpretation. 1979: 17

Internal reasons statements are statements which, in order to be true of an agent, there must be some ‘motive’ (Williams is yet to spell out what this means) that will be satisfied by acting as the reasons statement commends. External reasons statements have no such condition – when we make external reasons statements we are saying that agents have reasons *regardless of* whether any motive of theirs would be furthered. As we will see, Williams will argue that external reasons statements cannot be true, and he voices this intention to argue so early on in the piece:

> I shall also for convenience refer sometimes to ‘internal reasons’ and ‘external reasons’… but this is to be taken only as a convenience. It is a matter for investigation whether there are two sorts of reasons for action, as opposed to two sorts of statements about people’s reasons for action; indeed, as we shall eventually see, even the interpretation in one of the cases is problematical. Ibid.

Williams goes on to elaborate on the 'internal interpretation', and such an elaboration gives us claim 1) as I have listed it. Williams' first step is to elaborate on what he means by an agent's 'motives', namely that he does not merely mean by this, 'desires':

> The simplest model for the internal interpretation would be this: A has a reason to ϕ iff A has some desire the satisfaction of which will be served by his ϕ-ing… we might call it *the sub-Humean model*. The sub-Humean model is certainly too simple. My aim will be, by addition and revision, to work it up into something more adequate… Basically and by definition, any model for the internal interpretation must display a relativity of the reason to the agent's *subjective motivational set*, which I shall call the agent's S. Ibid: 17-18.

It is important to note Williams' comment 'by definition' here. Williams is drawing a definitional or analytic link between the idea that a reason must be capable of motivating, and the idea that a reason is relative to the subjective motivational set. The subjective motivational set, then, includes by definition *anything that is capable of motivating*. We can assume, given other things Williams says (which we will come to shortly), what the precise relationship between 'motives' and the 'subjective motivational set' is. When we are motivated, we seek to satisfy an element of our subjective motivational set. If, for instance, I am motivated by a desire, the *satisfaction* of that desire is what my motive aims at. Williams goes on to indicate that elements of an agent's subjective motivational set include desires, her values, long term goals, and positive emotional reactions toward things:

> I have discussed S primarily in terms of desires, and this term can be used, formally, for all elements in S. But this terminology may make one forget that S can contain such things as dispositions of evaluation, patterns of emotional reaction, personal loyalties, and various projects, as they may be abstractly called, embodying commitments of the agent. Above all, there is of course no supposition that the desires or projects of an agent have to be egoistic; he will, one hopes, have non-egoistic projects of various kinds, and these equally can provide internal reasons for action. Ibid: 20-1.

Importantly, (and Williams notes this early on in the piece) acting on our motives does not *always* ensure the satisfaction of our motivational set, even though motives aim at such satisfactions. Let's say I am motivated by a desire to drink a glass of gin and tonic to quench my thirst. I reach for the glass of gin and tonic, but unbeknownst to me, it contains petrol. This is an example of an instance in which acting according to my motives will not achieve the satisfaction that is aimed at – due to a false belief of mine. Motives, then, must meet *certain conditions* if acting on them is to bring about what they aim at: the satisfaction of an element of the motivational set. Only motives that meet such conditions count as reasons, and Williams makes it clear that full knowledge of facts is one such condition:

> The agent believes that this stuff is gin, when it is in fact petrol. He wants a gin and tonic. Has he reason, or a reason, to mix this stuff with tonic and drink it?... On the one hand it is just very odd to say that he has a reason to drink this stuff… On the other hand, if he does drink it, we not only have an explanation of his doing so… but we have an explanation of the reason-for-action form…This consideration might move us to ignore the intuition we noticed before… I do not think, however, that we should do this. It looks in the wrong direction, by implying in effect that the internal reason conception is only concerned with explanation, and not at all with the agent's rationality. But it is concerned with his rationality… So I think we should rather say:
>
> A member of S [the subjective motivational set], D, will not give A a reason for ϕ-ing if either the existence of D is dependent on false belief, or A's belief in the relevance of ϕ-ing to the satisfaction of D is false. Ibid: 18-19

The fact that Williams does not consider all motivating reasons to constitute normative reasons is an important point to note. Williams only believes that motives constitute reasons if acting in accordance with them will actually satisfy the relevant element of the subjective motivational set. And in order for one's acting on a motive to be guaranteed to satisfy an element of the subjective motivational set, it must survive a deliberative process that meets certain conditions: one of which is possessing full information. But Williams goes on to specify other elements of this deliberative process that motives must survive in order to succeed at what they aim at: an agent must not only have correct beliefs, but must be drawing valid inferences from those beliefs, and must be using their imagination to its maximal capacity, with regard to other possible courses of action they may not have considered (Ibid: 19-20). This how Williams second central claim, 2), comes about.

A little bit of extrapolation is required to infer claim 3); the claim that there are no substantive requirements of rationality. There are a number of reasons for drawing this conclusion from Williams. Firstly, the constructivist (as opposed to realist) nature of Williams' account of practical reasons lends itself to such a conclusion. In constructivist accounts reasons are constituted by the stances of rational agents, where 'rationality' consists in an agent meeting a variety of procedural requirements. In a constructivist account, any account of what makes a consideration a reason is parasitic on a prior account of what rational deliberation involves. Reasons for action are not 'out there' to be tracked by rational agents, but are, on the other hand, constituted by whatever conclusions rational agents might come to given the specifications of the 'rational' conditions.[20] An implication of constructivist accounts of practical reason is that there are no substantive requirements of rationality. Certainly, on constructivist accounts, it will turn out that some actions are rational and others are not. But this will be as a result of conclusions that would be yielded by a logically prior definition of rationality, and the procedures involved in rational deliberation. In a realist account of practical reasons, in contrast, the account of what makes an agent rational is parasitic on a prior account of what reasons we have. To be rational, on a realist account of practical reasons, is to have faculties that track or detect these reasons, rather than define or constitute them. The implication of realist accounts of practical reasons is that there are substantive requirements of rationality. There are some actions that are rational or irrational because the domain of practical reason requires them, and one's rationality or irrationality can be evaluated on one's ability to grasp (and then be motivated by) these requirements.

Williams' internalist account is clearly constructivist in nature. There are no substantive requirements of practical reason 'out there', which can deem us irrational for not recognising them. Rather, it is our own stances (in Williams' account, specifically, our 'motives') which, having been subjected to certain procedural requirements of rationality, *determine* what our reasons are. The constructivist nature of Williams' account of practical reasons is enough evidence to conclude that his account involves a denial of substantive requirements of rationality. But other evidence in favour of the fact that claim 3) is central to Williams is that, when Williams attempts to talk about what external reasons (which he goes on to deny the existence of) could possibly be, it seems to be something like a 'substantive requirement' that he has in mind. When Williams discusses what it would mean to believe an external reasons statement about ourselves, Williams says that this belief must involve the belief that a certain *consideration* just counts as a reason, even if it did not motivate us at all hitherto. (Although Williams admits to having difficulty conceiving what it could possibly mean for an external reasons statement to be true (1979: 22), he does say that external reasons are things that we can have beliefs about; particularly with

[20] For a discussions on these difference between realism and constructivism about practical reasons, see Wedgwood, 2002

regard to ourselves). Williams makes this point in his discussion of Owen Wingrave, whose motivational set contains no element that would be satisfied by his joining the army. Nevertheless, Owen Wingrave could come to have an external reasons belief that he should join the army because,

> …he believes of some determinate consideration that it constitutes a reason for him to ɸ. Thus Owen Wingrave might come to join the army because (now) he believes that it is a reason for him to do so that his family has a tradition of military honour. Ibid:22

If a determinate consideration (such as family honour) does just constitute a reason for me to act, this is equivalent to saying that the action that the consideration favours is a substantive requirement of rationality (in the example: that doing what is in one's family honour, at least in some relevantly similar circumstances, is required by practical rationality). So Williams appears to be suggesting that having a belief in an external reasons claim would involve believing that some substantive requirement applies to me.

Williams goes on to claim that such a belief can indeed motivate (ibid: 23), and therefore that acting upon such a belief can end up satisfying an element of the motivational set after all. But this is not enough to prove that external reasons exist. Williams says, on the other hand that the claim that an external belief can motivate an agent is '*so* plausible that this agent, with this belief, appears to be one about whom, now, an *internal* reason statement could truly be made: he is one with an appropriate motivation in his *S*'. (ibid) In order for it to be proven that external reasons exist, Williams argues, it must be the case not just that beliefs about external reasons motivate, but that they are *true* – in other words, that there exist substantive requirements of rationality, rather than mere beliefs about them. Williams hints that a realist account of practical reasons, (reasons which our rationality tracks or detects), would be needed if external reasons statements were to be true, when he says:

> The basic step lies in recognising that the external reasons theorist must conceive in a special way the connexion between acquiring a motivation and coming to believe the reason statement… that the agent should acquire the motivation *because* he comes to believe the reason statement, and that he should do the latter, moreover, because, in some way, *he is considering the matter aright*. Ibid: 24 (my emphasis).

In other words, for external reasons to exist, it must be the case that an agent's rational thought processes *rightly track* them, and upon discovery of such reasons, their motivations to act in accordance with such reasons follows. A failure to be motivated by considerations of a particular kind (whichever the externalist wants to claim constitute substantive rational requirements) then becomes part of the definition of what it is to be irrational.

Williams goes on to discuss what he thinks is an implication of this externalist/realist picture of rationality and reasons:

> …the external reasons statement itself might be taken as roughly equivalent to, or at least as entailing, the claim that if the agent rationally deliberated, then, whatever motivations he originally had, he would come to be motivated to ϕ. Ibid: 24.

In other words, if there are substantive requirements that are independent of the motivations of agents, then being rational *must* consist in being able to recognise these requirements and be motivated to act in accordance with them, regardless of any motivations that the agent currently has. What the external reasons theorist must show is that *being rational* consists not just in deliberating in a clear, well informed, and imaginative way from whatever motives one happens to have, but that *being rational* consists in one *being motivated by certain substantive considerations*. Williams himself confirms the validity of my interpretation in a later work when he writes, (with the intention of adding more clarity to 'Internal and External Reasons'):

> The claim that somebody can get to the conclusion that he should ϕ (or, the conclusion to ϕ)n by a sound deliberative route involves, in my view, at least correcting any errors of fact and reasoning in the agent's view of the matter… We are allowed to change – that is, improve or correct – his beliefs of fact and his reasonings in saying what he has reason to do. That is already enough for the notion to be normative.
>
> But if we are licensed to vary the agent's reasoning and assumptions of fact, it will be asked why we should not vary (for instance, insert) the agent's prudential and moral considerations as well. If we were allowed to adjust the agent's prudential and moral assumptions to some assumed normative standard, then obviously there would be no significant difference between the internalist and the externalist accounts. We would have incorporated into the notion of a 'sound deliberative route' anything the externalist could want… To the extent that the agent already has prudential or moral considerations in his S, of course, they will be involved in what he has reason to do. They will contribute to an internal reason. 1995 [1989]: 36-7.

In other words, responsiveness to substantive requirements (of a prudential or moral type, for instance)[21] is something the externalist might build in to the conception of what it takes to be rational. But the internalist, on the other hand, will keep the account of rationality strictly procedural. Parfit (1997) has also had the insight to realise that this is the key difference between Williams' internalism, and reasons externalism:

> Many Externalists would claim that, if we knew the relevant facts and were fully rational, we would be motivated to do whatever we had reason to do. This claim is not, as it may seem, a concession to Internalism. According to these Externalists, if
>
> R) we have a reason to do something,
>
> that entails that
>
> E) if we knew the relevant facts, and were fully *substantively* rational, we would be motivated to do this thing.
>
> To be substantively rational, *we must care about certain things*… 1997:100-101, emphasis added.

In summary, we have seen that Williams argues for three claims in 'Internal and External Reasons'. The first claim is that, for a reasons statement to be true, it must be the case that acting as the reasons statement directs ensures satisfaction of an element of an agent's motivational set. Accordingly, 'reasons' are just a kind of motive: a motive that (due to its meeting certain procedural requirements of rationality), when acted on, ensures the satisfaction of an element of the agent's motivational set claim 2).[22] An implication of this analysis of normative reasons is that there do not exist the kind of 'reasons' in which the externalist believes: that of substantive requirements of rationality. The stances (or 'motives') of agents do not track independently constituted rational requirements, but constitute them. In other words, there are no such things as motives that an agent is by definition irrational for failing to possess. While the externalist believes that certain types of motives can be irrational by virtue of their being *motives of a certain type* (for instance,

[21] It is worth my noting, at this point, that on the view of prudence I will be defending, the standards of prudence just do constitute reasons, because what is prudential or good for us is, as I will argue in chapter 4, just whatever would satisfy our motivational set. Accordingly, though I believe the standards of prudence to be reason-giving, this does not undermine Williams' view: that *substantive* requirements of prudence (such as what one might find in an objective-list account of prudence; see chapter 4) are not built into the conception of rationality.

[22] Finlay 2009 also emphasises this conceptual dimension of Williams 1979 work: namely that it essentially attempts to give a conceptual analysis of normative reasons in terms of explanations; that is, motives.

immoral motives), the internalist believes motives are only irrational if acting on them would fail at what they aim at: the satisfaction of an element of the subjective motivational set. And the way to test this is to see whether motives would survive the deliberative process Williams outlines.

There are aspects of Williams' 1979 paper which remain puzzling, and in closing this subsection I will mention some of these. Firstly, one might ask on what basis Williams justifies claims 1) and 2). One possible answer is that claim 2) itself provides the justification for claim 1). And there is some evidence for this in the text, as Williams does indeed appear to argue that the concept of a reason must essentially involve the concept of a motivation, because of the fact that motivations explain actions:

> This explanatory dimension is very important, and we shall come back to it more than once; if there are reasons for action, it must be that people sometimes act for those reasons, and if they do, their reasons figure in some correct explanation of their action. 1979: 18

It might seem then, that Williams's basis for analysing reasons as a kind of motivation is based on the fact that reasons must be explanations. But if this is the case, then Williams restriction on what *kind* of motives count as reasons is an arbitrary one – for if the reason Williams thinks that reasons are motivations is essentially because reasons must explain action, then this would entail that *all* motives (not just motives that guarantee the satisfaction of an element of S when acted on) are reasons. This would then leave claim 1) without justification.

Alternatively, we could see things the other way round, and see claim 2) as being justified by claim 1). In other words, Williams provides the restrictions on which motives constitute reasons that he does precisely because he wants to achieve the result that reasons claims are true if an element of the agent's motivational set is satisfied. But this reading has other problems. Firstly, it is hard on this reading to see why Williams' 'explanatory dimension' is as important as he says it is. Indeed, the explanatory dimension is unnecessary. For, according to claim 1), a reasons statement can be true of an agent *regardless* of whether an agent is motivated to act in accordance with it or not, and, therefore, whether an agent's action could be explained by such a reason or not.

A more serious problem raises its head. If Williams' analysis of reasons as given by claim 2) rests on the logical priority of claim 1), then one must ask what the basis for claim 1) is. 1) seems now to amount to an assertion that we just have reason to satisfy elements of our motivational set. But this would seem like a *substantive* requirement of rationality, and would so render Williams' constructivist analysis of reasons and rationality in claim 2) superfluous (if not contrary to the very reasons for believing 1)). Why shouldn't Williams just instead take the realist routes and say that there is a substantive requirement of rationality (to satisfy elements of our motivational set), and analyse an agent's rationality in terms of their ability to do this? Furthermore, if Williams allows that there is one substantive requirement of rationality, then there would seem to be no principled way of saying that there could not be others – such as moral requirements (this is the essence of one of Korsgaard's (1986) objections to Williams). It would seem then that claim 3 in Williams is clearly false.

In summary, the 'explanatory dimension' that seems to underlie claim 2), renders claim 1) false. And if claim 1) is Williams' primary claim, then claim 2) seems unnecessary, and claim 3, false. So how do we reconcile these seemingly incompatible strands in Williams? In section 2), I will attempt to do just this by citing Williams' subsequent work, which sheds a bit of light on these apparent inconsistencies. I shall hope to show that, once these difficulties are made sense of, Williams' view is a very plausible one.

1.3 Situating other philosophers across the internalist/externalist divide

At the outset of this section it is worth noting that the difference between reasons internalists and externalists has been understood differently by different philosophers. Darwall classifies himself as a reasons 'internalist' (1983: ch 5), but by this he means simply that he subscribes to the platitude discussed in chapter 1: that a judgment about one's reasons will motivate (ibid: 32). But what Darwall goes on to say – that there are reasons to be motivated by the considerations that would move an impartial agent – clearly places him in the externalist camp in Williams' sense. Nagel is similar. For Nagel clearly believes there are substantive rational requirements, and that altruism is one of them (1970: 1). But Nagel also makes it clear that he believes that considerations constitute *reasons* because they are capable of motivating in some capacity, which is why he explains his search for the reason-giving power of morality in terms of a 'motive' that can be found (ibid: 7-12). One might then be tempted to interpret Nagel as arguing something like the following: that there are substantive rational requirements to be responsive to, in virtue of the fact that such requirements are to be found among our motives – as part of our motivational set. Thus there is a substantive requirement: to satisfy our motivational set, and obey whatever rational requirements do this. But the spirit of the following passage from Nagel clearly goes against this idea:

> Kant's effort to produce a categorical imperative is an attempt to discover requirements on action which apply to a man on no conditions about what he wants, how he feels, etc… The position which I shall defend resembles that of Kant in two respects: First, it provides an account of ethical motivation which does not rely on the assumption that a motivational factor is already present among the *conditions* of any moral requirement. On this view the possibility of appropriate motivation must be guaranteed by the truth of the moral claim itself – but *not* because the existence of such motivation is included in advance among the independently comprehensible truth conditions of every moral claim. There are reasons for action which are specifically moral; it is because they represent moral requirements that they can motivate, and not vice versa. Ibid: 13.

We see here that Nagel is keen to avoid the idea that reasons to be moral are contingent on prior desires. So in what sense does he believe that moral requirements are, or must be, 'motivating'? A later passage from Nagel gives an answer:

> The more central and unavoidable is the conception of oneself on which the possibility of moral motivation can be shown to depend, the closer we will have come to demonstrating that the demands of ethics are inescapable… the principle of altruism… is connected with the conception of oneself as merely one person among others. It arises from the capacity to view oneself simultaneously as 'I' and as *someone* – an impersonally specifiable individual. Ibid: 19

What Nagel is arguing is that a realisation of a fact – that one is a person among others – is capable of motivating. Now, importantly, Williams would not deny this either. Williams would agree that becoming aware of facts has the potential to change one's motivational set. Is there, then, no difference between Williams' and Nagel's view?

The answer to this question, I believe, is 'no'. Nagel appears to be claiming both that substantive rationalist requirements exist, and that these (moral) requirements should motivate: the realisation that one is a person among others *should* (rationally) motivate one toward altruism. Whereas Williams would say no such thing; the conception of oneself as one person among others *may* motivate one to be altruistic, but it does not, on pain of irrationality, *have to*. It seems then that Nagel is an externalist in Williams' sense. He is committed to showing that there are such things as substantive requirements, which lead him to make the claim that, upon recognising them, we will be motivated to act accordingly. Thus Nagel does what every other externalist does: he builds into the definition of 'rationality' from the start a capacity to be motivated by considerations of a certain type, which he claims constitute substantive requirements.

An objection Korsgaard makes against Williams is that his internalism simply begs the question against external reasons, or, to be more precise, substantive requirements of rationality (Korsgaard uses the term 'requirements of pure practical reason', (1986: 21-5) by which, given the context and arguments of the paper, we can take her to mean 'substantive requirements of rationality'). What Williams argues in favour of, according to Korsgaard, is a platitude mentioned in chapter 1, which she calls the 'internalism requirement': that, in order for something to count as a reason, it must be able to motivate an agent insofar as that agent is rational (ibid: 11). Being the platitude that it is, *any* practical reasons theorist (internalist or externalist) ought to accept it. The internalism requirement does not, however, entail that there can be no substantive requirements of rationality. Reasons must motivate rational agents. But being 'rational' could well involve meeting substantive requirements, as well as procedural ones, as Korsgaard goes on to point out. The internalism requirement cannot place limitations on what principles might count as principles of practical reason (ibid: 22)

Korsgaard is right to say that the internalism requirement does not preclude the possibility of substantive requirements of rationality. But Williams is not defending *merely* the internalist requirement (that reasons must motivate agents insofar as they are rational) where 'rational' is given a 'whatever that might mean' interpretation. Williams is defending a specific *version* of the internalist requirement, where he cashes out 'rationality' in very specific, and purely procedural, terms.

The question of whether Michael Smith belongs in the internalist or externalist camp in Williams's sense is one to which I have an answer that is not completely confident.

Certainly, he subscribes to the rationalist conceptual claim, which has also occasionally been referred to as ‘reasons internalism’ (Brink 1989: 37-45). But the reasons internalism we are talking of here is the view that all reasons are internal reasons. And where Smith stands on this question is not entirely clear. On the one hand, Smith voices a near-agreement with Williams’ procedural account of rationality. Smith’s only addition is that the agent’s desires be ‘systematically justifiable’ (1994: 158-161, 1995: 114), by which he means that the deliberative process ought to lead to the agent acquiring desires that form the largest possible coherent set. So far, it could be suggested, there is nothing necessarily externalist about this. But other arguments from Smith suggest a distinctly externalist flavour. Firstly, Smith criticises what he calls the agent-relativity of Williams’ account, on the back of his own concept of reasons, which entails a claim of his that the very concept of normative reasons implies convergence (1995: 122-125, 1994: 172).[23] If the concept of normative reasons implies convergence, then this must mean convergence *on some particular conclusion about what to do*. Thus, Smith can be taken to be implying that the concept of normative reasons involve the concept of substantive rational requirements. Accordingly, I will treat Smith’s objections to Williams as objections to reasons internalism itself, and treat those objections in section 3.

Finally, philosophers on the ‘internalist’ side (though the following philosophers adopt a version of internalism that is narrower than Williams’, such as the Humean or instrumentalist account, as I have said in chapter 1) include Foot (1972), Harman (1975,), Joyce (2001) and Schroeder (2007).

2. In Favour of Reasons Internalism

2.1 Truth conditions of reasons judgments

We saw, at the end of our look at Williams’ ‘Internal and External Reasons’, that there are various elements of the paper that do not sit well together. The aim of this section is to reconcile these elements, and show that Williams’ reasons internalism is a consistent and plausible view. Throughout this section, I will look at some arguments that can be made in

[23] Sobel, 1999: 143, makes this observation about Smith’s concept of reasons as well.

favour of reasons internalism and externalism, and conclude that reasons internalism comes out on top.

Let's recap on the apparent difficulties in Williams' paper. On the one hand, Williams seems to think that the concept of a reason is essentially the concept of an explanation of intentional action and therefore a motive. But this explanatory basis cannot support his further claim that only motives of a certain kind, (motives which guarantee the satisfaction of elements of the motivational set when acted upon) constitute reasons. But, on the other hand, if Williams wants to insist that only these motives count as reasons because satisfying our motivational set is just something we have reason to do, this claim seems equally baseless (and arbitrary given that Williams bars all *other* substantive requirements from existence), and his procedural analysis of rationality seems superfluous.

I believe this tension can be resolved, if we reconstruct Williams's argument in the following way. If an agent judges that she has a reason to ϕ, this can be said to be a reason she acts *for* (in the familiar terminology, her 'motivating reason'). But to *judge* that one has a reason to ϕ (or have the concept of a normative reason to ϕ) involves the judgment, according to Williams, that we would be motivated to ϕ upon deliberating carefully about it, arriving at full knowledge of relevant facts, and upon following certain other procedures like the use of our imagination. Williams hints at this in later work when he writes:

> The grounds for making this general point about fact and reasoning… are quite simple: any rational deliberative agent has in his S a general interest in being factually and rationally correctly informed… So the claim that he has a reason to ϕ... introduces the possibility of that reason being an explanation [of his action]; namely, if the agent accepts that claim (more precisely, if he accepts that he has more reason to ϕ than do anything else). When the reason is an explanation of his action, then of course it will be, in some form, in his S, because certainly – and nobody denies this – what he actually does has to be explained by his S. 1995 [1989]: 37-9

Here we see Williams making the claim that making a judgment about one's reasons involves making a judgment about what one would be motivated to do after going through the procedures he thinks necessary counting as 'rational': when he says that a rational deliberative agent has in his S a general interest in being factually and correctly informed, this is a way of saying that, if I as a deliberating agent, want to know what my reasons are, this will mean that I will have some belief that being informed (and meeting his other procedural requirements) will have something to do with this. In other words, my judgments about what my reasons are will be judgments about what I would be motivated to do when properly informed, and so on for the other features of Williams' procedure: having deliberated correctly, and used my imagination. Why 'about what I would be *motivated* to do'? Because, as Williams notes, the reasons that we act *for* motivate us. And we saw in chapter one that the basic terminology of motivating and normative reasons assumes that judgments about what our normative reasons are, motivate us (given that they are called *motives*). But we would not call just *any* motive a reason: only those motives that would survive the kind of deliberative process that Williams outlines. And, to relate this back to another point, such reasons arrived at will, necessarily, constitute satisfactions of our motivational set. The same necessary conditions under which motives count as reasons are the same conditions under which motives count as those which will satisfy elements of our subjective motivational set.

We can see, then, that there is no conflict between the explanatory dimension and Williams' account of what makes a motive a reason. When we judge that a motive is a reason, it explains our action, it is a reason that motivates us, and becomes a reason we act *for* (assuming we follow through on the motive). But part of what *judging that a motive is a reason involves* is judging that the motive is one that we would retain if it could survive certain rational procedures. This analysis of what judging oneself to have a normative reason involves is fairly sound. Suppose I believe I have a reason to act on an existing motive. How might my conviction be shaken? Maybe you could shake my conviction by saying that I have not *thought* or deliberated clearly or carefully about the action I am about to do. Maybe you could also shake my conviction by saying that there is some fact that I do not know about the action that I am about to do (as there is in Williams' gin drinker example). Maybe you could cite other procedural problems with my decision to act: surely, if I used my imagination more, I might realise that there are other ways and other courses of action that might be beneficial to achieving the ends that I want. Unless I was conceptually confused about what it meant to believe I had a normative reason, one would expect me to change my mind if I believed any of the things you said were true.

This resolution of the apparent tensions in Williams' view also gives us a fairly compelling argument for reasons internalism. I will give it here:

1. If judgments about what it is to have a reason involve judgments about what one would be motivated to do if such a motive survived a certain deliberative procedure, then true judgments about what it is to have a reason involve true judgments about what one would be motivated to do if such a motive survived a certain deliberative procedure.
2. Judgments about what it is to have a reason involve judgments about what one would be motivated to do if such a motive survived a certain deliberative procedure.
3. Therefore, true judgments about what it is to have a reason involve true judgments about what one would be motivated to do if such a motive survived a certain deliberative procedure.

In a more general form, the argument can be put as follows:

1. If a judgment about X involves a judgment about Y, then true judgments about X will involve true judgments about Y.
2. A judgment about X involves a judgment about Y.
3. Therefore, true judgments about X involve judgments about Y.

Such an argument, in my view, constitutes a compelling case for reasons internalism. If the argument cannot be faulted, a big presumption in favour of reasons internalism has been gained. But, in 2.2, we will see that externalists might raise doubts about premise 2.

2.2 An externalist reply

When we are making judgments about what we have reason to do, externalists might claim, we are *not* merely making judgments about what considerations or motivations would survive a deliberative process. We are, to use Wedgwood's phrase: *trying to figure out what is best* (2002). We are *using* our deliberative procedures precisely because we expect such procedures to track substantive requirements (ibid: 141). To put the idea as Williams puts it: the externalist may have in mind not only that there are certain procedural requirements we must meet in order to count as rational, but that there are certain substantive requirements too, such as moral requirements (1995: 36). And indeed, an externalist might claim that the interest to act in line with certain standards like the moral (or the aesthetic, or the conventional) is precisely one of the interests we have when we make judgments about what we have reason to do. This objection amounts to a claim that Williams's analysis of what it is to make a judgment about one's reasons is inadequate.

The externalist's suggestion may be supported by a particular example. Take John, who is trying to work out whether he should donate money to charity. In other words, he is trying to work out whether the judgment that he has reason to donate to a certain charity is a true judgment. In working this out, John may *not* merely be working out whether or not the motive to donate to a certain charity survives certain rational deliberations of his, but John may also have an interest in doing something *altruistic*, and may be trying to work out whether his giving to this particular charity really would benefit others (suppose John has given money to 'charities' before, who turned out to be fraudulent or incompetent). John, possessing the belief that altruistic acts are substantive requirements of rationality, then, is not merely trying to work out what motives of his would survive a Williams-style deliberation, but is trying to work out whether giving to this charity is a *substantive rational requirement*. Thus, the externalist could use my kind of argument against me:

1. If judgments about what it is to have a reason involve judgments about what one's substantive requirements are, then true judgments about what it is to have a reason involve true judgments about what one's substantive requirements are.
2. Judgments about what it is to have a reason involve judgments about what one's substantive requirements are.
3. Therefore, true judgments about what it is to have a reason involve true judgments about what one's substantive requirements are.

I think we will have to admit that the externalist has a point here. The phenomenology of making judgments about our normative practical reasons very often involves a feeling of wanting to meet certain external standards. However, the fact that this is (merely) *often* the way that making a judgment about reasons feels does not get the externalist what she wants.

The point at which the externalist's counter-argument fails is at the simple fact that not everyone is like John. It is completely possible to imagine a kind of person who has no inherent belief or respect for any kind of altruism (or any other feature of action, for that matter). And yet this person would still most likely agree that the motives of hers that count as reasons are the ones that would survive the kind of deliberation that Williams has outlined. Imagine this person is about to steal from a collection plate. We might condemn her, tell her she has a reason not to do this, because this is a very cruel and un-altruistic act she is doing (if we reasoned like this, we would of course, be talking like externalists who assume that altruistic or moral acts are substantively rationally required). It is entirely possible for her to agree with us that her act is cruel and un-altruistic, but persist in the conviction that she has reason to perform the act, without being conceptually confused in the slightest. But if, however, we convinced her that there might be a fact about stealing the collection money that she has overlooked, or that she has not thought hard enough about what she is doing, she would certainly suspend her judgment that she has a reason. The point I am making is that, although *some* competent users of the concept of a normative reason (good-hearted John being an example) make judgments about substantive requirements when they make judgments about reasons, not every competent user does. But every competent user of the concept of the normative reason *must* have the concept of a motive that would survive the kind of deliberative process Williams talks of.

It might be rebutted at this stage that it has been thought that *it is* conceptually confused to acknowledge that an action morally right, and yet fail to judge that one has a reason to perform the action (Smith 1994, Prichard 1912). Thus, it could be said, there is at least *one* kind of substantive requirement that is built into the concept of being rational – moral requirements. But, as I shall argue in chapter 3, this, to my mind, rests on the fact that the term ‘moral’ is often used in a peculiar sense. According to a view defended by Hare (1952) one’s ‘moral’ judgments simply express those values or ideals one believes in most earnestly, which entail self-directed imperatives when made first-personally. In a similar vein, Smith, 1994: 71, also makes the claim that a change in one’s moral convictions constitutes a change in one’s ‘fundamental values’. But one’s ‘moral judgments’ in this sense is simply a synonym for what one might take to be one’s reasons and are not judgments about what I call *morality proper*: judgments about that system of standards that generally condemns harmful actions, and requires refraining from harm. Paridigmatically moral concerns, the concerns of morality proper, typically involve concerns to do with being altruistic, refraining from doing actions of a certain type (and chapter 3 will expand on this idea and give it more precision). But one’s personal ideals and commitments need not have any altruistic concerns, or any concerns about others' welfare. In other words, we arguably have two separate concepts of one's 'moral judgments' here: one concept of moral judgments as one's deeply held values, whatever they might involve, and another denoting judgments about the requirements of 'morality proper' that system of norms concerned with the wellbeing of others and the impact of one's actions on this. One's 'moral' commitments in the first sense need not imply that one has 'moral' commitments in the second sense. And it is only on the definition of moral requirements in this first sense that one could say that it is conceptually confused to judge that ϕing is morally required but fail to judge that one has a reason to ϕ. But if we're talking about moral requirements in the second sense, as the requirements to do actions of a certain altruistic, welfare-concerning kind (as in the John example), it would be an uphill battle to show that there was any conceptual confusion in claiming one had no reason to do one's moral requirements in this sense. (Others who have paid similar attention to the definition of 'morality' and made a similar claim to mine here one here are Brink 1989: 37-51, Cullity 1997: 108, Williams 1981: 22-4 Foot 1972: 310-2).[24] The externalist's counter-argument, and the example that supports it, must rely on the assumption that, if moral requirements are among the substantive rational requirements we have, what must be meant is moral requirements in the second sense (for, if the claim is taken to be given in the first sense, the claim would be that it is rationally required to act in accordance with one's ideals and values… but this is nothing more than what an internalist would happily grant). Thus, the externalist could not make the counter-argument above and marshal in support of it the familiar claim that it is conceptually confused to judge that ϕing is morally required but fail to judge that one has a reason to ϕ. To argue like this would commit a fallacy of equivocation between the first sense of 'moral requirement' and the second sense of 'moral requirement'.

[24] Brink's amoralist arguments, and his claims that the amoralist is conceivable, rely, for their plausibility on the assumption that the claim that one *'morally ought'* to do something denotes a particular range of concerns, which one may, or may not (as Brink's amoralist argument intends to show), be reason-giving. Foot, also, sees no inconsistency in doing what one believes to be immoral, on the understanding that morality (like etiquette) is a system of rules with particular standards, about which one may have the opinion that they are either reason-giving or not. To flout morality is to be a *villain* (ibid:310) – someone who disregards the welfare of others – rather than to disregard one's personal values, as a according to the first sense of 'moral requirements'. And Williams paints the picture of Gauguin, one whose personal values and ideals are quite other than that of what I have called morality proper. Williams' example is important to this debate simply as it illustrates that there is no conceptual necessity in one's moral concerns in the first sense (one's personal ideals) being moral concerns in the second sense (the concerns of morality proper) – an assumption that I will accuse Smith (1994) of making, in chapter 3.

2.3 What we want from a theory of practical reason

To reiterate, what the external reasons theorist has to show is that one would fail to be rational *simply* by failing to be motivated by a consideration of a certain type (even after having gone through a deliberative process). As Parfit (1997) neatly characterises the externalist's view, externalism says that we have reasons to perform actions or refrain from performing them on the basis of features that those actions have (as opposed to features about ourselves or our motivational set). Suppose, then, one is an externalist, and one thinks that one has reason to do an action because it has this feature: because, for instance, it is *altruistic*.

The question of *how* one would establish that the feature of altruism provides a substantive rational requirement brings to mind another conceptual shortfall of reasons externalism. For supposedly, not just *any* features of actions confer substantive rational requirements: there must be a way of working out which features of actions are reason-giving and which are not. And this, supposedly, is precisely what a theory of practical rationality is supposed to do: it is supposed to help us answer the question of which kinds of actions are reason-giving, and which are not. Reasons internalism serves this purpose, but reasons externalism *presupposes* answers to that question from the outset.

Reasons internalism, on the other hand, does what we presumably want a theory of practical reason to do: it gives us a principled way of working out which actions are rationally required and which are not: it says that those motives to perform actions that would survive a certain procedure would count as rational. Thus it does not assume answers to the question of what kind of actions we should do into its theory of rationality. A theory of rationality should help us to satisfactorily *answer* substantive questions about whether we have reason to do actions that are altruistic, or follow the rules of etiquette, and the like, not assume answers to these questions by definitional fiat. But by building into the conception of rationality responsiveness to certain substantive considerations, this is exactly what reasons externalism does: answers our questions by definitional fiat. It is little wonder that Williams criticised appeals to external reasons as amounting to little more than 'bluff' (1979: 26).

This advantage that reasons internalism has over reasons externalism is an important one. For it is common practice for us to question (or at least to have the propensity to question) whether we have reason to commit certain acts *just* because they have a given feature, or conform to a particular standard. Suppose that I am normally inclined to act in accordance with what I believe to be morally right. In other words, if I think that an act has the feature of moral rightness, I tend to be motivated to act in accordance with it. But suppose, on one occasion, morality demands a lot of me. There is an act that I think is morally right, and yet acting will be very costly. We tend to think it would be perfectly legitimate for me to ask: 'do I really, after all, have reason to do this act?', or 'do I, after all, have reason to do something just because it is morally right?' (keep in mind that we typically have the second sense of 'morally right' in mind when we ask this question, lest the accusation again arise that such a question is not a sensible one). The practice of asking (often weighty) life questions which concern the rationality of our actions, and which question the notion that a certain action may be rationally required of us *just* because it has a certain feature, is a practice we engage in all the time. Any theory of practical reason that implied that such a reflective practice was illegitimate would have a serious drawback. But reasons externalism does imply, that at least on some occasions, such a practice *is* illegitimate: for whatever features of actions, or standards the externalist wants to incorporate into the notion of being 'rational', it will turn out that it is conceptually confused to ask the question of whether one should perform an action just because it has that feature. This, I would suggest, goes strongly *against* normal intuitions about what practical rationality involves. The openness to question dogmas that certain actions ought to be done simply because they have a certain feature is commonly thought to exhibit the very *height* of reflectiveness and practical rationality. This, certainly, is what many great philosophers and reformers have been admired for.

It is certainly an understandable motivation that reasons externalists have, to try to show that the moral life (among other things) is one that practical rationality also supports. It is no accident that the most ardent defenders of reasons externalism have adopted the theory for the end of showing that moral behaviour is, in the end, rationally required (for instance, Nagel 1970, Darwall 1983, Gewirth 1978, Parfit, 1984, Smith, 1994). The idea behind the approach is quite simple: if one builds into the definition of what it takes to be rational a motivation to do what is moral, then it can't help but be the case that one has reason to do what is moral. But such a victory, I hope I have shown by now, is rather cheap. I too confess to a desire to find it to be the case that the moral life is (or is largely) rationally justified, but I would suggest that there is a far more satisfying way of securing this result than what the reasons externalists offer. My approach is to embrace reasons internalism, which is the more plausible theory of rationality for a whole host of reasons, and then to show that moral standards are, as a matter of contingent fact, reason-giving (or are at least reason-giving most of the time). Surely this kind of answer to 'the normative question',[25] which relies in no way on definitional fiat, is far more satisfying.

It is worth saying at this point that reasons externalists are not the only ones who are guilty of ensuring the convergence of moral requirements and rational requirements through cheap definitional means: internalists have done this too. While the externalist's move is to build into the concept of 'rationality' a responsiveness to the requirements of morality proper, the trick of certain internalists (Harman, 1975), is to build into the concept of 'morality' just the concept of acting in accordance with whatever reasons one has: a redefining of 'morality proper' into 'morality-as-one's-ideals-or-reasons'. Of course, this internalist move has very different implications: it tends to lead to relativism, rather than to the absolutist theories of the externalists. But the mistake is similar in one respect: to try to secure for 'morality' (whatever one might take that to mean) as tight as possible a relationship with practical rationality by definitional fiat. In both cases, the result is unsatisfying. The externalist approach is unsatisfying because the 'why be moral?' question has essentially been dismissed as conceptually confused rather than answered. And the internalist approach is unsatisfying because the subject has been changed: it is no longer morality in the sense of what I have called 'morality proper' that is being talked about, but morality proper was what we were interested in. I will discuss this more at length in chapter 3.

My approach, on the other hand, is to avoid pulling any definitional punches, and to keep the concept of moral requirements, and the concept of normative practical reasons entirely separate. Moral requirements are to be understood as belonging to a certain *system of*

[25] This is the term Korsgaard uses to describe the question of whether morality is rationally justified, 1996: 7-49

standards which are unified around a certain cluster of welfarist concerns. Normative practical reasons are as the internalist says: they are motives which, if acted on, would secure the satisfaction of an element of the subjective motivational set, which means also that they are motives which would survive a particular deliberative procedure. Although the concept of a moral requirement, and the concept of a reason for action are entirely different, it will become clear by the end of this thesis that we probably (as a contingent matter of fact) have reason to make moral concerns an important concern in our lives.

2.4 'Ought' implies 'can'

I conclude this section by bringing to light one final problem with reasons externalism, which reasons internalism does not suffer from.

A serious problem I believe reasons externalism faces is that it implies that the motivational set of an individual is in and of itself rationally criticisable. Reasons externalism is committed to claiming that individuals should be motivated to do particular things or meet particular substantive requirements. If individuals are not motivated as such, they are 'irrational' – their motivational set is not as it should be. That reasons externalism implies that the motivational set itself can be rationally criticised is highly problematic. A reasonable intuition says that in order for something to be rationally criticisable, it must be the case that it satisfies the 'ought implies can' principle. In other words, if the motivational set is rationally criticisable, it must be something over which we have control.

I must be careful as I make this claim, not to contradict anything said so far. For it is certainly the case that elements of our motivational set can change through processes of deliberation. Suppose I am initially a cold and callous person, but, after much reflection and use of my imagination, come to appreciate and empathise with the horrible effects my actions have on other people. Through this process, my motives change, and with this, the elements in my motivational set. Once, my desires were satisfied by my performing cruel and self-centred actions. Now, my desires are satisfied by my performing kind and compassionate actions. So the statement that we *do not have voluntary control* over our motivational set is not meant to amount to a statement that our motivational set cannot change as a result of something that we voluntarily do (i.e., engage in the deliberative process).

In what sense, then, is it the case that our motivational sets lie beyond our voluntary control? My answer is as follows. When we undertake processes of deliberation, we do not necessarily have control over the *way* in which our motivational sets will be changed or not changed. Take the example of a second callous person who goes through a process of deliberation. Like me, they picture the same facts, deliberate just as clearly as me, but their fundamental desires do not change. It remains a fact about them that no intrinsic, informed desire of theirs consists in making other people happy. Their motivational set is stubbornly callous.[26] There are many other elements of our motivational sets that we could think of as fixed, without having to cite such extreme examples. It is a relatively fixed feature of my motivational set that I am attracted to men, and not women. I could subject this desire to any amount of rational deliberation and this fact about me probably would not change: that is to say, my attraction to men just is an entrenched element of my motivational set and, as such, cannot seriously be said to be something that I have chosen. In virtue of the fact that I cannot choose to either have or not have this element in my motivational set, it surely cannot be the sort of thing that can be rationally criticised. Given that motivational sets, in a sense, just *are the way they are*, in much the same way that a rock just *is the shape it is*, how can it make sense to rationally criticise the motivational set's elements, or the rock's shape?

An important qualification should be made here. The idea that a person's motivational set is not rationally criticisable does not prevent us from being able to morally criticise it – a result, I think, that we should want. On my view, to morally criticise a person means to judge that they, their behaviour, and even their motivational set, fails certain standards. And the idea that one could fail a certain standard does not seem to require, as rational criticism does, that one has voluntary control over one's meeting it. Suppose I wish to qualify for the Olympic high jump event, and in order to do so, I must be able to jump over a bar that is 2 metres in the air. The bar sets a standard – any jump below it fails to conform

[26] Though this kind of person is conceivable, I will be arguing, later on, that every human person plausibly has, as a core element of their motivational set, the propensity to gain pleasure from (and thus for desire to be generated for) close relationships, and compassionate behaviour. Thus, it is unlikely that there are many (perhaps any) human persons who have 'stubbornly callous' motivational sets to this extent, which would remain unchanged by a proper, Williams-style process of deliberation. But this is to be understood as a contingent fact about our motivational sets, not a substantive requirement that is built into the definition of rationality. Certainly, for instance, a creature with a 'stubbornly callous' motivational set is conceivable, even if such a creature is unlikely to be human.

to it, and any jump above it succeeds in conforming to it. Suppose I do not make 2 metres. Despite my earnest efforts and training attempts, I fail: my failure and lack of giftedness is beyond my voluntary control. However, the realisation that my not being able to make the jump is beyond my voluntary control would not cause any one to retract the statement that I fail this Olympic standard. It is perfectly reasonable to be able to say someone can fail a standard without this needing to imply that such a failure is within their voluntary control. And since moral requirements are standards, we can say the same thing. Thus it is possible to say of somebody's immoral action both that it was rational (supposing it satisfied an element of the subjective motivational set), but that it also constituted moral failure. Thus internalism does not imply that people's actions are not morally criticisable – a result, I believe, the internalist ought to want.

We do not rationally criticise inanimate objects because we believe (quite sensibly) that it does not make sense to rationally criticise something for facts about itself that it cannot be responsible for. Similarly, there are certain things that we do not criticise persons for. We do not rationally criticise a person for not being taller than they are, or for having been born male or female – these are facts about themselves over which they simply do not have control. Similarly, the facts about the ways in which our motivational sets respond (or would respond) to deliberation, are not facts about ourselves that we can control. Thus, one cannot be held rationally criticisable if one's motivational set fails to change in a certain way. Reasons must be the kinds of things that are *relative to* the motivational set, and not the kinds of things which criticise it.

3. Externalist Arguments

3.1 Two arguments from Shafer-Landau

Nevertheless, externalists insist that reasons externalism has some intuitive pull. I shall survey what I think are some of the most important arguments externalists have offered, but conclude that none of them make a convincing enough case for reasons externalism. The first two arguments I will survey come from Russ Shafer-Landau (2003).

Shafer-Landau's first argument asks us to picture a truly miserable person: she is shut off from the world, she has little excitement, companionship, or pleasures in her life. It is reasonably plausible to think that she has reason to change her life: that she has reason to seek uplifting interaction with others, that she ought to get out more and see what life has to offer: if she did, she would experience more pleasures of a variety of kinds. Surely this is enough to say that she has reason to 'get up and get out'. And yet, nothing in her current motivational set, shaped as it is by her depressed state, would move her to get out. Shafer-Landau asks us to even imagine that she knows all the facts: that she can adequately imagine the pleasures that would be hers if she only sought them, but, because of her depressed state, would not seek them (ibid: 185-6)

I agree, with Shafer-Landau, that this agent probably has a reason to 'get up and get out there'. But I deny that the agent in this scenario really has properly subjected her 'depressive' motives to stay put to the deliberative process that, according to the internalist, would constitute its being a reason. According to an account of pleasure that I will give in chapter 4, something is a pleasure for someone if, by definition it would, generate desires (for, as we will see in chapter 4, there are serious problems with defining pleasure in other ways). If the agent really *knew* what these pleasures were like, if she really *had* imagined them adequately, then her motives to stay put *would* dissolve, and she would acquire a motive to 'get up and get out there'. If the knowledge of these 'pleasures' did *not* have an effect on her, we could not properly call them pleasures for her. Importantly, it is not enough merely that she is aware that certain activities are pleasurable for other people: she must be acquainted with them herself, and be bringing to her mind the full awareness of what the pleasures are like. If, because of her depressed state, she cannot do this, this simply entails that she is too depressed to undertake part of the rational procedures Williams says that her motives must be subjected to. Thus this scenario does not provide a true counter-example to the idea that one's reasons are determined by the motives one would have after having gone through the kind of deliberative process that Williams describes: the depressive state of mind itself is a barrier to going through such a deliberative process.

Shafer-Landau's second argument asks us to consider a thoroughly misanthropic person. Such a person is motivated to engage in 'the worst kind of horrors' (ibid: 187). We could imagine that such a person has subjected their motives to the kind of deliberative process Williams describes, and yet retains their evil motives. Now, Shafer-Landau claims, we would want to hold that such an agent is *blameworthy* or *deserving of punishment* for their evil deeds. But to hold them as blameworthy, we would need to suppose that they failed to respond to a normative practical reason that they had. Surely, Shafer-Landau thinks, we would hold them blameworthy. So we must suppose that there was a reason, *despite their well informed and clearly deliberated on motives*, for their having acted otherwise. In other words, we must assume that there was an external reason for their having acted otherwise.

I agree that we would hold the agent blameworthy and deserving of punishment. But the vital question is: blameworthy and deserving of punishment *in what sense*? Surely, the sense in which the person is blameworthy is, *morally* blameworthy. That is to say, the agent *fails certain moral standards*. But, unless we are assuming that moral judgments are equivalent to judgments about normative practical reasons (something Shafer-Landau himself denies), then there is no reason to think that the failure to conform to a moral standard is equivalent to the failure to respond to a normative practical reason.

But perhaps Shafer-Landau's concern can be pressed: if failure to conform to moral standards is not equivalent to failure to respond to reasons, then are we really justified in our practice of punishing or blaming those who flout moral standards? Once again, there seem to be two senses in which we might be justified in blaming or punishing those who flout moral standards: justified in a moral sense, or justified in a rational sense. I think it is easy to see that we are justified in our practice of blaming and punishing moral wrongdoers in both senses, and that reasons internalism does not threaten the morality or the rationality of such practices, which would be an unwanted result indeed. For punishment or blame to be morally justified, according to the standard-based view I endorse, is just for it to be the case that punishing and blaming moral wrongdoers is in fact a standard of morality. And, although I am yet to give my account in the following chapters of what makes something a moral standard, or does not, there seems to be no reason, on the face of it, to think that the practice of blaming and punishing moral wrongdoers would not be true moral standards. Intuitively, there seems to be no barrier to saying this. If Shafer-Landau's criticism is that internalism implies that agents are not morally criticisable (and therefore not punishable, even by moral lights), we have already seen, at the end of 2.4, that this is not the case.

Would we be *rationally* justified in punishing such a wrongdoer? If the 'we' in this question is made up of people who desire to see the standards of morality kept intact for the benefit of ourselves and society, then I think we can perfectly well say that we are rationally justified, on a fully internalist picture, in punishing our wrongdoer. Both our desire for safety from this person, and our empathic outrage on the behalf of his victims may generate motives to punish him that could certainly well survive the process of rational deliberation. What about the practice of blaming him? This too, seems rationally justified. For he is clearly responsible for his actions: as a free agent, he could have done otherwise, so the responsibility of his action falls squarely on him.

In short, the idea that the practice of moral punishment and blame involves a commitment to external reasons is, I think, simply false. It is also false that the idea of *failing to meet a particular standard* involves a failure to do what there is external reason to do. Reasons are different from standards. It is possible to have a standard present without having a reason present.

3.2 Cuneo's Normative Web

A most interesting defence of reasons externalism is to be found in the work of Terrence Cuneo (2007).

Cuneo argues for a rationalist version of moral realism. He identifies the externally reason-giving nature of moral facts[27] as something that anti-realists identify as a 'problematic' feature about moral facts (2007: 8). However, he argues, epistemic facts have this same feature: if it is an epistemic fact that belief X is warranted, this gives one an external reason for having belief X (ibid: 62). If the anti-realist is going to discount moral facts on the basis that they are externally reason-giving, then the anti-realist must also discount epistemic facts. For it would seem just odd to say of epistemic reasons that they are not external. Suppose I have been given good evidence to believe that Mount Everest is the tallest

[27] Once again, 'external reasons' is not among Cuneo's terminology, but it is clear that external reasons are what he has in mind when he discusses the 'categorical reasons' that moral facts give as not depending on any desires, cares or interests of an agent (see in particular p 6, p 38-9)

mountain in the world, but, (out of a stupid nationalistic pride) persist in my false claim that Mt. Kosciuszko is the tallest mountain. One would certainly want to claim that there is an external reason for me to believe that Mt. Everest is the tallest mountain: a belief that has nothing to do with my motives, and everything to do with the facts. If external reasons exist in the epistemic realm, why not also in the rest of the practical realm?

One way that one might respond to Cuneo is to argue that, although epistemic reasons are 'external' in a sense, practical reasons are not – and thus the parallel that Cuneo is attempting to draw between epistemic and practical reasons cannot be drawn. But I think a much better argument can be made. That is that there is a parallel between epistemic and practical reasons, that epistemic reasons are a particular species of practical reason (given that they concern a subset of what we ought to do; namely, what we ought to believe), but that such a parallel supports the idea that epistemic reasons are, after all, *internal* reasons.

Epistemic reasons are reasons that count in favour of beliefs, and epistemic reasons count in favour of beliefs in virtue of a belief's propensity to represent reality. These are two platitudes that Cuneo grants about beliefs (ibid: 53-61) that I am prepared to accept also. Cuneo concludes from this that there must be external reasons to believe what is true, or have beliefs that are true. But I think this conclusion is unwarranted, and that an alternative conclusion compatible with reasons internalism is available. This is that we have *internal* reasons to believe what is true, and perhaps necessarily so, because *beliefs* have, by definition, a direction of mind-to-world fit. In other words, the role we intend our beliefs to play is one of representing the world *truthfully* to us. Because, in forming beliefs, we have interests in representing the truth to ourselves (as this is what we take beliefs to be *for*) then the reasons we have to believe things that are true are *internal* reasons: our beliefs being true serves an interest of ours, an interest we have when forming our beliefs. Thus we have epistemic reasons to believe what is true not because there is some external requirement on us to represent the truth, but because representing the truth is what beliefs, properly understood, are *what we are interested in doing when we form them*.

Cuneo of course disagrees with this:

> 'Suppose, for example, I were to reproach you for having ignored considerations at your disposal that indicate that p is true. Suppose you were to reply by saying that, since you don't care about believing

> whether p is true, these considerations don't give you any type of reason to believe that p is true. Strange response! In so replying, we would rightly wonder whether you were being sincere, intent on provoking, or worse, in a lucid state of mind.' Ibid: 61

Cuneo is right in saying that such a response in light of being shown the likely falsity of one's beliefs would be strange. But, according to what I have argued, he is right for the wrong reasons. The reason for criticising a person for not caring about whether her beliefs are false lies not in the fact that she has ignored some external requirement, but that she is in a sense contradicting herself. By engaging in the very practice of *having* beliefs, she has, if she is not conceptually confused about 'beliefs', committed herself to representing the world to herself. It is not possible then to both want to do this and not care about whether one's beliefs really *do* represent the world. Such a mistake is akin to willing an end without willing its means.

What, then, is the appropriate response to my stubborn persistence to believe that Mt. Kosciuszko is the tallest mountain? If I have been given evidence of the falsity of my belief, but persist in my wilful blindness, it is questionable whether my 'belief' really is a proper belief: whether it really is a mental state with a mind-to-world direction of fit. A sensible thing for me to say is that I *want* it to be the case that Mt. Kosciusko is the tallest mountain in the world – I *want* to believe something that has shown to be false, and this explains my stubbornness. But my wanting to believe something is a mental state with a world-to-mind direction of fit, a desire-like state. Accordingly, the standards for deeming a *belief* rational or irrational, technically speaking, no longer applies to it. We might think the desire that Mt. Kosciusko be the tallest mountain in the world is a silly or irrational desire. But this will simply mean, according to internalism, that it is a desire that would not survive the process of deliberation outlined by Williams.

Epistemic reasons, then, do not seem to presuppose any commitment to reasons externalism. Internalism can explain the rationality of beliefs just as well. What it is to have a belief is to have a mental state with a mind-to-world direction of fit. Thus, in forming a belief, our end or interest is in representing the truth to ourselves. Such an end or interest is a motive of ours, relative to our motivational set.

3.3 Smith's objections to agent-relativity

Smith takes the fact that rational agents could, (according to his view) regardless of their starting motivations, come to the same conclusions about what is rational, as evidence against the agent-relativity of reasons (more precisely, the relativity of reasons to motivational sets). But this is no evidence against agent relativity at all. Nothing in Williams' theory precludes the possibility that rational agents with different motivational sets could come to the same conclusions about what they have reason to do, given that deliberative processes themselves have the potential to change motivational sets. It may well be the case that all rational agents might converge on what they conclude are reasons, regardless of what their motivational sets are like to start with. But all that this might prove is that it is a contingent fact about deliberative processes that they have similar impacts on the motivational sets of human beings as they reason. If this is the case, then it is not the case that some reasons are, as Smith says 'agent-neutral' , but rather, that they are agent-relative despite the fact that all agents happen to reach similar conclusions. In order for reasons to be agent-neutral, it would have to be the case that deliberative processes have similar effects on the motivational sets of all agents *as a necessary fact*: which is what Smith seems to suggest in saying that the *concept* of a reason does not allow for agent relativity, and involves convergence.

Smith's other claim, that reasons-neutrality is a result that we ought to *want* from a good analysis of normative reasons too, is unsubstantiated. In 'Internal Reasons', Smith makes the following argument:

> The point is important, for it suggests that when we talk about reasons for action we quite generally take ourselves to be talking about a common subject matter: reasons *period*. We are thus potentially in agreement or disagreement with each other about what constitutes a reason and what doesn't. This is why, when we find ourselves in disagreement… We always have the option of engaging in argument in the attempt to find out who is right and who is wrong. Other people's opinions about the reasons that there are thus constitute potential challenges to my own opinions. I have something to learn about myself and my own assessment of the reasons that there are by finding out about others and their assessment… All of this is flat out inconsistent with the claim that our concept of a reason for action is quite generally relative to the individual; that it typically means reason$_{me}$ out of my mouth, reason$_{you}$ out of yours… It suggests rather that our concept of a reason is stubbornly non-relative. 1995: 123-4

This passage captures what seem to be Smith's major motivations for rejecting agent relativity. But these motivations are ill-founded. The relativity of reasons, as Williams cashes it out, has none of the implications that Smith seems to think it does. Smith supposes that agent-relativity of reasons implies that whenever we are talking about reasons, we are not talking about the same 'subject matter' – such that whenever you and I talk about reasons we will be talking past each other, so that nothing I might say about what your reasons are could have any relevance to you. But this is incorrect. According to agent relativity, in Williams's sense, there *is* a 'shared subject' in discussions about reasons. The shared subject is yours and my motivational set. This may indeed yield the conclusion, quite often, that your reasons for action and my reasons for action differ. But it *does not* entail that I am talking past you or that you are talking past me when we make reasons claims about each other. You can be making a claim about what I have reason to do by saying an element of my S would be satisfied by my ϕ-ing. And I might disagree with you about this; but this is still a *genuine* disagreement, and not a mere talking past one another. Furthermore, the idea that I could tell you about what your reasons are (and you could tell me about what my reasons are) is not precluded by Williams' theory either. And this is the case even when our reasons do differ. Agents, according to Williams, can be mistaken about what their own reasons are, by failing to understand how their motivational set would be satisfied. If I truthfully say that you have a reason to ϕ, because an element of your motivational set would be satisfied by ϕing and you disagree, then, according to Williams' theory, I am right about your reasons and you are wrong about your reasons. This is not because I take ϕing to be a reason for *me* (as Smith's criticism implies), but because there are facts about you that make it a reason for you. In short, Smith seems to have misunderstood the sense in which reasons are 'relative' in Williams. Reasons are not relative to different actual agents own *opinions* about their reasons, but relative to different motivational sets of agents.

Nor does agent-relativity entail (as we have already seen) that there are reasons for action that you and I could not both share. Williams seems to be concerned that an analysis of reasons should yield the conclusion that there are reasons that we all turn out to share. And this may be a fair concern. But an analysis according to which reasons are agent-relative in Williams' sense *can* do precisely this. It may well be the case that all rational agents could have reasons for action that they all share, due either to contingent similarities in motivational sets at the outset of deliberation, or to contingent similarities in the outcome that rational deliberation has on those motivational sets. Agent relativity does not prevent some reasons from being universal.[28]

3.4 Conclusion

In conclusion, we have seen that reasons internalism gives us everything we could want from a theory of practical reasons. It is a beautifully unified theory in that the same reasons why actions are reason-giving is the very same reason we are motivated to act in accordance with our reasons-judgments: namely because our reasons are motives that we regard to have survived a particular deliberative procedure. Reasons internalism allows us the possibility of demonstrating, very genuinely (that is, without definitional fiat) the possibility that certain normative standards (including moral standards) are justified. Reasons internalism allows for the possibility that there are reasons we all share, allows that we are rationally justified in our moral practices in blaming and punishing, and allows that we always have reason to believe what is true. None of the strength reasons externalism claims to have are exclusive to it, and many of the weaknesses that reasons externalism suffers from do not assail reasons internalism. I conclude, then, that reasons internalism is the more plausible theory of practical reason.

If this is the case, the threat of moral error theory looms. The claim that external reasons do not exist was a key premise of the error theory argument surveyed in the previous chapter. If moral success theory is to be vindicated, then, the other major premise: the rationalist conceptual claim, must be convincingly argued against. I intend to do this in chapter 3. In the following chapter I will give an account of moral judgments that will both pave the way for an account of moral facts, and exclude any commitment to the existence of external reasons.

[28] As for Smith's claim that the very concept of normative reasons requires agent-neutrality, Sobel has an effective reply 1999:44-5. At least some of our reasons, such as our reasons grounded in tastes, such as preferences for icecream flavours or music styles, are agent relative, though they are still normative. If we are right in thinking that *some* reasons are agent relative, then there is nothing incoherent about them *all* being relative: thus agent-neutrality cannot be a conceptually non-negotiable feature of normative reasons.

3: Moral Judgments

1. A Content-Based Account of Moral Judgments

1.1 The Content of Moral Judgments

We understand that there are differences between moral claims and other kinds of normative claims. We typically recognise the claim ‘you ought not chew with your mouth open’ as a claim about what etiquette requires of us, rather than as a claim about what morality requires of us. Similarly, we recognise the claim ‘you ought not move your rook diagonally’ as a claim about what the rules of chess requires of us, rather than a claim about what morality or etiquette requires. And we recognise the claim ‘it is wrong to steal’ as a typical moral claim – a claim about what morality requires of us. The very fact that speakers recognise these differences between types of normative judgments brings me back to a point made in chapter 1: that normative judgments are propositions about systems of standards: there are different *types* of normative judgments because there are different systems of norms, different systems of standards, to which they belong.

It is clear that at least some component of what we understand moral judgments to be has to do with content. In order to understand the meaning of moral terms, in order to be a competent user of moral discourse, it must be the case that one is able to distinguish and differentiate moral judgments from other types of normative judgments, such as etiquette judgments, or fashion judgments. What is it about a normative judgment that makes it a *moral* judgment? What we must now ask is: what are some basic things that all competent makers of moral judgments assume about moral norms? In what way do competent users of moral concepts understand moral norms as being different from other norms?

It should be obvious from the start that the rationalist conceptual claim – the idea that moral judgments are judgments about reasons – is unhelpful for distinguishing moral judgments from other kinds of judgment. At best, the rationalist conceptual claim is less than sufficient for making sense of the difference between moral judgments and other kinds of judgments, at worst, unnecessary. Why is the rationalist conceptual claim insufficient for making sense of the differences between moral judgments and other kinds of judgments? It is simply because the rationalist conceptual claim says nothing about content-based parameters of moral judgments. Suppose I believe it is morally right that I keep my promise. According to the rationalist conceptual claim, this will involve the belief that I have a reason to keep my promise. But my belief must amount to *more* than this. For I may also have the belief that I have reasons that are not paradigmatically *moral* reasons: I may believe I have reasons to wear pink shirts, to see the Wallabies play, or to brush my teeth. But none of these are *moral* reasons. The claim that moral judgments involve making judgments about reasons (even if it were true) would need supplementation: specifically, one must believe that one has a *moral* reason. One must believe that a paradigmatically *moral* consideration counts as a reason. And it takes an understanding of how the *content* of moral judgments differ from the content of other kinds of normative judgments to cash out just what is meant by a *moral* consideration, as opposed to a prudential consideration, an etiquette consideration, or any other kind of consideration. Even Smith, an ardent rationalist, makes this point in *The Moral Problem*, and admits that what makes a *moral* reason different from a *prudential* reason has something to do with its content (1994: 183-4).

I will, of course, end up arguing that the rationalist conceptual claim is altogether false. It is not necessary, when one makes a moral judgment, to make the judgment that a consideration of a certain type (what I will refer to as a paradigmatically moral consideration) constitutes a reason for action. All that is required, when one makes a moral judgment, is that one makes the judgment that the consideration entails that performing (or refraining from performing, as the case may be) the action is a standard of morality. Suppose I believe that giving to charity is morally right, because it is altruistic. According to the rationalist conceptual claim, I am judging that the consideration (the fact about the action) 'it is altruistic' gives me a reason to give to charity. On my account, however, all that need be being believed for the moral judgment to be genuine is the belief that the consideration 'it is altruistic' is what makes giving to charity morally required, or, in other words, a moral standard.

But this will all be spelled out in good time. For now, we must begin the discussion by asking the question of just what kind of considerations are moral considerations. When we make moral judgments, according to my account, what kind of feature must we be judging the thing we are morally evaluating to possess? The answer to this question will help us make sense of the idea that morality, and moral judgments are delimited (at least partly) by their content. We know that the norms of etiquette are, roughly, concerned with the governing of social interactions in certain settings. Etiquette concerns things like manners, rituals, what is appropriate to say in a given context and the like. We know that the rules of chess are rules of a certain game, played on a board, and that the rules are in place to constitute a game in which strategic, probabilistic, and geometric thinking must be utilised in order to win. What, then, can we say, the norms of morality are 'about'? What are the 'concerns' of the norms that we take to be *moral* norms? What are we making judgments about when we make moral judgments?

Initially, it may be asked whether there even is such a thing as norms taken to be *the* norms of morality, and it may be doubted that there any *one* concept of morality, given that there appear to be many and varied systems of morality over the world, and throughout different times and places. But this question (which may arise either from particularist or relativist instincts) of whether there is such a thing as 'one true morality' does not undermine our purpose here of saying that morality as a whole is one distinct concept. There are many different 'moralities' or moral systems around the world, for sure. But the point is that we recognise these normative systems all as *moral* systems, and the judgments accepted and passed on within each system, as *moral* judgments (whether or not our own moral judgments diverge).[29] Our task here is to elucidate what makes all these differing systems, systems of *morality* – what is it that they all have in common such that we recognise them as *moral* systems, and the norms within them as containing *moral* norms?

A common hypothesis, and one that I will endorse, is that morality is a system of norms concerned with *welfare*. The idea that, at a pretheoretical level, the content of moral norms has fundamentally to do with the well-being of others is found throughout numerous philosophical writings, of which the following are just a few:

> We should first distinguish moral "oughts" from other types of "oughts." Some different normative concepts are associated with prudence, some with rationality, and some with aesthetic norms. Moral norms primarily concern our interactions with others in ways that have significance to their well-being. Thus while it is true that we *ought* to eat at least five servings of fruits and vegetables a day, this ought is not a moral one. If we fail to do this, we have harmed only ourselves – so it is a failure of prudence, not of morality. Also, one *ought* not to hang a psychedelic black velvet painting over one's colonial fireplace. However, doing so is not a moral failure. If anything, it is an aesthetic failure. But if we do something that could harm or benefit someone else, then arguably this is a *moral* matter. Someone who wrongfully harms another does something that he or she ought not do in the moral sense of "ought." Driver, 2006: 1

> As these questions suggest, from among the diverse meanings of 'morality' and 'moral' a certain core meaning may be elicited. According to this, a morality is a set of categorically obligatory requirements for action that are addressed at least in part to every actual or prospective agent, *and that are concerned with furthering the interests, especially the most important interests, of persons or recipients other than or in addition to the agent or speaker*. Gewirth, 1978: 1 (emphasis added).

> In the last chapter, I suggested that there is a commonsensical conception of moral facts, fundamental to which are platitudes of two kinds. In the first place, there are what I called the 'content' platitudes. These platitudes tell us that there are intimate relations that obtain between moral facts and human well-being. For example, according to the content platitudes, actions of a certain range generally display moral properties inasmuch as they foster or undercut (or express intentions to foster or undercut) human flourishing or the participation in the states and activities that are components thereof. Cuneo, 2007: 52

> What is striking about this literature is that, from a young age, children distinguish the moral violations from the conventional violations on a number of dimensions… the explanations for why moral transgressions are wrong are given in terms of fairness and harm to victims. For example,

[29] This point is also made well by G.A. Warnock, 1971, ch 1.

> children will say that pulling hair is wrong because it hurts the person. By contrast, the explanation for why conventional transgressions are wrong is given in terms of social acceptability… Nichols 2004: 6

Another hypothesis, supported by Jonathan Haidt and Craig Joseph (2004) is that there are a certain set of hard-wired intuitions that we recognise as 'moral' intuitions along the lines roughly given by the above quotes. Such intuitions arise from evolved emotional reactions, such as empathy and compassion toward those who are suffering, and desire for reciprocity and fairness (and anger at when it is flouted). Although other emotional reactions are cited (for example, the emotion of disgust), which gives rise to as giving rise to moral intuitions (such as intuitions and norms revolving around the notion of sexual purity for example), the pair of emotional reactions most cited by authors on this topic are the emotions involving compassion and fairness (ibid: 57-9) Such hard wired human emotional reactions give rise to intuitions about rightness and wrongness that are embodied in just about every moral system around the world: that harm toward others is undesirable, to be minimized, and should not be inflicted unless there is some outweighing good to be had; that resources should not be distributed any old way, but should be based on principles of desert or need; that certain virtues such as kindness, temperance, and courage should be cultivated. So, although moral disagreement is notable, there can be found to be much moral *agreement* underlying, and even explaining, such differences. Turning our mind to a concrete example will help to illustrate this point. In ancient Eskimo tribes, it was common practise for elderly members to euthanize themselves, so that the delicate amount of resources the tribes had in their harsh living environment could be spent on younger members of the tribe who had more of a future ahead of them, and a chance to keep the tribe going (this example is taken from Rachels, 1996 [1986]) In Middle Eastern cultures, on the other hand, the elderly live with their extended family right up until their death, and often remain the heads of the family until then, to provide order, wisdom and stability. Although these cultures diverge drastically in their practices and views on how the elderly ought to be treated, we can see shared moral intuitions pertaining to harm minimization, and to reciprocity, behind them. In the Eskimo culture, the elderly are dispensed with so that harm to the rest of the tribe may be avoided. In Middle Eastern culture, the elderly lead households until their death for the very same reason – so that harm, which might arise from a lack of wisdom in the younger members, might not befall the family. Similarly, we see intuitions of fairness at work as well. The elderly in the Eskimo culture choose to die because they have lived a long life already – it is only fair that younger members get the chance to do the same. The elderly in the Middle Eastern culture deserve respect, deference and long term care because this is a fair exchange for the wisdom they have to offer their family. This 'welfarist' core in the divergent moral systems, according to my contention, is what makes both systems recognisably *moral* systems.

Other evidence, particularly for the involvement of the emotion of empathy in the making of moral judgment, is cited by Shaun Nichols. Roughly, Nichols' the idea is as follows: that moral judgments are constituted by judgments about rules prohibiting and permitting actions that, respectively, are likely to cause negative and positive affect in observers as a result of perceived suffering or benefit in others (2004: 18). In other words, Nichols' 'sentimental rules' account characterises moral judgments (and accordingly, under my definition, moral claims too) in terms of content – in terms of commending beneficial actions and condemning harmful actions.

It is worth going into a little more detail about Nichols' theory of moral judgments, as it raises important questions for our claim that morality is delimited strictly by content. According to Nichols' account, moral judgments have two components. The first component involves the agent distinguishing between right and wrong actions on welfarist considerations, or the propensity to cause harm. As Nichols recalls the empirical research on people's ability to distinguish between moral and conventional violations, he attributes to them 'an internally represented body of information, a "normative theory" prohibiting behaviour that harms others…' (2004: 16). In my terminology from chapter 1, what is happening is that agents are making standard-based judgments: judging as 'correct' or 'incorrect' actions with certain features, and these features have to do with the propensity to harm. But Nichols also cites a second feature of moral judgment: an affective mechanism that is implicated in moral judgment. We learn to take notice of, and recognise harm-causing actions due to the fact that we experience a degree of negative affect when they are perceived (2004: 16-18). Judgments about moral violations, then, are judgments about standards which, if flouted, are likely to cause negative affect.

This raises a question about the claim that moral judgment is strictly delimited by content, for it might be plausible to think, on the basis of Nichols' claims, that there is one formal characteristic of moral judgments: an experience of (perhaps inherently motivating) positive or negative affect. Such a presence of positive or negative affect when moral violations are perceived may support the rationalist contention that judgments about reasons are present when moral judgments are made, after all (the fact that we are judging that there is a *reason* to refrain from harm may be touted as explanation of the presence of negative affect when harmful actions are perceived). However, I will argue essentially that only the first component of Nichols' account of moral judgment is essential to the concept of a moral judgment. It is worth noting that Nichols himself may have no quarrel with this,[30] - he intends his account to be an *empirical* claim about moral judgments, and believes that the second element of his account of moral judgments (the presence of affect) is *conceptually* dissociable from one's capacity to have an internalised normative system (ibid: 19) As I develop my account of moral judgments, this is a claim I will come back to: not only that positive and negative affect are conceptually dissociable from standard-based judgments about harm, but that other formal elements, such as judgments involving commitment to any kind of reason, are dissociable also.

In section 1.3, I will attempt to give some more clarity to the ideas already gestured at by the philosophers surveyed in the excerpts above: the idea that moral judgments are importantly related to judgments about welfare. But before I do, it is worth briefly dismissing a relativist idea about the content of morality, which was briefly surveyed in chapter 2.

[30] Nichols stresses in a footnote in Sentimental Rules, that he is not making a conceptual claim about what moral judgments involve, but an *empirical* claim – this actually entails that Nichols and I have no conflict. In fact Nichols is agreement with me when it comes to the conceptual claim: 'I am not suggesting that it is a conceptual truth that psychopaths fail to make moral judgments. A great deal of work in metaethics focuses on the idea that it is conceptually possible that someone might have mastery of moral concepts without having any concomitant motivation to act morally or any concern for others. I think this is right and of some significance for evaluating certain claims about internalism.' (ibid: 19).

1.2 Lest we change the subject

In chapter 2, I mentioned two senses in which the term ‘morality’ can be used. Two senses which, properly understood, denote two very different concepts. The concept that I referred to as ‘morality proper’ in chapter two is the normative system I have just been discussing. It is the normative system of standards concerned somehow with welfare. Accordingly, moral judgments (judgments about ‘morality proper’) are distinguished from other kinds of normative judgments by virtue of their content along the lines given by philosophers like Driver and Nichols, which I hope to clarify further in the next subsection. According to another concept to which the term ‘morality’ is confusedly (and unfortunately) applied, is the concept of one’s personal ideals or values, one’s ‘morality’. Of course, the content of one’s own personal ‘morality’ could be anything. Williams’ figure of Gaugin (1981:22-23) is an example of one whose ‘personal morality’ has nothing to do with the concerns of morality proper, but with his own artistic pursuits. Such a concept of ‘morality’ lends itself to Harman’s relativist claim: that morality is relative because one’s reasons, or one’s personal values, are relative (1975).

There are two ways of describing how this ‘personal morality’ concept operates in moral judgments. One suggestion is a subjectivist or relativist one: that when we make judgments we are making judgments about what our personal ideals are, or perhaps our reasons (which, if we, following Harman, believe the internalist interpretation of such judgments to be the only sensible one, will commit us to a moral relativity entailed by the agent-relativity of reasons). To be clear from the outset, this ‘personal morality’ sense of morality is *not* what is being discussed in this chapter. It may be asked what, exactly, is wrong with this relativist or subjectivist picture of morality. My best answer would perhaps be: ‘nothing, except that this concept of morality is not the one that most of us are really interested in when we engage in moral discussion.’ When we debate the rightness or wrongness of abortion, war, giving to charity, or killing an innocent person in order to promote good consequences, we are clearly involved in discussing ‘morality proper’, rather than reporting what our personal values or reasons are. We are interested in discussing whether or not any of these acts have the kind of feature or characteristic that would make them morally required. This claim of mine can be supplemented by a common criticism faced by subjectivist and relativist views, namely that we have *genuine disagreements* when discussing moral matters, in a way we would not have were we simply using the term 'morally right' as a place-holder for one's personal ideals, or whatever one took to be internally reason giving. When I say that war is morally justifiable in some circumstances, and you say it is not, we are *disagreeing*. On the subjectivist 'personal ideals' view of morality, of course, there is no disagreement: you are simply saying that war is not according to your ideals, and I am saying that war sometimes is according to my ideals. But the fact that we take moral disagreements to be *genuine* is evidence that it is morality proper, morality-the-system-of-standards-somehow-concerned-with-welfare, that we are discussing. More precisely, on my account, we are disagreeing about which prohibitions or requirements are standards that belong to this system of morality, or do not.

There is a second way we could understand the 'personal ideals' concept of morality as being embodied in moral judgments, on a non-cognitive understanding. According to such views, although we do not *report* our personal ideals when making moral judgments, we nevertheless *evince what our personal ideals are* when we make moral judgments, giving voice to desire-like states about what we ourselves and others should do.[31] What is wrong, it might be asked, with taking this non-cognitivist idea on board but adopting an ecumenical view of moral judgments: that there are two elements to our judgment, both a cognitive and a non-cognitive element, which employ the two senses of 'morality' respectively? On this view, the cognitive element is our making judgments about which standards are standards of morality, on the basis of welfarist considerations. And the second, non-cognitive component is our expressing our personal ideals.[32]

My response, essentially, is that such an ecumenical view asks moral judgments to be too much. Firstly, there is no conceptually necessary connection between the two features: I could well make the cognitive judgment that the requirement to ϕ is a standard of morality on the basis of welfarist considerations, without ϕing being among those personal ideals to

[31] There are a variety of forms these desire-like states could take, and I use the term 'non-cognitive' with some hesitation, given that some expressivists (see Horgan and Timmons, 2006a, 2006b) prefer not to use the label. For my purposes, however, I will be content to describe non-cognitivism about moral judgments to include the emotivism of Stevenson (1937), the prescriptivism of Hare (1964), and the expressivism of Gibbard (1990), Horgan and Timmons (2006a, 2006b), as well as the quasi-realism of Blackburn (1993). All such views have in common that they can be taken to be examples of views which hold that a speaker's moral judgments indicate or evince what some of her personal ideals are, whether they take the form of pro-attitudes, prescriptions, or other-desire-like states. All such views, then, are committed to the concept of morality-as-personal-ideals.

[32] A prime example of such a hybrid view is defended by Copp, 2001.

which I express any allegiance. For the purposes of the coming discussion, I will regard only the first, cognitive feature of moral judgments to be *essential*, and the second feature only contingently present in many moral judgment makers. One may want to reply that genuine moral judgments must meet both conditions (of being judgments about standards, and expressions of one's personal ideals) even though they are not connected by necessity. But, as I will argue further on, such a move is unjustified and potentially *ad hoc*.

Lest we become confused, or change the subject, the reader must keep in mind that it is 'morality proper' that I have in mind in the discussion of subsection 1.3. When I give the account of moral judgments throughout the next section, I simply intend to be giving an account of what judging a requirement (or recommendation, or prohibition) to be a standard of morality involves.

1.3 My content based account of moral judgments

I have said that morality is chiefly concerned about *welfare*. The tightest formulation of this idea has been that given by Nichols: moral judgments are made from an internalised normative system which prohibits harmful actions (and supposedly praises beneficial actions as well). In this section, I will give this idea even more precise formulation. Although I will be defending the idea that moral judgments must be principally concerned with welfare, one should keep in mind that I do not intend this to be a narrow consequentialist thesis. We make often moral judgments about actions which can go against consequentialist intuitions (judgments which, for example, respect certain deontological side constraints), but which do not fail to be *moral* judgments for that. So although morality's cluster of concerns has principally to do with welfare, there are a number of other moral concepts besides consequentialist ones that I think can said to be 'welfarist'. Morality can be said to be a system of norms that commends people acting in ways that would ensure reciprocity, and fairness, or evince compassion or good will toward each other, as well as being concerned with the *promotion* of interests.[33]

Also of interest in this section is the task of distinguishing moral judgments from prudential judgments. We have said that moral judgments are about welfare. But then again, so are prudential judgments. As I present my account of moral judgments, I will comment on how

[33] For a discussion of the potential breadth of application of the term 'welfarist', see Keller, 2009

prudential judgments differ from them. Basically, what I will say will support Driver's contention above that morality is a system of norms which concerns how one ought to treat other people in such a way that serves their interests, while prudence is a system of norms pertaining to how one serves one's *own* interests.

To get our account going, let's recall some of the intuitive differences between moral judgments and normative judgments of other kinds, highlighted by the excerpts from 1.1. Often, when we express moral judgments, people can initially mistake our meaning, thinking that we are conveying some other kind of normative judgment. When this happens, we qualify our statement by citing considerations of a certain type, in order to make our meaning clear. Take the following example. Suppose I say to my friend John, 'you ought not cheat Mike out of his money'. Let's imagine, for the purposes of this example, that John is a sociopath who doesn't always find it easy understanding moral concepts, or when people are making moral judgments. Accordingly, John answers 'I see what you mean. It would be wrong to cheat Mike out of his money because it will come back to bite me.' John has mistaken my moral judgment for a prudential judgment. So I reply: 'no, what I mean is that you *morally* ought not to cheat Mike out of his money.' Given his sociopathy, John answers: 'What do you mean?' Now, at this point, what am I most likely to say? I am probably most likely to say something like: 'I mean you ought not cheat Mike out of his money *because it is not fair* to him! You ought not cheat Mike out of his money because *it hurts him*! That is what's wrong with it!' In other words, in order to make my use of moral concepts clear to John, I point out certain types features as the act as those which confer the moral status I am attributing to the act – features to do with fairness, goodwill toward others, or minimisation of harm. After having done so, John might be able to accept my statement: 'yes, I morally ought not cheat Mike out of his money' (even if, given his amoralist tendencies, John is still motivated to cheat Mike anyway. We will come to amoralist controversies shortly).

Consider another example. Imagine I invite two friends, Jenny and Sarah over for dinner, and I have my treasured collection of frozen funnel-web spiders on the mantelpiece near the dining table. Imagine Jenny arrives first, sees the spiders, and with a grimace, says 'ooh… you shouldn't have those spiders there.' Once again, it is not immediately clear what kind of normative judgment Jenny is expressing. Suppose she goes on to say: 'those spiders look a bit tacky where they are – they'd be better off in the study.' It becomes apparent that Jenny is making an aesthetic evaluation. But suppose Jenny says, 'don't you remember, Sarah's little brother died from a funnel web bite just recently? She tried to save him but couldn't get him to the hospital in time, and she's been experiencing post-traumatic stress ever since. If she saw those spiders, she'd flip out.' It is now clear that Jenny is making a moral judgment: I ought not have my spiders on display where Sarah can see them because it will cause her a great deal of damage in the form of psychological distress.

How can we formalise the intuitive distinctions drawn in the examples above? I believe the following formulation gives us a good way to do so. I call the following claim 'the moral consideration claim' (MoCC), as the name of the doctrine that moral judgments are made on the basis of paradigmatically moral (welfare-based) considerations.

> **Moral Consideration Claim (MoCC):** A makes a positive moral judgment about X iff A judges that there is a positive relation between X and the wellbeing of others. A makes a negative moral judgment about X iff A judges that there is a negative relation between X and the wellbeing of others. A makes a neutral moral judgment about X iff A judges that there is neither a positive or negative relation between X and the wellbeing of others.

The different elements of MoCC need some explaining. A is the agent making the judgment. X is the thing being morally judged, and X can plausibly include actions, people, or states of affairs. For simplicity's sake, I will be discussing MoCC in relation to *actions*, but one should keep in mind that any of the coming arguments made about moral judgments about actions can be made for moral judgments about people and states of affairs too. I have distinguished between *positive*, *negative*, and *neutral* judgments, to be as clear as possible. A positive moral judgment is a judgment about the rightness, goodness, or ought-to-be-done-ness of something. Examples include: 'it is right to keep one's promises', 'one ought to treat others as one would be treated', 'it is good to give to charity', 'it is morally obligatory to save someone's life if doing so would come at little cost to oneself,' 'Mary is a good woman' 'John's character ought to be emulated', 'a society that looks after its poor is a just society.' A negative moral judgment is a judgment about the wrongness or badness of something. Examples include: 'It is wrong to break one's promises', 'one ought not steal from one's neighbour', 'it is bad to murder somebody' 'it is morally prohibited to slander someone to get ahead at work,' 'Mary is a bad woman', 'John's character ought not be emulated', 'South Africa under apartheid was an unjust state of affairs'. A neutral moral judgment is a judgment that something is permissible; neither good nor bad, right nor wrong. Examples include: 'It is permissible to ask your neighbour if you can borrow their car', 'it is neither morally right nor wrong to eat cheese on toast.' There are more questions about the above statement, which I shall address in turn: 1) what is a positive or negative relation to wellbeing? 2) What is the difference between moral and prudential judgments? 3) What makes moral judgments true or false? 4) Do moral judgments necessarily need to be based on *welfarist* concerns? 5) What is the scope of morality's application and to whom does it apply?

What does a positive relation to wellbeing involve? What does a negative relation to wellbeing involve? These terms 'positive' and 'negative' are deliberately vague: they are supposed to be place-holders for any consideration that could plausibly be said to have anything to do with welfare at all, which people invoke or have in mind when making moral judgments. Sometimes, when people make moral judgments, they have the direct promotion of good/bad consequences in mind. So one thing that 'positive' and 'negative relations include is the direct promotion or damaging of welfare, respectively. Another thing that a 'positive' or 'negative' relation can include is the respecting or infringing of a person's rights, respectively. But regardless of this, it is plausibly a welfarist notion, as is a rule-consequentialist's idea of rules which, if followed, promote wellbeing. Another recognisably deontological principle that might be invoked when making moral judgments is that of retribution. One might judge that people who have harmed others ought to be punished because of the harm they have caused. It is thought that punishment on these grounds gives adequate weight to the harm that has been inflicted on the victim.[34] So another positive relation between an action and the wellbeing of others can be said to be the embodying of a respect toward the (damaged) wellbeing of others even though, in this case, our action consists of harming or making someone worse off (the perpetrator). Another possible positive relation an action can have to the wellbeing of others is its being

[34] Perlmutter discusses this justification for retributive punishment in saying: 'there are some crimes that are so serious, so offensive to the moral community, that their perpetrators deserve death for committing them' 1999: 123. The offensiveness, presumably, has to do with the extreme damage done to (the victim's) welfare.

motivated by a virtue in which the agent acting *intends* to help others: if an action is done out of *kindness*, *compassion*, or *self-sacrifice*. Accordingly, another negative relation between actions and the wellbeing of others are actions with motives behind them such as malice, spite, or callousness. I have tried to list a range of different considerations from consequentialism, deontology, and perhaps even virtue ethics, which people who can be said to be making moral judgments evoke as the basis for the moral rightness (if the relation is positive) or wrongness (if the relation is negative) of that which they are judging.

I have said above that 'positive' relations to wellbeing might include a number of things: the direct promoting of interests, the respecting of rights, the obeying of duties, the following of rules on which well-being of individuals within a moral community rests, and the possessing of certain virtues. It is reasonable to ask whether particular deontological concepts such as rights and duties are even welfarist concepts at all, and that a positive relation between an action and wellbeing must necessarily be a consequentialist one of promotion or probable promotion. A complete answer to this question will not be given until chapter 5, where I will identify a unifying feature that I believe all positive relations to well-being have in common. But I think I can put forward a good reason at this stage for assuming that deontological principles are welfarist (and therefore that judgments made on the basis of such deontological considerations constitute genuine moral judgments on the view I am defending). What is it about a particular being that determines whether it has rights, or whether others have duties toward it, or not? While humans and quite possibly the majority of animals have rights and are owed duties, it is far harder to say that inanimate objects like rocks, chairs and cans have rights and are owed duties, without seriously stretching the concept of a right or duty. The difference, it seems to me, is that humans and animals are the kinds of things that are capable of being well-off or not well-off, whereas the inanimate objects are not. Rights and duties, then, attach to beings *because* welfare also attaches to such beings. Take a familiar moral example where duties and rights are invoked, in order to bring this out: the scenario in which a doctor has the chance to cut up an innocent patient and farm out his organs to five needy patients. It is natural to think that it would be wrong for the doctor to do this, on the basis that a duty that doctors have to patients is being breached, or because the innocent patient's rights are being infringed – whether or not the doctor could directly promote more good consequences. Such deontological intuitions, however, would not be present if we did not regard the innocent patient in the scenario as something with welfare that could be damaged by the doctor's actions. What reason would there be to suppose the doctor had a duty to refrain from cutting up the patient if such a thing did the patient no harm? Or what rights could the patient have if it were an inanimate object with nothing at stake in the doctor's decision? There seems to be good reason, then, for regarding the upholding and infringing of rights and duties as positive and negative relations to welfare, respectively. There is good reason to deny that positive relations to wellbeing are equivalent to consequentialist verdicts. It is possible to make a moral judgment on the basis of considerations to do with rights and duties, and for such a judgment to be a moral judgment according to the parameters set out by MoCC.

Everything I have listed as being potential 'positive' relation between an action and wellbeing have certain intuitive similarities with each other. The same goes for the negative relations. But the reader might yet be unsatisfied with this. The reader might crave something that can be said to be a *unifying* feature by virtue of which all putative positive relations are positive, and by virtue of which all putative negative relations are be negative. Later on, in chapter 5, I will argue that the unifying feature all positive relations between an action and the well-being of others is that a particular kind of ideal observer would have a particular kind of response to them (the same goes for negative relations). Accordingly, I will argue in chapter 5 that the ideal observer's responses serve as the *truth-makers* for moral judgments. But the ideal observer's responses is not a concept we need to employ now, to build an understanding of what moral *judgments* consist in. Despite the fact that others have advanced ideal observer theories as analyses of moral judgments (see in particular Firth 1952), this is not something I wish to do. A good theory about what makes a moral judgment must be less stringent than this, and must not require that the folk have the esoteric concept of an ideal observer in order to qualify as having moral concepts which are employed in moral judgments. In order for the man on the Clapham Omnibus to make a judgment that X is morally right, it is enough that he be either making a judgment that X promotes welfare, respects rights, or evinces a certain motivation (or all of the above at once), without having to understand that what all these 'positive relations' have in common is that they would evoke certain responses from a certain ideal observer. By analogy, one could say that it is enough for a three year old child to be making a judgment that X is water so long as they are making a judgment about the stuff that comes out of the tap, or which falls from the sky, or which fills the rivers and lakes, without having to understand that what all this 'watery stuff' has in common is that it is H2O.

I now turn to the second question: what is the difference between moral and prudential judgments? We have specified a number of ways in which actions can be positively or negatively related to well-being. But prudential judgments are supposedly concerned with this as well. So what is the difference? The difference between moral and prudential judgments about actions can be spelled out as follows: when one judges an action as morally right or good, one is judging that the action has a positive relation to an interest bearer(s) *who is not the agent performing the action*. When one judges that an action is prudentially right or good, one is judging that the action has a positive relation to *the agent performing the action*. The analogous comments apply for immoral and imprudent actions and negative relations. If the judgments in question are about *persons* rather than actions, we can say a similar thing. A judgment about the moral status of a person's character will consist of a judgment about how their character impacts or consists of intentions toward others' wellbeing. A judgment about the prudential status of a person's character will consist of a judgment about how their character impacts or consists of intentions toward themselves. Whenever I mention MoCC in the arguments to come, MoCC ought to be interpreted as excluding prudential judgments as I have defined them just now.

Given the distinction I have given between prudential judgments and moral judgments, a question might arise: is it not possible for there to be such a thing as a duty to oneself, and possible to make judgments about such things? In other words, is it not possible to truly make the judgment that benefitting oneself is morally right, and that harming oneself is morally wrong? I believe it is possible for such judgments to be made, and to also be true, but not in a way that denies the distinction between moral judgments and prudential judgments as I have spelled it out in the above paragraph. If one is making a moral judgment about an action by which an agent harms herself, such a judgment will involve a judgment that *others* are indirectly affected by the individual's harming herself. If one is making merely a prudential judgment about an agent's action, they are merely making a judgment about the agent's action and its impact on her own wellbeing. Consider the following example. I might judge that Sam ought to give up smoking. If I am making a moral judgment (if my 'ought' is a moral 'ought'), to the effect that Sam has a moral duty to himself to stop smoking, this might be because I think others around him will be negatively affected by the poor health Sam brings on himself by smoking. I might think Sam owes it to his children to live long enough to see them accomplish as many of their life milestones as possible, for instance. Whereas my judgment that Sam ought to quit smoking is a prudential judgment if I merely judge that he is harming merely himself by smoking. It is possible to judge that prudentially good actions are also morally good, and thus constitute 'duties to oneself'. But prudentially good actions are morally good in virtue of the fact that their benefit is connected to the benefit of others besides just oneself.

On to the third clarificatory question. There is obviously a difference between making a moral judgment and making a *true* moral judgment. MoCC is not to be interpreted as a doctrine about what it takes for moral judgments to be true, but rather as a doctrine about what it takes for a normative judgment to be of the *moral kind*, whether true or false. For a judgment can classify as a moral judgment according to MoCC and still be deemed (perhaps by someone making a contrary moral judgment) a *false* moral judgment. The following example will help make this clear. Suppose I come from a cultural background that practices female genital mutilation (FGM). Suppose I endorse the practice, and make the judgment that FGM is a good practice. Suppose you press me for the reasons why I think this is the case, and I say 'FGM is vital for ensuring faithfulness in marriage, which is an important social good.' My judgment is indeed a moral one because it consists of or implies a judgment about a positive relation to the wellbeing of others. But it is still, quite plausibly, a *false* moral judgment. You could argue against me that there are more welfarist considerations *against* FGM than in favour of it: for instance that the suffering caused by it outweighs any benefit that I think might arise from it, or that marital faithfulness can be encouraged in society through far more humane and effective means, or that it is not fair that women be denied sexual enjoyment while men are allowed it. You might also point out that my claim rests on false empirical premises (a reason why many of our moral judgments are false) about the nature of women: that I am wrong to think that women are more 'inherently promiscuous' than men are. In saying that a moral judgment is false, what one is saying is the judgment attempts to morally evaluate something on the basis of certain welfarist features, but somewhere, the evaluation goes wrong (either because more welfarist considerations support a contrary judgment, or because the judgment rests on false empirical claims). As I said before, the ultimate test as to whether the welfarist considerations appealed to for the truth of a moral judgment are 'adequate' or not depends on how a certain ideal observer would respond to what is being judged. Such a view will be unpacked in chapter 5.

The example in the previous paragraph might, however, raise another thought: there are many norms from other cultures and from within our own that seem to have a moral 'flavour' to them (they are taken very seriously, violators of the norms are punished or ostracised), but which might not, on the face of it, appear to be based on welfarist considerations. Normative judgments about sexual practises are one of the key examples. Condemnations of sex before marriage (or sex that is too casual), homosexuality and incest are all examples of judgments that are usually assumed to be moral judgments, but which philosophers have thought has no welfarist basis.[35] If MoCC is true, what are we to do with this? I believe an easily overlooked fact is that many judgments condemning sexual practices (certainly in popular discourse) *do*, in fact, have welfarist bases: in regards to sex before marriage, or sex that is too casual, the argument is made that sexual intimacy is enhanced, the more discriminating one is about who one has sex with, alongside the argument that sexual breakups wrench bonds formed during intercourse and put strains on future relationships. When it comes to homosexuality, more tacit forms of harms are also appealed to, as with incest. Often, religious justifications are given for sexual mores, but even these mask welfarist judgments: that life goes well for us if we follow 'God's design' for human sexuality. Now, these judgments about the moral status of sexual practices may or may not be true – this is beside the point. The point is to show that many of them *are*, contrary to what is often assumed, judgments based on positive or negative relations to well-being. This means that they qualify as moral judgments according to MoCC, and this in turn gives MoCC even more plausibility. I think it would be a weakness of MoCC if it turned out that many of the judgments about sexual practices turned out not to be moral judgments. There are, of course, many *other* judgments for which we might have to bite the bullet and discount them from being moral judgments: judgments that wiping a toilet bowl

[35] One notable ethicist who has voiced this assumption is Peter Singer (1993: 2-3)

with your nation's flag (ceteris paribus, and providing no one is watching or would be distressed) is wrong, for instance. But, given how many intuitions we have succeeded in accommodating, this is not too big a bullet to bite.

Finally, the last question. To whom, and to whose actions, can moral judgments be reasonably applied? Do moral judgments apply only to human beings? Can we have such things as duties toward non-human animals? It should be noted that the last two questions are *not* two different ways of asking the same question. The question of who moral judgments can *apply* to, and who moral judgments might *concern* are two different questions. I think it can be said that moral judgments cannot be sensibly directed *toward the actions of* animals. For normative judgments evaluate the actions of beings with an appropriate level of agency that most animals (if not all) plausibly lack. It might be reasonable to morally condemn a human being for hacking apart another human being – the former could have done otherwise, and is capable of internalising a normative system prohibiting such behaviour, as Nichols would put it (as we will soon see in the next section, this is true even of psychopaths). But it hardly makes sense to make a moral judgment of a *tiger* who hacks a human apart. The animal does not have the kind of agency that is responsive to normativity in the way that the human killer is.[36] Although animals are not moral *agents*, however, they can be said to be within morality's realm of concern, for they are interest-bearers. For moral judgments are judgments about how our actions are positively or negatively related to the well-being of other interest-bearers, and animals are interest bearers. It is not conceptually confused, then, to make moral judgments about the actions of humans toward animals, such as 'It is wrong to eat meat', 'it is wrong to use animals for experiments'. What has been said here raises further questions about the scope of morality, of course. Does the category of 'interest-bearer' involve *all* animals, or just very intelligent or sentient ones (we may have duties to cats and dogs, but do we really have duties to molluscs?). Do 'interest-bearers' include plants? Is God an 'interest-bearer'?

[36] In case the reader is worried that these comments contradict an earlier claim made in chapter 2 – that rational but not moral criticisability requires voluntary control – let me allay this concern. In chapter 2 it was said that a motivational set could be rationally, but not morally, criticisable on the basis that the features of one's motivational set are not under voluntary control. But this is nevertheless consistent with the claim that certain human *actions* are under voluntary control, which must certainly be required for such actions to be morally assessable. Thus there are two levels of voluntary control of interest here: the first as it applies to the motivational set, and the second to the control one has over one's actions. Humans lack only the former, so are open to moral criticism, while animals arguably lack the latter also, so are not.

What about hypothetical extra-terrestrial beings who are nothing like us? Would they count as interest-bearers too?

I will give some brief, but (for now) unprincipled answers to these questions. I exclude both animals who lack sentience, and plants, from the category of 'interest-bearers'. Certainly, such organisms can flourish or be well-off in a biological sense. But I am making the assumption that (mere) biological life is morally irrelevant.[37] We may have duties to preserve flies, molluscs and plants due to the fact that the ecosystem requires this. But the moral value of the ecosystem being sustained comes back to the fact that we as humans, and other sentient animals like us, will be worse off without this. In short, plants and non-sentient animals are not interest-bearers in the way that we are. A more principled argument will be given for my drawing of this line in chapter five, but from now I think it ought to be provisionally accepted on intuitive grounds: the 'interests' of plants and non-sentient animals are so unlike the 'interests' of sentient beings that it is probable that these concepts are entirely different and that the latter constitutes a quite a metaphorical use of the word 'interests'.

God, depending on the conception that we are working with, could be said to be an interest-bearer. Some conceptions of God are of a completely impassible, deistic, and distant entity such that it might be a mistake to call him an interest bearer – at least one whose interests could be in any way connected to ours. But raw, Biblical conceptions of God paint him as a being who can be grieved, pleased, and who partakes in both our suffering and our wellbeing. Such a being could be said to be both a moral agent and an interest-bearer: a being whose actions can be morally evaluated and a being who can plausibly be said to be harmed (grieved) through our actions, and in relation to whom our actions can be morally judged. In regard to other imaginable beings, the same things can be said. Insofar as they are agents, their actions can be morally evaluated. Insofar as they are interest bearers, they can be said to be beings toward whom our actions can be morally evaluated. Other

[37] Of course, we *may* have duties that involve preserving currently non-sentient forms of life on the basis that they have a sentient future like ours, as various pro-lifers argue in the abortion debate (see Finnis 2011: 17-24, and Marquis: 2011: 51-62). Again, the moral justification for preserving the lives of embryos comes back to a consideration that revolves around the characteristics that I have said that 'interest bearers' possess: consciousness, sentience and the like (even if not currently realised). If, however, there existed a creature that was like an embryo in all respects except that it would never develop the relevant features, then we would plausibly not have moral duties to such beings.

complex questions arise in relation to this, but for the remainder of this chapter, it will not be necessary to go into these.

MoCC, then, gives a neat and plausible formulation of the intuitive differences we draw between moral judgments and other kinds of normative judgments. The very least that can be said about MoCC is that it, or something very like it, forms at least part of any plausible account of what moral judgments are. As we have seen, even committed rationalists will agree that moral judgments are partially delimited by content. The question to be asked now is: does the idea that moral judgments are judgments about *reasons* need to be added to this? In section 2, I will answer, 'no'.

2. The Rationalist Conceptual Claim

2.1 Considerations and reasons, again

We have formulated a content-based doctrine about what makes a moral judgment: MoCC. A similar formulation of the rationalist conceptual claim could be as follows:

> **The Rationalist Conceptual Claim:** A makes a positive moral judgment about X iff A makes a judgment that there is a normative practical reason for X to be done. A makes a negative moral judgment about X iff A makes a judgment that there is a normative practical reason for X not to be done.

Once again, this is worth being clear about. According to RaCC, judging that an action is right is to judge that there is a reason for the action to be done. More specifically, if I judge that it is right that I do X, then I am judging that *I have a reason* to do X. If I judge that it is right that a third party do X, I am judging that the third party has a reason to do X. If I judge that it is wrong that I or that a third party do X, I am judging that I or that a third party refrain from doing X. The only point at which RaCC is not analogous to MoCC is on the question of what neutral moral judgments are. It would not make sense to say that, according to RaCC, if an action is morally permissible, there is neither a reason to perform it nor a reason not to. Consider, for instance, the judgment 'it is morally permissible for me to go to the movies'. This hardly implies that there is no *reason* for me to go to the movies.

The fact that going to the movies will bring me enjoyment may well count as a normative practical reason for me to go.

On this note, recall the distinction between considerations and reasons that was made in chapters 1, section 2.3. A consideration is a fact about an action which *may or may not* count as a reason in the normative sense. Take a given action: eating a piece of chocolate cake today. There is a fact about this action, a consideration: that doing so will cause me to ingest 500 calories. This consideration may, or may not count as a normative practical reason (to either eat or refrain from eating). Whether the consideration counts as or constitutes a reason depends on whether or not it meets certain conditions specified by the right account of practical reasons. (According to me, internalism is the correct theory of practical reason, so the fact about the cake's calorie content will count as a reason in favour of or against eating the cake depending on what my motivational set is. If I desire to lose weight, the cake's calorie content will count as a reason to refrain from eating it. If I desire to gain weight (suppose I want to move up a weight grade fast for my next wrestling tournament), then the cake's calorie content will count as a normative reason to eat. In summary, considerations are facts about actions which *may count* as normative practical reasons depending on whether they meet certain conditions). Now, moral considerations are just facts about an action of a certain type: facts about the positive or negative relations between actions and the wellbeing of others. A *moral reason* in the normative practical sense is a moral consideration that *also* meets the conditions necessary for it to count as a practical reason. If I judge that there is a moral consideration in favour of me giving to charity (that is, a positive relation between my action and the well-being of others), this need not mean that I also believe that such a consideration counts as a *normative practical reason* – I merely believe that there is a fact about the action of giving to charity that counts as a moral consideration which, *from the moral point of view*, counts in favour of the action. But if I judge that I have a moral reason to give to charity, this means that I judge that there is a moral consideration, and that the consideration *constitutes a reason for me to act* – I take the moral point of view to be reason-giving.

To be clear, these are the two alternatives we are comparing in this section. The first alternative is to simply accept MoCC, and argue that moral judgments are simply delimited by content. The second alternative is to accept a combination of MoCC *and* RaCC:[38] to judge that I ought to ϕ would be to judge that ϕing has a positive relationship to welfare, *and* that I have a reason to ϕ.

2.2 Amoralists again

But what has just been said raises a familiar question. Is it even *possible* to judge that there is a moral consideration present without a judgment about reasons being present also? In other words, is it even *possible* to judge, for example, that there is a positive relation between ϕing and the well-being of others, and *not* also take this to be a reason for ϕing? Is it possible to judge that there is a negative relation to the well-being of others and *not* also take this to be a reason for refraining from ϕing?

The obvious answer to this question, I believe, is, 'yes'. All we need to ask ourselves to establish this is the following question: 'is it conceivable that there could be an agent so callous, so unfeeling toward the plight of others, that they would not view facts about the wellbeing of others to be reasons for themselves to act?' I think the answer to this question has to be 'yes'. We *can* imagine people, characters, who do not care about others at all. The concept of a thoroughly uncaring person is not a conceptual confusion. We can give such a figure the familiar label: the amoralist.

At this stage, it is worth asking some questions that might initially raise doubts about the conceivability of the amoralist. The thesis that I am putting forward is that judgments about the relations between actions and the well-being of others, and judgments that such facts constitute reasons, are conceptually or *a priori* dissociable from one another (call this the 'dissociability thesis' or DT). In order to see whether DT is really true, we have to briefly play devil's advocate about this concept of the amoralist. To begin with, let's distinguish between two putative kinds of amoralist, so that we may ask whether either is conceivable. These two kinds are the psychopath, and what I will call 'the reflective amoralist'.

We will first see that the figure of the psychopath could be marshalled to threaten DT, rather than confirm it. Psychopaths are rarely motivated, and rarely take as reason-giving,

[38] The choice is not one between accepting MoCC, or accepting RaCC by itself, as we have already seen that even those who accept RaCC need to accept MoCC or some other similar content-based claim to make sense of the difference between *moral* reasons and other kinds of reasons.

facts about the wellbeing of others. How, then, one might ask, does the psychopath *threaten* DT? Doesn't the existence (and mere conceivability) of psychopaths rather confirm DT? The answer to these questions depends on what the true source of a psychopath's callousness is. A popular lay suggestion is that psychopaths don't actually possess the concept of another person, and thus do not possess the concept of 'another person's wellbeing'. The reason that they don't take facts about the well-being of others to be reason-giving is because they don't possess the concepts of the well-being of others in the first place. If this suggestion were true, this might undermine DT: psychopaths would *not* constitute an example of people who can make judgments about the relations between their actions and the wellbeing of others while not taking such facts to constitute reasons, because they do not even possess the concept of another's well-being. But this popular suggestion is very likely false. As Nichols has pointed out, the perspective-taking skills of psychopaths are quite developed, which explains why they are good at manipulating others (2004: 79). It is quite evident that psychopaths have possession of the concepts of other people, and of their benefit and harm: it is their lack of *empathy*, not their lack of awareness of certain facts about actions, which explains their callousness. Another challenge that might be marshalled in an attempt to undermine DT is the evidence that psychopaths tend to fail the moral/conventional test (ibid: 76). Psychopaths tend to evince no distinction between moral (harm-based) wrongs, and conventional wrongs (wrongs having to do with social unacceptability). Once again, the psychopath's failure to distinguish between harm-based wrongs and conventional wrongs does not necessarily stem from the fact that it is inconceivable that psychopaths could internalise a normative system prohibiting harmful actions (as Nichols emphasises, ibid: 19). It more likely stems from the fact that harm-based wrongs do not *strike* a psychopath as worth noteworthy. In the same way, I may fail to succeed in making distinctions between ancient Chinese and ancient Japanese etiquette prohibitions: not because I am unable to possess the concept of these differences, but because, due to my lack of interest (analogous to the psychopath's lack of empathy), the differences between these two kinds of prohibitions do not *stand out* and I am merely, as a matter of contingent fact, not in the habit of recognising them. It is plausible then, that psychopaths constitute an example of a person about whom DT could apply: someone who is conceivably capable of recognising facts about the relations between actions and welfare, but not taking such facts to be reason-giving.

Another kind of amoralist we can imagine, however, is a figure I would like to call 'the reflective amoralist'. The reflective amoralist, like the psychopath, does not take considerations about the well-being of others to be (in and of themselves) reason-giving. But unlike the psychopath, the reflective amoralist was not born with a lack of empathic potential. Like most people, the reflective amoralist may have even started out learning to make moral judgments the way most people learn to: by identifying actions that cause them a degree of negative affect through their empathic mechanisms, as harm-based (and thus moral) wrongs. But, at some point in the reflective amoralist's life, she made a decision that she was not going to care too much about the welfare of others. She was going to live a life with narrow, selfish concerns, and only care about the well-being of others insofar as doing so related to these narrowly selfish concerns. An amoralist like this, again, is certainly conceivable, and supports DT even more than the figure of the psychopath does.

At this point, an objection might arise. If empathy is implicated in the reflective amoralist's moral judgment making, doesn't this mean that they are judging that they have a reason – *an internal reason* – to act in accordance with their moral judgment, given that empathy is perhaps inherently motivating and thus constitutive of an element of the subjective motivational set? There are two reasons why the answer to this question is 'no'. When a reflective amoralist (or anyone, for that matter) has an empathic reaction as they make a moral judgment, this does not entail that *their moral judgment consists in the judgment that they are judging that they are having an empathic reaction*. If it were the case that making a moral judgment entailed *judging* that one has an empathetic reaction, then, yes, there would be a conceptual connection between making a moral judgment, and judging oneself to have a particular kind of internal reason for acting. But all that Nichols has claimed is that an affective mechanism is *implicated* for (most) people when they make moral judgments, not that they are making a judgment about the latter. What this fact entails is that, most likely, when we make moral judgments there is, *as a matter of contingent fact*, an internal reason for us to act in accordance with our moral judgments. But this is entirely compatible with DT, which denies only a *conceptual* connection between two kinds of judgments.

A second reason that the answer to the above question is negative is that it is not even necessary that empathic reactions *are* always implicated in moral judgment making, and this is especially so for the reflective amoralist. Even if the reflective amoralist *initially* learns to make moral distinctions on the basis of negative or positive affect that is caused in her, this need not mean that she needs such affect to draw such moral distinctions on every instance thereafter. Suppose, when she first notice one child teasing another at age three, negative affect in response to the harm being done arises, and as a result she makes her first ever judgment about the moral status of teasing: that teasing is wrong. Years later, at age sixteen, and after she has become a reflective amoralist, she sees another instances of teasing, and she makes the same judgement: that teasing is wrong. However, this time, no negative affect comes as part of the package (having been an amoralist for some time now, her empathy is beginning to be significantly dulled). She no longer needs her empathic reactions in order to make moral judgments: at the very most she merely needs to *remember* that she judged teasing to be wrong the last time she saw it, that there are no major descriptive dissimilarities from between the first instance of teasing and other instances of teasing, and thus that there cannot be a difference in moral status. Once the normative system is internalised via empathic reactions, it does not necessarily need the continuation of empathic reactions to sustain it. Again, there seems to be nothing wrong with the idea that an occasion of making a moral judgment need involve either motivation from an empathic source, or a judgment that one has reason to act in accordance with one's judgment. (The above 'memory' argument shows why even sentimentalist views like Slote's do not vindicate motivational internalism, despite his assumption to the contrary 2010: 46).

Enough has been said on this point. Making a judgment that there is a positive relation between an action, and the wellbeing of others, does not entail that one will make the judgment that one has any kind of reason to perform the action. And making a judgment that there is a negative relation between an action, and the wellbeing of others, does not entail that one will make the judgment that one has a reason to refrain from the action. Judgments about what we have reason to do will motivate us. But it is perfectly conceivable that there exist people who remain unmotivated after recognising positive relationships between actions and the wellbeing of others: I hope to have shown that the figures of the psychopath and the reflective amoralist show this.

It is worth noting that similar sounding amoralist arguments have been presented by philosophers against RaCC in the past (see in particular Brink 1989, Foot 1972, and Shafer-Landau 2005). One might think that arguments between those who endorse RaCC and those who endorse amoralist arguments are only bound to end in a stalemate: those who adopt RaCC believe that moral judgments just are judgments about reasons, and that the amoralist is inconceivable (and that all apparent amoralist judgments are really just making "inverted commas" judgments.[39] moral judgments when they remain unmotivated by them). Those who deny RaCC believe that moral judgments are merely judgments about welfare-based considerations, and that the amoralist, who does not care about these things, is conceivable. How is this debate to be resolved?

2.3 The amoralist question resolved

We have taken some steps, so far to resolving it. We have established that MoCC is true, and that one necessary condition of making a moral judgment is that it is based on welfare based considerations. We have also established that welfare based considerations, like any kind of consideration, are dissociable from the concept of a normative practical reason (the dissociability thesis). The question we must now ask is: is there any reason to think that one must be taking a moral consideration to constitute a reason, in order to count as making a moral judgment?

Let's clarify what the two options we have before us. Our first option is to accept MoCC: to say that what it takes to make a moral judgment is simply to make a moral judgement about the positive or negative relations between an action, and the well-being of others, and that this confers the moral status that the action has. Our second option is to accept both MoCC and RaCC: to hold that the only people who *count* as making moral judgments are moralists: people who make judgments about positive/negative relations between an action and the well-being of others *and who also judge that such considerations are reason-giving*. Intuitively, the second option looks quite suspect. Once we have already noted that judgments about paradigmatic moral considerations are conceptually dissociable from judgments about normative practical reasons, it looks quite arbitrary and *ad hoc* to say that only those who count moral considerations as reason-giving count as making moral claims. What possible grounds could there be for insisting that this is the case?

Some of the grounds offered by Michael Smith will be looked at in section 2.4 and 2.5: among them the argument that moral judgments are thought to be morally motivating.

[39] To use Hare's (1964: 164) phrase.

But, as we will see, most of these arguments either beg the question (that is, they require the assumption that moral judgments are in fact judgments about reasons), or they ignore MoCC and make the confusion between the two different senses in which the term 'morality' can be used, which was mentioned in 1.2.

So arguments in favour of this extra, necessary condition for making a moral judgment will be found wanting. But are there any good *positive* arguments that we can offer in favour of thinking that MoCC outlines the only necessary conditions for a judgment being a moral judgment? I will outline what I take to be a good presumptive argument for the remainder of this section.

Imagine an ethics classroom, in which a discussion about abortion is taking place. The goal of the ethics discussion is to try to get to the truth about the moral rightness or wrongness of abortion. Imagine that there are four participants in the discussion: two pro-choicers on one side, and two pro-lifers on the other. The pro-choicers make some familiar arguments which take into consideration certain points about the welfare of certain interest-bearers: a woman has a right to determine what happens to her own body, particularly if a pregnancy has resulted from rape. Surely, the pro-choicers argue, the fact that the woman is a completely self-conscious individual capable of having her life disrupted by an unwanted pregnancy entails that abortion is morally permissible in this circumstance. The two pro-lifers argue against them, both of them citing considerations about welfare for their own arguments. The foetus, they say, ought to be considered to be a person and as such can be said to have interests and well-being. Surely the greatest way of harming another individual is by killing them and taking away their future, and so the foetus has a right to life that outweighs the mother's right not to be inconvenienced or harmed in some other lesser way, even when the pregnancy has resulted from rape. All four parties to the discussion look like they are potentially tracking moral truth. Of course, it is probably going to be the case that one side is wrong, or at least not as right, as the other side.[40] But we can see that all four of them are in the ball-park when it comes to tracking moral truth: we take all of their judgments as what may be called 'serious candidates' for the tracking of moral truth. None of them are saying anything that just *could not plausibly be regarded as* morally correct – like a person who claims that scratching your left ear before your right ear is morally

[40] Assuming there is no indeterminacy in this case.

obligatory, for instance. All four disputants in the discussion are making judgments on the basis of considerations to do with welfare, and are therefore in the game of tracking moral truth. Suppose we were to find out that one of the pro-lifers is an amoralist. Should we think that the two pro-choicers should only bother debating with the moralist pro-lifer, but forget about the amoralist's arguments, because she is not really making moral judgments anyway? Surely not! To say this would be most implausible. The amoralist pro-lifer, in the scenario, is capable of making exactly the same arguments as her moralist counterpart: abortion is deliberate killing of a being that can sensibly be said to have interests and well-being, so abortion is morally wrong. The fact that the amoralist doesn't particularly *care* about the well-being of others, and would see no reason for *her* not to have an abortion *if she got pregnant*, does not seem to alter this fact. She would simply say of herself, if she were to have an abortion, that she is doing something morally wrong, but that she doesn't see this to provide her with a reason not to have an abortion, because she just doesn't care about other people's welfare, and gets along just fine in life not caring. But the fact that she can make exactly the same arguments about the wrongness of abortion as her moralist counterpart, and can do so just as clear-headedly, would surely have to imply that she not only capable of tracking moral truth just as well as her moralist counterpart, but tracking moral truth *because she is making the same kind of judgments as her moralist counterpart*. She would be capable, on this understanding, of saying she *would* have reason to refrain from having an abortion were she more concerned about doing the morally right thing! If her pro-life moralist counterpart is right (or at least mostly right) in the debate, and is attempting to track moral truth and is succeeding, then so is she. But if this is the case surely we must conclude that moralists and amoralists alike are capable of making moral judgments.

Of course, there is a well-known objection to the idea that amoralists are making genuine moral judgments, despite being able to track moral truth. This objection is given by Michael Smith,[41] who makes his point through the following scenario. Smith asks us to imagine a person who has been blind all her life, but who nevertheless is able to track truths about colour very well (perhaps even better than those who can see). She is connected to an apparatus which allows her to feel, through her skin, which objects reflect which

[41] See Smith, 1994: 68-71

wavelengths. Despite the fact that her judgments about the objects track truths about colour, however, it would be wrong to suggest that she is therefore making judgments about colour. This is because she simply fails to possess the colour concepts that ordinary users of colour terms possess, by which they track truths about colour. The way in which ordinary users track truths about colour is through use of colour concepts, which are irreducibly visual experiences. The way in which the blind person tracks truths about colour is through a means besides the use of colour concepts. It is therefore sensible to say that ordinary users of colour terms make genuine colour judgments, whereas the blind person does not. This, Smith thinks, is what we should say about amoralists. Though they are capable of tracking moral truth, they are not doing so through the use of ordinary moral concepts. Therefore, they are not using moral judgments.

I accept Smith's claim that the blind person in the scenario fails to make colour judgments. What I reject is that the amoralist's case is analogous to the blind person's case. For, the basis for the claim that the blind person does not possess colour concepts is the claim that she tracks truth *in a different way* to the way ordinary users of colour concepts do. The visual data is the means by which ordinary users track truth, and this visual data is what constitutes the colour concepts. The blind person lacks colour concepts, tracking truth a different way, because she lacks the visual data. But note, we *cannot* say this about the amoralist's case. The moralist uses particular concepts: considerations to do with the well-being of others, in order to track moral truth; call these 'moral concepts'. The amoralist pro-lifer in the above example uses *the same concepts as the moralist counterpart* and is tracking truth *in the same way*. Given this, we must surely conclude that Smith's colour analogy fails as an objection to the idea that amoralists track moral truth by making moral judgments. Amoralists do make genuine moral judgments: moral judgments that track (or are capable of tracking) moral truth by means of the same concepts that ordinary users (that is, moralists) employ in tracking moral truth: welfarist concepts. Thus, the possibility that an amoralist could make a genuine, and true, moral judgment is a real one.

In summary, the claim that only moralists count as making moral judgments is implausible. We have seen that the best explanation for why amoralists track moral truth is *because they make moral judgments*: they use the same concepts that moralists do when tracking moral truth. One might, in a last-ditch attempt to save RaCC, contend that amoralists only make *partial* moral judgments. This objection could grant that amoralists use the same concepts as moralists do when tracking moral truth, but argue that these truth-tracking concepts are not the only concepts relevant for making moral judgments. To make a moral judgment, one must not only track truth in the same way that 'ordinary' users of moral concepts track it, but one must also take such considerations to be reason-giving.

But now the obvious question is: what basis is there to persist with the thought that something more than these content-based concepts are relevant for the making of moral judgments? There had better be some good positive support for such a view if there is to be any vindicating of RaCC. It is now time to move on and examine some commonly given arguments in support for RaCC, the validity of which I have already subjected to questioning. The most notable arguments come from Michael Smith's 1994 work, *The Moral Problem*. But even these arguments, I shall show, are unpersuasive.

2.4 Smith and confusion over 'morality'

A reason that philosophers have found the rationalist conceptual claim appealing is due to the stubborn intuition that moral judgments are supposed to have great practical clout. Smith believes that moral norms have the practical significance that I have accorded only to reasons judgments. One of the reasons that Smith thinks this is revealed when he claims that a change in our moral opinions constitutes a change in our *most fundamental values* (1994: 71). Thus, unlike other kinds of norms, which may be ignored or not, morality, like practical reasons, are supposed to indicate *what we should actually do* – this is the role that our most highly held ideals or values play. One's most highly held ideals are the ideals that one takes to be reason-giving. It is thus easy to see why Smith's claim that moral judgments reveal one's most highly held ideals would support the rationalist conceptual claim. But we can see from this argument of Smith's that he is speaking of 'morality' in the 'personal morality' sense. Needless to say, Smith is no subjectivist or expressivist, so does not adhere to either of the ways discussed in 1.2 that the 'personal-morality' concept has been accounted for in moral judgment. But Smith adheres to the notion in his own way: a change in our moral convictions represents a change in our most fundamental values, by which he means, a change in our judgments about what we have reason to do. But this 'personal-morality', like the other uses of the concept, is a very separate one from the concept of 'morality proper.'

Smith does tip his hat to 'morality proper' when he acknowledges that morality is partially defined by its content (1994: 183-4) – when Smith discusses the body of *moral* reasons we have (as opposed to *prudential* reasons) he does seem to be discussing something that approximates what I have called 'morality proper'. So Smith does seem to use morality in the two different senses: one is that 'morality' refers to one's most highly held set of ideals, and the other is its content of having to do with the promotion of the well-being of others. My contention is that, by moving between these distinctions, Smith engages in something of a fallacy of equivocation in trying to show that the requirements of 'morality proper' are reason giving: 'Morality' denotes our highly held ideals and thus is reason-giving. 'Morality' also denotes a system of standards concerning welfare. But what Smith doesn't acknowledge is that, between different uses of the term 'morality' on different pages, he has changed the subject.

In order for Smith's argument to be successful, and for no fudging to be taking place, Smith would have to show that it must be *conceptually necessary* that people have 'morality proper', and 'one's highly held set of ideals' are the same concept: that is, he would have to show that it is inconceivable that anyone *not* have the content-based, paradigmatic welfarist concerns of morality proper as one's personal ideals. That this is can be shown is highly unlikely; and this, I think is what the key point of the famous 'amoralist challenges' are.

2.5 Motivational internalism

When one makes a moral judgment, one will be motivated to act accordingly. Or so this is according to a doctrine known as *motivational internalism*. If motivational internalism is true of moral judgments, it may indeed indicate that moral judgments are reasons judgments. For why would moral judgments have such an effect on the motives if they were not judgments about reasons? Therefore, moral claims must be claims about reasons (this line of reasoning can be traced through Smith's *Moral Problem*, 1994: 6-7. See also Nagel 1970: 7, 13).

To be clear, the argument that the rationalist would have to propose is as follows: whenever someone makes a judgment about what I have called 'morality proper', one is motivated to act accordingly. The best way to respond to this argument is to simply deny that motivational internalism is true.

But before I do this, I must briefly draw attention to two differences in the way that motivational internalism has been presented as a view. Michael Smith presents motivational internalism in the following way:

> If an agent judges that it is right for her to ϕ in circumstances C, then either she is motivated to ϕ in C or she is practically irrational. 1994: 61

Expressed such a way, motivational internalism could be taken to be presupposing one of two doctrines. One doctrine it could be taken to be presupposing is that it is a substantive requirement of rationality to be motivated to do what one judges to be morally required (hence one counts as 'practically irrational' absent this). I have already dealt with arguments to this effect in chapter 2. But another doctrine that this motivational internalist doctrine could be presupposing is just the rationalist conceptual claim itself: because one's moral judgments are judgments about what one has reason to do, one will be motivated by them, or else one will be, by definition, practically irrational (for it would be practically irrational to fail to be motivated by what one judges to be a reason).[42] As such, this version of internalism cannot be appealed to by Smith in order to *show* why the rationalist conceptual claim is true, given that the latter needs to be presupposed. However, other formulations of motivational internalism seem to render it equally problematic:

> The… internalist about motives claims it is a conceptual truth about morality that moral judgment or belief motivates. According to the internalist, then, it must be conceptually impossible for someone to recognise a moral consideration or assert a moral judgment and remain unmoved. Brink, 1989: 46

[42] As it happens (and my arguments in favour of the 'motivational link' from chapter 1 suggest this), I doubt that this particular kind of practical irrationality is possible. If one makes a judgment about one's reasons, one will be motivated to act in accordance with it, and failure to be motivated will constitute evidence that a judgment about one's reasons was not really present. Of course this leaves other kinds of practical irrationality entirely possible: practical irrationality that consists in failing to recognise one's reasons for actions in the face of good argument, which on Williams's view would amount to a kind of failure in the deliberative process, or a refusal to even engage in it. But these forms of practical irrationality presuppose that the agent is failing to make a judgment about their reasons that they should make (or making a judgment that they should not make), rather than that they have made a reasons-judgment which fails to motivate.

According to Brink, if non-conceptually confused, sincere moral judgments necessarily motivate, then surely this is a difference between moral norms and all other norms is one that requires explaining. And, as its supporters will claim, the rationalist conceptual claim seems well placed to explain it: if moral judgments necessarily motivate, then perhaps that is because moral judgments are judgments about reasons. As I argued in chapter 1, there is a motivational link between making a judgment about one's reasons, and being motivated to act accordingly. If motivational internalism is true of moral judgments too, then perhaps this is good reason to suppose that moral judgments are reasons judgments.

So, is motivational internalism true? As Brink says, what the motivational internalist about moral judgments needs to be able to prove is that it is conceptually impossible to make a moral judgment and remain unmoved by it. Intuitively, the answer is, 'no'. If something that distinguishes morality proper from other norms is what it is *concerned* with (and, as I have pointed out, even Smith and other rationalists concede this): welfarist concerns such as reciprocity, avoidance of harm to others, and the like, then one would expect it to be the case that *these concerns* are also necessarily motivating, if moral judgments are. But there is no reason to suppose that *these concerns* are necessarily motivating. Once again, the existence of callous people and amoralists (let alone the conceivability of them) vindicates this.

2.6 Smith's argument from moral fetishism

There is one more argument from Smith to consider before we abandon the rationalist conceptual claim, and this is a further argument that Smith gives in favour of motivational internalism. Since you can't have motivational internalism without accepting the rationalist conceptual claim, this argument can be seen as an argument that would lend support to the rationalist conceptual claim. Smith claims that motivational internalism gives us the right account of the moral motivations in the 'good and strong-willed person'. The motivational externalist, on the other hand, is committed to the claim that self-conscious moral motivation is a moral virtue, when, on the contrary, there are good reasons to think it a vice. Smith's argument for this can be summarised as follows (1994: 72-75). When a good and strong-willed person judges that she morally ought to vote for the social democrats (for example), she is motivated to do the right thing *de re*, not *de dicto*. In other words, she is motivated in particular to vote for the liberal democrats, not simply to 'do the right thing, whatever that may turn out to be'. And her motivation to vote for the liberal democrats follows *directly* from her moral judgment. The externalist, however, must give a different account of moral motivation. According to the externalist, the person who is good and strong-willed is motivated to do the right thing *de re* because it is derived from an existing motivation to do the right thing *de dicto*. According to the externalist, our voter's motivation does not arise simply from her moral judgment that voting for the social democrats is the right thing to do, but because this motive is derived from a more general motive to 'do the right thing'. If this is the case, Smith thinks, then we have 'a straightforward *reductio*' against motivational externalism (ibid: 75). For the motivation to 'do the right thing' is arguably a vice, as we tend to think that:

> Good people care non-derivatively about honesty, the weal and woe of their children and friends… people getting what they deserve, justice, equality, and the like, not just one thing: doing what they believe to be right, where this is read *de dicto* and not *de re*. ibid.

Smith reminds us of William's man, who is faced with the choice between saving his drowning wife, or a drowning stranger. If his thought process is 'this is my wife, and in situations like this it is morally right to save one's nearest and dearest, so I should save my wife,' (as many an impartialist ethicist might want) we would accuse him of having 'one thought too many', or would find such a thought process morally objectionable. His thought should simply be '*She's my wife!'* The man should be 'non-derivatively' concerned about saving her, rather than having a motivation to save her that is derived from a broader motive to do the right thing by impartialist rules. Smith thinks the externalist's view about moral motivation in general is problematic in an analogous way to the impartialist's views about the motivations Williams's man should have when saving his wife. We are having 'one thought too many', and this is morally objectionable. On the externalist's view, the moral motivations of good people constitute a fetish for morality, for doing-the-right-thing-de-dicto, and this is a problem for motivational externalism.

I believe Smith's argument can be responded to in a number of ways, and that, on the whole, it is unpersuasive. It seems that Smith is making an objection to what he presumes is the phenomenology behind the thought processes of someone who is motivated to do the right thing de-dicto, given the analogy he draws with Williams's man. Smith seems to think that, on the externalist's view, ordinary moral motivations like caring for one's near and dear, donating to charity, and the like, are accompanied by such objectionable thought processes: 'I must care for my children and friends because it is the right thing to do,' or, 'I said I would keep my promise to help my friend move house, and since keeping promises is morally right, then I ought to help my friend move house.' The idea behind Smith's objection seems to be that the moral motivation to do the right thing *de re* should be more instantaneous than this, and is prevented from being so by the desire to do the right thing *de dicto*. I agree that, in such ordinary situations, the desire to do the morally right thing *de re* should be instantaneous, but I think Smith is wrong in supposing that the externalist is committed to the view that they can't be. Having a standing motivation to do the right thing *de dicto* need not render *de re* motivations any less instantaneous than they would be on Smith's account.

We can see that this is so when we consider a familiar non-moral example. Suppose I see a bus hurtling toward me, and I have a desire to jump out of the way immediately. We can imagine my having the instantaneous thought: 'I must get out of the way!' (Or maybe just: 'doge!'). We can also agree, however, that my desire to jump out of the way is a derived desire: a desire to preserve my life. So we can say that my desire to save my life *de re* (my desire to avoid the bus) is derived from a desire to save my life *de dicto* without having to suppose that my conscious thought process must be: 'Because I desire to preserve my life and because being hit by the bus will frustrate this desire, I must jump out of the way.' The reason for this is simple: the desire to preserve my life is so ingrained, and the belief that moving buses take lives is so ingrained, that they don't need to be re-hashed in my mind. Similarly, the externalist can say that, in good people, uncontroversial judgments such as 'it is right to keep promises to friends' and 'It is right to care for family and friends', are so ingrained, obvious, and unquestioned to the agent that they need not re-hash them in conscious thought processes. The only conscious thought processes there need to be are expressions of desires to do the right thing *de re*.

A second objection that can be levelled against Smith is that the desire to do the right thing, *de dicto*, rather than being a vice, is a virtue. We consider the ability to change one's moral judgments and motivations, even when such a change may be emotionally painful, as a virtue. I put it to Smith that standing desires to do the right thing *de dicto* must be what explains such changes in people with integrity. Take the following example. Richard is a wealthy baby boomer who is a powerful CEO in the fossil fuel industry. Richard has no scruples about his position: human induced climate change is a myth, those who say otherwise are against progress and the betterment of the human race, and against the provision of millions of jobs that keep families going around the world. In other words, Richard believes that his role in the fossil fuel industry is completely morally justified, and his motivations fall in line with this judgment. He does his job diligently, and plays a part in the funding of many climate change sceptic organisations. But suppose, over time, Richard's moral judgments begin to be rattled and eventually changed. He reads up on the science more and more. The carbon content of the atmosphere has been raised from 280 parts per millionth to 400 since the industrial revolution. Such a change is far more dramatic than any other climate change in earth's history, and Richard begins to see the deception in the sceptical arguments he has funded. More grim truths hit him: the rate at which humanity is pumping carbon into the atmosphere is four times faster than it can afford to be if a global catastrophe is to be averted within the next generation. This means that the human race cannot afford to dig up any more coal or oil. It has to stay in the ground. Richard's very job ought to be moving toward obsolescence, but instead Richard's efforts have prolonged it. It gets worse: fossil fuel consumption is increasing, not decreasing. If disaster is to be averted, the fossil fuel industry must shrink dramatically within the next five years. Richard is hit with the most unpleasant moral conviction of his life: he must quit his job, divest, vote for a carbon tax, and encourage others to do the same. Spare a thought for poor Richard: his whole life and values have been built on the fossil fuel industry and the benefits he believed it to be bringing to the world. Through his whole life he has been convinced that he had made all the right choices, and now must publicly admit, through his actions, that he has been wrong. Suppose Richard does do the right thing, and makes the above sacrifices (for he is a 'good and strong willed' person). Given his painful predicament, what could possibly have explained his radical change but a desire to do the right thing, *de dicto*? Richard, being a good and strong-willed man, has always had desires to do the right thing, both *de dicto* and *de re*. Previously, he thought that being involved in the fossil fuel industry was the right thing to do, *de re*. Now he thinks that shunning it is the right thing to do, *de re*. Both motivations to do the right thing at the different times must be explained by a standing desire to do the right thing, *de dicto*.

Naturally, the reader will object at this point that the same can be true on Smith's internalist account: Richard's change in motivation can have changed simply by virtue of his change in moral judgment. But I say in response that, if Richard merely possessed a desire to do the right thing *de re*, and no desire to do the right thing *de dicto* it would be very unlikely that his moral judgments *ever would have changed in the first place* given the pain involved. Let me explain.

We are all familiar with the human tendency to harden our existing moral judgments whenever changing them would involve pain (whether it be simply the shame of admitting we were wrong, or the pain of giving up certain choices and privileges that our moral stances allow us). Furthermore, this desire to avoid such pain can induce us to close our ears to many sensible arguments and objections to our position. We can imagine how Richard's story could have gone. He could have chosen not to investigate the real science behind climate change when doubts about his position niggled at him. He could have chosen to continue listening to the voices of sceptical friends and authors and avoid mixing with those who would challenge his views. There are many post-hoc rationalisations he could have made in favour of his position. We can see this phenomenon at work in many scenarios. Many women who desire unlimited sexual freedom obviously find attractive the view that abortion is morally unproblematic. Such women often will not even countenance reasonable questions about whether foetuses or embryos have rights to life (that one would even investigate this ethical dimension to abortion debates is dismissed as morally offensive). Many men who desire unlimited sexual freedom are often wilfully blind to the observations about the harms that the pornography and prostitution industries wreak, and will continue to put up arguments (however weak) in support of their 'rights as individuals' (and if guilt about the well-being of prostituted women should arise, the argument is fabricated that they are acting freely too, and enjoy their choices). And yet such men and women can still have very powerful motivations to do what they take to be the right thing *de re* – such motivations are easy to act on when one's moral views are so convenient!

The kinds of people who are more likely to change their moral judgments in the face of clear (even though painful and confronting) arguments, and who are less likely to be susceptible to conviction hardening and post-hoc rationalisations, I propose, are people who have desires to do the right thing *de dicto*, not simply *de re*. Take Richard again. Suppose Richard had no clear desire to do the right thing, *de dicto*. If this were the case, he would have had little or no desire to doubt his moral judgments that fit with his existing motivations to do what he took to be the right thing, *de re*. It would have been easy for Richard to persist in his existing moral judgments and motivations were it not for thoughts like: 'but I do, after all, want to *do the right thing*… so I'd better make extra sure I *am* doing the right thing and look honestly at the other side of the argument!' If Richard were motivated only to do what he took to be the right thing, *de re* – working for the fossil fuel industry – there would be little reason for such thoughts to arise. It takes a forceful desire to do the right thing *de dicto* to help us scrutinise our *de re* motivations and judgments, especially when we are so attached to them. The point is simple: far from being a vice, having the motivation to do the right thing *de dicto* is a virtue found in the most morally reflective, good, and strong willed people, contrary to Smith's assertion.

3. Joyce's Error Theory

3.1 Joyce's argument for moral error theory revisited

In chapter 1, I gave an overview of Joyce's argument for moral error theory. I classified Joyce as one who subscribed to the rationalist conceptual claim: that moral judgments are judgments about reasons, a view I have just dismissed. However, there are other available ways of interpreting Joyce's argument that do not require him to be committed to the rationalist conceptual claim. Indeed, since *The Myth of Morality* Joyce has somewhat distanced himself from the claim that moral judgments are judgments about reasons, and has argued that moral discourse might be committed to external reasons in some softer way (see, for instance, 2006: 62-4)

So although one possible interpretation of a premise in Joyce's argument is an affirmation of the rationalist conceptual claim, another interpretation is available. I now turn to the question of whether this 'other interpretation' vindicates error theory. So far, I have argued against the rationalist conceptual claim, and thus argued against the idea that moral discourse is committed to external reasons in the sense that moral judgments *are* judgments about reasons. But a question worth addressing in closing this chapter, and one that Joyce would surely urge us to ask, is whether there might be any *other* sense in which users of moral discourse are committed to external reasons, which might vindicate error theory.

3.2 Implying and Implicating

One way of understanding Joyce's claim that external reasons might be a conceptually non-negotiable feature of moral discourse is that speakers, in expressing moral judgments, make *implications* concerning external reasons. To start this discussion, I must make it clear just what type of 'implication' is being discussed here. There is a distinction to be made between implications that are made in virtue of the meaning of words, and implications that are not. Paul Grice drew this distinction as one between 'implying' and 'implicating' (Davis, 2013) and I will use the same terminology here. Consider that the statement 'John is a bachelor' implies 'John is a married man'; that is to say, anybody who makes the statement 'John is a bachelor' implies 'John is an unmarried man', because the concept of a bachelor *just is* the concept of an unmarried man. This type of implication is the first type of implication: an implication that is made in virtue of the meaning of the word.

Consider a second example. Imagine that someone asks me: 'are you coming to dinner tonight?' to which I reply: 'I have to work.' The statement 'I have to work' does not *mean* 'I cannot come to dinner', but, given the context, the question asker will know that this is what I am saying. In this situation I am *implicating* (rather than implying) that I am not coming to dinner – I am using words and concepts to convey something other than what is in their meaning. To recap, the distinction between implying and implicating is as follows: when one implies X, one does so by saying Y, which entails X. When one implicates X, one does so because one uses Y to communicate X (by virtue of the context of one's utterance of X, or shared assumptions that one's hearers understand) even though Y does not mean X.

The main difference between implying and implicating can be expressed in the following way: terms or concepts *themselves* are responsible for the fact that we imply certain things when we use them. But we *speakers* are responsible for implicating things, for which the concepts themselves are not responsible.

Having made the distinction between implying and implicating, I am now in a position to look at Joyce's arguments that he uses to suggest that moral discourse is somehow committed to external reasons. An apt interpretation of what he means by this, we will see, is that we use moral judgments to *implicate* judgments about external reasons. Whether or not moral claims *imply* external reasons claims in the *first* way has already been discussed, and dismissed, because I have argued in this chapter so far that normative reasons, and moral standards, are two entirely different concepts. But now we should see if the suggestion that speakers use moral judgments to *implicate* judgments about external reasons would vindicate error theory. See Joyce's following comments:

> Consider again our moral condemnation of a felon. We say (at the very least) "You ought not to have done that." We cannot end matters there, or we have nothing with which to counter the felon's "So what?" Indeed, if *all* we had to say on the matter was "You simply *mustn't*!" – accompanied by some table-pounding – then the felon's query seems positively reasonable. We seek something that might *engage* the criminal. Even if it is something that does not succeed in actually persuading her, we want something the ignoring of which would be in some manner illegitimate on her part. Looking to provide her with a *reason* appears to be the only possibility.
>
> However, we do not want the claim "You have a reason not to do that" to be nothing more than an utterance licensed by our having taken the moral point of view, otherwise we've said little more than a reiteration of "You ought not do that," and we may still quite reasonably be ignored… So when premise (2) links "having a reason" with a moral "ought", it is intended to be something other than an institutional reason; it is what I have been calling up until now (rather deplorably) a "real" reason.
>
> One rejects and violates the rules of gladiatorial combat, one rejects and violates the rules of morality. Yet we do not think of them on a par. We invest the moral judgment with an extra authority, and it is this fugitive thought that we must try to nail down. The best place to seek the fugitive is in an account of non-institutional reasons. 2001: 44-5

Joyce thus identifies some aspects of contexts in which we make moral claims that suggest we use them to implicate external reasons statements: for we make moral judgments in such a way that we want them to have a certain 'oomph'. We want to engage those to whom we direct our moral judgments, to suggest that our judgments 'cannot legitimately be ignored', that a felon ignores 'real reasons' in ignoring our injunctions. In conjunction with the understanding that moral claims are non-evaporable, what we end up implicating is that our moral judgments are reasons for action. Recall the question 'are you coming to dinner?' and the response 'I have to work'. I am implicating that I am not coming to dinner, despite the fact that the claim 'I have to work' does not *mean* 'I am not coming to dinner'. Now consider an analogy from moral discourse. Suppose someone asks me 'Do I have reason to rob the bank?' and I reply 'It is morally wrong!!' Because I (and perhaps my hearer as well) am making the comment particularly forcefully – as Joyce says – as if it cannot legitimately be ignored, I am implicating that my hearer has a reason to *refrain* from robbing the bank; even if, as I have argued in the main part of this chapter, one cannot say that something's being morally wrong *means* that there is a reason to refrain, or that a judgment about wrongness just is a judgment about reasons. If we typically use moral thought and talk to *implicate* judgments about external reasons (having granted that moral judgments do not *imply* them by virtue of conceptual connection), might moral error theory be vindicated? I hope to show, in the next section, that the answer is, 'no'.

Before I make this argument, however, it should be noted that, even if it is *typical* practice of users of moral discourse to implicate judgments about external reasons, it is far from obvious that users of moral discourse *always* do this (and perhaps this is what the error theorist needs? If so, error theory is already off to a poor start). While conversations like the one in the above paragraph are common, so are conversations (perhaps, between anti-rationalist philosophers or amoralists) like these, in which no implicating takes place. Suppose Andrew says to me 'I have to plagiarise. That is the only way I can pass my assignment. I'm sure I can get away with it. Should I do it?' Suppose that I, given my convictions about a lack of a conceptual connection between moral standards and reasons, reluctantly say, 'look Andrew, maybe you will get away with it. Maybe I can't provide you with a reason; except to say that I morally disapprove, and I will think less of you for doing this.' I have not used my moral judgment about the wrongness of Andrew's act to implicate any external reasons. I have even denied that I might provide Andrew with a reason at all (on the assumption that even *my disapproval and thinking less of him* would not count as a reason for him). But I have nevertheless earnestly voiced my moral judgment. The practice of using moral judgments to implicate judgments about external reasons, then, is not something that users of moral discourse do all the time. But I am happy to grant that they *typically* do, or do so *most of the time*. Given this, is moral discourse fatally flawed? Is moral error theory warranted?

3.3 Implications and error theory

Even if we do assume that we typically use moral claims to implicate external reasons statements, it is not the case that this can support Joyce's error theoretic conclusion. In order to argue convincingly for moral error theory – that the truth conditions of moral claims depends on something problematic – I believe one must show that *moral discourse itself* is responsible for committing its users to this problematic concept. But if all Joyce can show is that speakers typically use moral terms to implicate external reasons statements, then such a point cannot be demonstrated. What is shown, on the other hand, is that *speakers are responsible for committing moral discourse* to something problematic, not the other way round – this my contention for this section.

In response to this, an error theorist of Joyce's ilk may accept that *users of moral concepts*, not *moral concepts themselves* are to blame for the fact that moral discourse is problematic, but bite the bullet and say that this still supports a moral error theory: for the end result is that moral discourse is still hopelessly problematic. But I will now argue that, if the moral error theorist must resort to this, moral error theory is not a very convincing view.

Moral error theory is a revisionary view. What this means is that it seeks to take one of our most fundamental beliefs (our belief that our moral claims are at least sometimes true), and argue to a surprising conclusion: namely that none of these beliefs could ever be true. Error theory is a revisionary view because it proposes that our belief in the existence of true moral judgments should be revised or given up.[43] And certainly, if it is shown that the very *meaning* of moral terms commits us to something problematic, then a convincing case can be mounted that we give up on there being truthmakers for our moral claims. But if it is merely the case that we use moral terms to *implicate* something that turns out to be

[43] This is not to say that error theory leads necessarily to eliminativism: the recommendation that we cease to engage in moral thought and talk. All I mean to say here is that the error theorist is trying to say at least that we ought to abandon the belief that our moral claims can be true. Whether we should then treat moral thought and talk like a useful fiction (like the fictionalist believes), or cease moral thought and talk all together (as the eliminativist believes) is a separate issue.

problematic, then the moral error theorist's claim that we revise our belief that our moral claims could be true becomes utterly unconvincing, for there seems to be a more straightforward alternative: why should we not instead give up our practise of implicating that we do? After all, if we gave this up, nothing significant about the *meaning* of moral claims would be given up.

An analogy might be helpful in making this point. Take the term 'woman'. The meaning of 'woman' is simply: a female human. However, for the majority of earth's history, the word 'woman' has been used to implicate a number of (largely negative) things: somebody who is intellectually and perhaps even morally inferior to a man, somebody whose only use is bearing and rearing children. It is also quite likely that, for centuries, the majority of users of the term 'woman' (arguably, even women themselves too) made these implications whenever they used the term. We can imagine an example of a conversation from past times going as follows: 'I'm going to leave your list of expenses with your wife to calculate the total,' to which the other conversant replies, 'Are you sure that's a good idea? She is a woman, after all.' Thus 'woman-discourse' was commonly used with such implications running through and through. Over the last century, our beliefs about women have changed, such that implications like the above example are (hopefully) made far less frequently. We have come to discover that women *are* as intellectually and morally capable as men, and that a woman's usefulness to humanity can extend into the public as well as the private sphere. Given these revelations, what are we to revise? Given that none of the things that we used the term 'woman' to implicate turn out to be accurate, should we throw our hands up in the air and say that woman discourse is hopelessly flawed and become error theorists about 'woman-discourse'? Surely not. Surely, what we ought to revise is our practise of using the term *woman* to make the erroneous implications in the first place. It will turn out that we can still talk about women, still make women-claims, because the core meaning will remain unaffected. Implications are, after all, cancellable!

Similarly, if it turns out that moral speakers by and large have been using moral judgments to make problematic implications, we should revise this practise, rather than revising our belief that moral judgments can be true. To take the latter route of revision seems to put the philosophical blame on moral discourse where it is not deserved; rather than putting the blame on *ourselves*, the users of moral discourse, for making mistakes about how we have used moral terms. These mistakes we have made – these implications – not being part of the meaning of moral terms, are totally optional for us to use or not use as we please. If speaker A makes the claim ‘John morally ought not steal’, and uses the claim to imply the presence of an external reason of John’s, while speaker B makes the claim ‘John morally ought not steal’ without making any such implication, neither one has failed to participate in moral discourse. *Both* ways of speaking are permitted by moral discourse: moral concepts do not permit *only* the problematic way. Thus the error lies not with moral discourse, but with the completely *unforced* mistakes that users of moral discourse make.

At this point the error theorist may simply dig her heels in and argue that, to truly be participating in moral discourse, one *must* make the implication that speaker A makes. But it is hard to see that there could be any basis for making such a claim. Such an implication is not required by the meaning of moral terms, so why, exactly, should one have to make it? Different users of the same discourse should be able to differ on the implications they load into it, without being accused of using a different discourse all together (and the implausible claim to the contrary would indeed be an implication of this error theoretic line of argument).

Even if it is the case that most of us do use moral make implications about external reasons, I do not think this can support moral error theory, which is essentially the view that overturns our belief that our moral claims can be true. It seems much more sensible to claim that it is our practise of making problematic implications using moral terms, that should be overturned. Given that this practise is not required by anything about moral claims themselves, but rather by the whims and inchoate thoughts often responsible for the implications we make, it seems that the propensity of moral claims to be true should not be held hostage by these.

3.3 Conclusion

In this chapter I have tried to show that the rationalist conceptual claim is false. Moral judgments can be made, and can be true or false, in a way that has nothing to do with the judgments about external reasons – or any reasons at all. Nor will it help the error theorist to soften their claim about how moral discourse is committed to external reasons: in making this premise more plausible, the error theorist makes her conclusion less plausible. [44]

To reiterate, this is not to deny that people very often *do* take their moral judgments to provide them with reasons: all I have been arguing is that there is no *conceptual demand* to do so. Nor is it the case that people lack reasons to do what is, in fact, morally right. Indeed, I have suggested already that at least much of the time we do have reason to do our moral duty.

The conclusions of this chapter and the last are significant. Even though the notion of the external reason is seriously problematic, this need not render moral concepts problematic. There can be such things as true moral judgments, if we understand moral judgments along the lines carved out by the moral consideration claim.

Chapters 4 and 5 will constitute the more positive part of my project. Having defended a view about moral judgments that renders them safe from a certain brand of moral scepticism, I will go on to give an account of what the truthmakers of such judgments are. In other words, I will give accounts of what we would need to know, in order to know whether a moral judgment is true. Firstly, we would need to know what *welfare* consists in, and secondly, we would need to know what the unifying feature of *positive or negative relations* between morally evaluated things, and welfare, is. Chapters 4 and 5 will deal with these questions.

[44] The arguments I have made in the above section are similar in some respects to the arguments of Simon Kirchin, who notes this tension in moral error theory: the more plausible one makes one part of the error theory argument, the less plausible other parts become, and vice versa. See Kirchin, 2010.

4:
Welfare

1. What do we want from a Theory of Welfare?

1.1 Parfit's tripartite distinction

In this chapter, I will discuss the topic of human welfare. Sections 1 and 2 will be occupied by discussions of *theories* of welfare, while the last section, 3, will briefly discuss some common *sources* of welfare. It is important to understand the difference between these two things: the *sources* of one's welfare are those things that one must get or have in one's life in order for one to be well-off. Things like friendship, achievement, a loving community, physical heath, or education may all plausibly be thought of as sources of welfare. A theory of welfare, on the other hand, attempts to say *why* it is that such things contribute to one's welfare, and what, after all, welfare is.

Parfit (1984: 493-502) separates theories of welfare into three broad categories: hedonism, desire-satisfactionism, and objective list theories. Hedonists claim that one's welfare consists in the amount of pleasure one has in one's life, and desire-satisfaction theories claim that one's welfare consists in the satisfaction of desires. Objective list theories claim that there are some things that are good for a person *regardless* of how this might be connected to pleasure, or desire satisfaction. Although some (see Arneson 1999: 114-7) have found Parfit's tripartite distinction to be problematic for various reasons, it is, for the most part, sufficient for my purposes.

In this chapter, I will do three things. In this first section, I will outline some desiderata that I think a plausible theory of welfare needs to fulfil. It is important that such a theory of welfare meets these desiderata such that it is nicely consistent with the goals of my metaethical project. I will explain how one type of theory, objective list theories, do not fit the given desiderata, but how hedonism or desire satisfactionism can fit perfectly well. In section 2, I shall offer my own preferred theory of welfare, a hybrid of desire satisfactionism and hedonism. Though I believe it to be an interesting theory with certain advantages, the reader should keep in mind that there may be other, more well-established hedonist and desire satisfaction theories that can fit with my overall project. Section 3 will discuss the sources of welfare, as shown in a number of studies of positive psychology. It will be shown that there are, in all likelihood, widely shared goods across humanity.

1.2 A theory of welfare and reasons internalism

There are two desirable features I believe a plausible theory of welfare must have. The first is that it honours *reasons internalism*, which, as I have argued in chapter 2, is the more plausible theory of practical reason. The second feature is opinion-independence: a plausible theory of welfare must entail that an agent's own well-being is somewhat independent of her opinion on the matter. I will explain each of these features in turn.

What would it mean for a theory to honour reasons internalism? What do theories of welfare have to do with practical reason at all? This will take a bit of unpacking. For a start, it means that a theory of welfare must satisfy what Rosati (1995: 300) calls 'the internalist requirement'. To avoid confusion with other doctrines mentioned in previous chapters, I will refer to this internalist requirement as the 'welfare internalist requirement'. The welfare internalist requirement holds that a theory of welfare has to define welfare such that welfare strikes most agents as something worth caring about, something capable of motivating us (the question of what is meant by most agents will be addressed shortly). If you tell me that X is good for me, or counts toward my welfare, then if I have an adequate understanding of what 'welfare' is according to your theory, I should be (somewhat) motivated to pursue X. If this welfare internalist requirement is apt (and I believe it is), then this reveals something else about judgments about welfare: that we take them to be indicative, or implying, judgments about normative practical reasons. Recall, in chapter 1, that it was said that one's judgments about one's normative practical reasons necessarily motivate, and that (therefore), if a judgment about what I should do motivates me, this indicates that I am taking such a judgment to indicate a normative practical reason for acting. So too, in the case of welfare. Because a judgment that something is in my well-being is likely to motivate me, then it is also likely the case that my judgments about what is in my well-being imply judgments about what I have normative practical reason to do. So then, one's theory of welfare (if the internalist requirement is correct) will have implications for what kind of practical reasons one thinks there are. It is important, then for my purposes, that the theories of welfare I can accept are theories that only imply internal reasons, and avoid implying external reasons.

Now, obviously it would be far too stringent a requirement of a theory of welfare that judgments about welfare were capable of motivating *any* kind of agent. We can imagine an agent so depressed, suicidal or numb such that considerations about her welfare would not motivate her no matter how good your theory of welfare was. Theories of welfare ought not be required to motivate agents like this: agents who, we could say *do not care about themselves* or have no self-love. But we can demand of a theory of welfare that it ought to motivate ordinary agents who do care about themselves, and who care about their existence – in short, who care about their well-being. Given that one's own welfare is something that most agents care about, *de dicto*, a good theory of welfare should define welfare such that the interest and care about it is retained. If you present a narrow theory of welfare – say, a hedonistic theory in which you say my welfare depends on my experiencing a certain sensation a lot – it counts against your theory that I am left cold by it, or end up thinking that, if this is all wellbeing is, then maybe it isn't that important after all. Our general interest and motivation toward doing what is in our interests or toward our wellbeing entails that judgments about our well-being are judgments about what we have practical reason to do.[45] The most plausible theory of welfare, then, should be in harmony with the most plausible theory of practical rationality.

[45] I admit that the case of our depressed, suicidal agent raises many difficult questions about the relationship between wellbeing and practical reasons. Although my contention is that all judgments about what counts toward wellbeing entail judgments about what one has practical reason to do (at least, pro tanto), the reverse is not the case: there could be judgments about what one has reason to do that do not entail judgments about what contributes toward welfare. Our depressed and suicidal agent may make the judgment that she has a

This leads to an important point: my rejection of objective list theories (for examples, see Hurka 1993, Finnis 1980, 1983). Objective list theories attract a variety of criticisms, but the most important criticism I have of objective list theories is that they imply the existence of external reasons. Recall, objective list theories claim that it can contribute to an individual's wellbeing or interest to pursue certain things, regardless of how they may be related to her desires or her enjoyment/pleasure. One's desires, and one's propensity to enjoy or take pleasure in things constitute two fundamental elements in Bernard Williams' *subjective motivational set* (I shall assume that propensity to experience pleasure, or another form of positive affect, is what Williams might be referring to when he speaks of one's 'patterns of emotional reaction', 1979: 20). In other words, then, objective list theories imply that one has reason to pursue things, regardless of their relationship to elements of the subjective motivational set. Thus, objective list theories imply the existence of external reasons, and for this reason I reject objective list theories from the outset.

Before I continue outlining some problems with objective list theories, it is worth pausing to see if the same worry that I've just mentioned might not apply to hedonist theories as well, particularly to hedonist theories like Bentham's, which define pleasure as a particular sensation (Bentham 2008 [1789]: 99-103, Sobel 2002: 240-1). What if a given individual doesn't want the 'pleasure-sensation' of Bentham's hedonism? Wouldn't it be just as external-reasons-implying, then, to insist that the experience of such a sensation is good for an individual whether she desires it or not? This is somewhat of a moot point, given that I will go on to argue that the view of pleasure as a sensation is inadequate. There are alternative ways that we can define pleasure, which makes it more plausibly an element in Bernard Williams' subjective motivational set: something that is in and of itself capable of motivating us. In short, hedonism does not imply external reasons, so long as it defines pleasure, or the propensity for pleasure correctly: as an element in the motivational set (a 'pattern of emotional reaction') capable of motivating us, and even forming the desires we have. I will elaborate on such a view about pleasure throughout section 2.

reason to kill herself, without this entailing a judgment that doing so would count toward her wellbeing (quite the opposite!) However, this does not undermine the main point here: that, for agents who *do care about welfare*, judgments about their well-being motivate, and must therefore count as a sub-species of judgment about what they have reason to do.

Other problems with objective list theories abound. One problem is that objective list theories, before they *can* even imply external reasons, might fail to meet the welfare internalist requirement in the first place. An objective list theorist may dogmatically claim that I ought to have certain goods in my life, and, depending on what she puts on her list, her claims may leave me completely unmotivated. Any theory according to which welfare turns out to be something that leaves us cold or alienated exhibits an obvious flaw. Another, closely related complaint about objective list theories is made by Sumner (1996: 45): objective lists may attempt to tell us *what* is good for us, but they, in general, do not do a very good job of saying *why* these things are good for us (this lack of explanation and justification explains our propensity to be 'left cold' by objective list pronouncements). In other words, objective list theories do not make very good *theories* of welfare, even if they may constitute plausible catalogues of the *sources* of welfare.

It might be objected that individuals are often mistaken about what is good for them: individuals often pursue loneliness when they should pursue friends, or pursue material gain when they should pursue true love, for instance. Why then, should an individual's failure to be motivated to pursue certain things count against the objective list theory? An objection like this, however, misses the point. The welfare internalist requirement does not say that individuals who care about welfare should be motivated to pursue sources of welfare that are the real sources. What the internalist requirement entails is that the *explanation* a given theory of welfare puts forth as to *why* the sources of welfare are what they are, will motivate the agent to pursue those sources (assuming the agent has properly understood the theory), if the theory is a good one. The only explanation the objective list theory gives (given its conflation of an inventory of sources of welfare into a theory is) as to why certain things are sources of welfare is 'they just are!' Unsurprisingly then, objective theories are likely to leave most of us cold.

1.3 Opinion independence and similar sources of welfare

The second criterion that I think a plausible theory of welfare should meet is that one can be well off or badly off, despite one's opinion to the contrary. Similarly, one may also be wrong about the sources of one's welfare: a heroin addict may claim that heroin is a source of wellbeing for her, and we should be able to say that she (or anyone else in a similar position) is wrong about this.

Although I will be defending what might be called a hybrid view of hedonism and desire satisfactionism, it should be made clear from the outset that neither of these views need imply that one's welfare is (purely) a matter of actual individual opinion. Suppose a version of hedonism is true. An individual believes she is well-off, but leads a life with very little pleasure. According to hedonism, this individual's opinion about her own welfare is inaccurate. Suppose a version of desire satisfactionism is true. On the face of it, it might be difficult to see how such a theory would not give full weight to an individual's opinion of her wellbeing: supposedly the individual knows what her desires are, and knows whether they are being satisfied. But desire-satisfaction accounts can make an important move here: they can hold that the satisfaction of an individual's *actual* desires is not what makes her well off, but rather, the desires that she *would* have if she were a fully informed, or an otherwise idealised, version of herself (see Rosati, 1995 for detailed discussion about these theories. See also Rawls, 1971: ch 7, and Railton 1986). Accordingly, we can see that this version of desire satisfactionism is consistent with the claims that an individual can be mistaken both about the quality of her life, and about the sources of her wellbeing. Another way desire satisfaction theories can be shown to be consistent with the claim that an individual's wellbeing is somewhat independent of her opinions can be seen in the possibility that we may not necessarily always know what we desire – some desires could be subconscious.

There is one small qualification to be made. Some theories hold that *satisfaction with one's life* is central to welfare (a view defended by Sumner 1996: 140-171). I should say at this point that I too consider satisfaction with one's life, and other pro-attitudes toward one's life, to be part of wellbeing. Thus, I believe a component of one's welfare is dependent on one's opinion of one's welfare, but *only a component*. Consider two individuals, Bob and Jane. Bob has many pleasures in life, and many of his desires satisfied, but has no particular opinions about how his life is going. Jane has the same amount of pleasures, and the same amount of desires satisfied, and also thinks that her life is going well: she has a degree of satisfaction with her life. Both Bob and Jane's lives are going pretty well, but Jane's is going better.

On top of being able to imply (a large degree of) opinion-independence about welfare, hedonism and desire satisfactionism are also consistent with the claim that there can be such things as 'absolute goods' for humans. That is, there are some deep similarities in sources of welfare across the human race. It is important to note from the outset that one does not have to be an objective list theorist in order to get the result that there are some things that are definitively good for us (regardless of our opinion, for the most part) and other things that are definitively bad. Hedonists and desire satisfaction theorists can both claim that things like friendship, close family relationships, challenging pursuits and achievements, and even the display of certain virtues are good for at least most people, while isolation, a lack of challenge or stimulation, and selfishness are bad for people (in fact this summarises the research I will be outlining in section 3). A hedonist will say that these 'good' things are good because they produce pleasure, and a desire satisfaction theorist will say that these things are good because they fulfil deep desires that we (or the informed versions of ourselves) have.

In order for my theory of welfare to fit nicely into my broader metaethical framework, opinion-independence is a desirable attribute for my theory of welfare to have. Recall, from chapter 3, the claim that morality is centrally concerned with welfare. Moral facts, whatever they are (more on this in chapter 5), will be at least partially underwritten by facts about welfare. If welfare is something that is entirely dependent on individual opinion, this opens the door for a great deal of relativism about moral facts: a result that I said, at the outset, I wish to avoid.

To see why this is the case, think of the following example. Suppose I have the opinion that it is essential to my wellbeing to steal native land from a tribe of Aborigines, so that I can build my mansion on it (suppose also that I have the opinion that it will make them no worse off for me to steal from them). Suppose the Aborigines have the opinion that it is central to their wellbeing to keep the land. If morality seeks to protect or promote welfare, and welfare is opinion-dependent in this way, then it may be hard to avoid the conclusion that stealing land from the Aborigines is morally right relative to me, and stealing land from the Aborigines is wrong relative to them. Consider another example, in which even *agreement* about the welfare of an individual implies relativism. Suppose Fred is a paedophile, who convinces a young victim Becky, that it is in her interests or wellbeing to be intimate with him. Fred thinks a sexual relationship with Becky is good for him, and Becky believes this to be the case too. Suppose a third party, Bill, finds out about this, and condemns the situation, saying that what is happening is harmful to Becky and therefore morally wrong. But if individual welfare is a matter of opinion, it is hard to see how this wouldn't imply that the sexual relationship is morally right relative to Fred and Becky, but wrong relative to Bill. To avoid results like these, one must avoid claiming that one's own welfare, and the sources of it, is entirely a matter of individual opinion.

What if individual wellbeing is not dependent on opinion, but nevertheless varies from individual to individual? What if a full-information version of desire satisfactionism is true, and you and I have different (informed) desires? Furthermore, what if such desires lead to clashes of interest? Given that this is entirely plausible (even if there are other desires we have in common) what are the implications of this for morality? I will not be able to answer this question fully until Chapter 5. But what can be said at this point is that, if this occurs, we have not a threat of relativism, but a threat of *indeterminacy* in moral facts. This threat may turn out not to be real: perhaps there is a clash of interests, but I stand to lose more than you do if things don't go my way: in which case morality might (but not *necessarily*)[46] side with me, and there will be a moral fact of the matter: namely that things ought to go my way. However, suppose that this is not the case. Suppose that there is a clash of interests, and no fair way of resolving it. Here, we would have a case of ethical indeterminacy: there is no fact of the matter about what morally ought to happen.

That different individuals' welfare may be at least to some extent be realised differently, then, may imply some degree of ethical indeterminacy. However, this is no great problem for me. Any reasonable ethicist ought to concede that there may be *some* degree of ethical indeterminacy. Of course, if there is *too much* ethical indeterminacy, this may amount to an undermining of morality itself, and may warrant error theory. But this undesirable extent of indeterminacy is unlikely to result, given that I will be arguing that there are some broad *similarities* in how welfare is secured for human beings. Although there may be differences between individuals with regards to how wellbeing is brought about (given different

[46] I say 'not necessarily' because considerations to do with desert or fairness might come into play: perhaps I will benefit more if things go my way, but the fact that you *deserve* things to go your way may entail morality's siding with you: perhaps I would get more pleasure out of owning your house than you would, but you nevertheless *deserve* to own your house more than I do, because you worked for it and paid for it.

personality types, interests, and aptitudes), the deep similarities, I hope it will be obvious, are far more significant than these individual differences. Once again, the full significance of this will not be discussed until chapter 5, but it is desirable to bring this issue to attention at this stage.

1.4 Welfare and the satisfaction of the subjective motivational set

In the next section, I will outline my preferred theory of welfare, a hybrid of hedonism and desire satisfactionism. One way of summarising my view, (and of harmonising it with the claim that judgments about welfare entail judgments about reasons) is to say that welfare consists in the satisfaction of one's subjective motivational set. In other words, the more elements of one's subjective motivational set that there are satisfied, the more well-off one will be.

There are three memorable elements that Bernard Williams discusses: 'patterns of emotional reaction', 'desires' and 'values' or dispositions of evaluation (1979: 20). Williams also mentions goals and personal loyalties, but these may be captured under the headings of desires or values. In what follows, I will be claiming that the satisfaction of pleasures, the experience of satisfaction and endorsement toward one's life and pleasures, and the satisfaction of desires, all contribute to wellbeing. I believe all these things discussed cover these elements of the motivational set mentioned by Williams: desires, 'patterns of emotional reaction' and values. Admittedly, the question of what values are is a complex one that has generated much debate. In such a debate, I side with those who regard values as a desire like state, or a pro-attitude. For example, if one desires that the world be at peace, then world peace could be said to be one of one's values. Similarly, if one has a pro-attitude of endorsement toward the fact that one is a generous person, then generosity could be said to be among one's values. The opposing view has it that values are *judgments* rather than the kind of non-cognitive attitudes I have claimed. I will not be able to defend my view about what values are, as the scope for this chapter just doesn't allow me to do such a debate justice. I have to rest content with the fact that I am in good company (Lewis 1989, Blackburn, 1998: 9-13, Prinz, 2007).

To round off this first section, then, the theory of wellbeing that I regard as correct, and the theory of practical reasons that I regard as correct, both say that one ought to satisfy one's subjective motivational set: that there is such a harmony with practical rationality is a desirable feature for a theory of welfare to have. I regard hedonism and desire satisfactionism to be the theories of welfare that make most sense of the notion that it is in one's wellbeing to satisfy elements of the subjective motivational set, and a hybrid of the two views, ideal. Once again though, there is no need to adopt the exact theory that I offer, in order to have a theory compatible with my aims of providing both a reasons-internalism honouring, and opinion-independent theory of welfare. If the reader sees insurmountable problems with parts of the theory I offer, then, this ought not to be taken to be bad news for the broader project.

2. My Preferred theory

2.1 Pleasure

Over this section I will outline my preferred theory of welfare. While the theory I present will not be able to address every possible objection, I hope to show through the course of the presentation of it that it has some key advantages. I will argue that a combination of pleasure and desire satisfaction is central to wellbeing, and will argue that the experience of pleasure is in fact equivalent to one type of desire satisfaction. In this section I will discuss some difficulties of defining pleasure, and settle on my preferred definition.

One of the most difficult challenges hedonist theories of welfare face is the challenge of defining pleasure. Hedonists hold that pleasure is the only thing that is intrinsically valuable for one's life, but competing definitions of pleasure, each with their drawbacks, make this seem like a difficult statement to hold to. A simple view is that pleasure is a distinct sensation: something like a tickle, or a feeling of warmth, or something like the feeling we have when getting our back massaged. But such a conception of pleasure faces the well-known heterogeneity problem (Sobel 2002: 241, Feldman 2004: 79, Mason 2007: 380). The feeling of drinking a cold lemonade, the relaxing feeling of sinking into a warm bath, the feeling of walking through the park with a loved one, and the feeling of reading a stimulating book are all regarded as pleasures, yet the sensations involved in each activity are very different. The same could be said about different emotions that we experience as ‘positive’, yet which feel very different: tranquillity, excitement, amusement, fascination, or tenderness. This creates the need for an answer to the further question: what do all pleasurable sensations have in common such that they are all referred to as ‘pleasurable’?

A number of different answers have been given, and I will consider one from Feldman. One Feldman’s view is that what all pleasurable experiences have in common is that we have a pro-attitude to them, or take *attitudinal pleasure* in them: ‘A person takes attitudinal pleasure in some state of affairs if he enjoys it, is pleased about it, is glad that it is happening, is delighted by it.’ (Feldman 2004: 56). Attitudinal pleasure must have an object, and this object may be a sensation, an experience, or a state of affairs, such as my friend’s being happy or my winning my race. A clearer way to put the distinction between sensory pleasure and attitudinal pleasure is to say that the kinds of things we can take attitudinal pleasure towards are things that can be expressed by propositions: being pleased ‘that I am successful’, or ‘that my friend is happy,’ or ‘that my team won’. Non-attitudinal, or sensory pleasure, on the other hand, are feelings without such propositional content: one simply has feelings of (for example) happiness, warmth, comfort, or exhiliration.

Initially Feldman’s view looks like a neat solution to the heterogeneity problem. Furthermore, the kind of hedonism that defines pleasure as a certain sensation may suffer from the same kinds of problems as an objective list theory: in dictating that a particular sensation is good for a person, regardless of whether she cares very much about it. But Feldman’s view avoids both these problems:

> For one, [attitudinal hedonism] implies that *being a pleasure* is not an intrinsic property of any sensation that in fact is a pleasure. Consider the cool, wet, tingly sensation you get when sipping a cold beer. Suppose it is a sensory pleasure. On the view I have proposed, that very sensation is a pleasure in virtue of the fact that you take [attitudinal] pleasure in it. If you had not taken [attitudinal] pleasure in it, it would not have been a pleasure. It’s being a pleasure *depends upon the fact that it stands in a certain relation to you* – the relation of being something in which you take intrinsic attitudinal pleasure. Feldman 2004: 80 (emphasis added)

What makes a sensation a pleasure has to do with its relation to our pro-attitudes: our being glad that, delighting in the fact that, we are having the given sensation.

I believe that Feldman has identified something important in his description of 'attitudinal pleasure'. But, while I believe that Feldman has identified in attitudinal pleasure a *particular type* of pleasure (one that I will be giving a privileged place to in my theory of welfare), I do not think he has successfully identified, in attitudinal pleasure, a common feature that all sensory pleasures have in common, as he claims. It is quite possible, quite intuitive, that there are experiences that we might consider sensory pleasures which we nevertheless loathe, or have a con-attitude toward. Consider the following example. Jeff is addicted to child pornography, and finds the sensations involved in consuming it enjoyable. Yet he has attitudes toward this activity that are quite negative: he believes he is doing something seriously morally wrong when he consumes child pornography, he is ashamed of himself when he does so, or views himself as a sicko or a sleaze. It would be most accurate, I think, to describe Jeff's situation as one in which he takes a sensory pleasure in the experience, but takes attitudinal *displeasure* in the fact *that* he engages in this experience (Davis, 2008 [1981]: 167 discusses a distinction like this in detail with use of similar examples). But if we are to say (as I think we should) that this is the correct way to describe Jeff's predicament, then Feldman's view that what makes a given sensation a pleasure is our having a pro attitude toward it cannot be right. Rather we should say that there are two types of pleasures: sensory and attitudinal, and that the question about what all such specimens of both kinds have in common such that they are pleasures remains unanswered.

The intuitive, but not very informative, thought is that what all sensory pleasures have in common is that they are somehow liked or enjoyed (we would say that Jeff's child pornography consumption is a sensory pleasure for him because he, in some sense, likes it). Feldman's way of making sense of this 'enjoying' or 'liking' is to point to the presence of a pro-attitude, but we have seen that this is inadequate. Pinning down what this 'liking' or 'enjoying' consists in raises the same heterogeneity problem all over again. The experience of 'liking' or 'enjoying' my golf game will look and feel very different to my experience of 'liking' or 'enjoying' a hot warm bath. One can say that I am liking or enjoying my golf game if I am vigorously engaging with it: looking forward to the next hole or eagerly concentrating on my next shot. But there is nothing 'vigorous' about my enjoyment of a warm bath: my enjoyment in this case probably consists of a relaxed and sleepy state, a tendency to smile as I feel the hot water dilate my blood vessels. And while the feelings of disliking the golf game or the bath are indeed different to the liking of each them, the two instances of disliking are as different to each other as the instances of liking are to each other. My not liking a golf game may consist of a feeling of boredom, irritation at my constant failure to hit straight, and an experience of wanting to be doing something else. On the other hand, my experience of disliking a hot bath consists perhaps a claustrophobic feeling when I am submerged in heat, or when I am wet, or both. In summary, to say that what all sensory pleasures have in common is that they are liked simply raises the question all over again: there are perhaps as many different varieties of liking or enjoying something as there are the sensations liked or enjoyed.

Given these problems, the view of pleasure that I will be adopting is the view that what all pleasures have in common is that they bear a certain relationship to desires. There are two kinds of pleasure, sensory and non-sensory (or, attitudinal), and what defines both of them is a (different) relationship each of them have to desires. But before I explain what this relationship is, I must talk a little bit about desires, and desire satisfactionism as a view about welfare.

2.2 Desires

I take the category of desires to include anything from the smallest, most remote wish, to the most eagerly sought after intention or goal.[47] A desire-satisfaction view of welfare says that our lives are well off insofar as our desires are satisfied.

From the outset, I would like to draw a distinction between two types of desires: experience-implying desires, and non-experience implying desires. An experience-implying desire is a desire the satisfaction of which involves an experience: by which I mean any sort of conscious state, anything experience that has a certain phenomenology accessible to the subject to whom the desire belongs. My desire for the delicious taste of chocolate, my desire to see my friend and spend time with her, and my desire to know what it is like to endure a marathon, or to know what it is like to intimidate someone and make them afraid of me, all have in common that the satisfaction of them involves some sort of experience: the chocolaty taste, the elated feelings and satisfaction I get when catching up with my

[47] Keller: 2004, 32-4 discusses the distinction between mere desires, and desires which also constitute goals, and I more or less accept his terminology.

friend and hearing her stories, the feeling of exhaustion during the marathon, or the sensations and thoughts I might experience on seeing someone cower before me. A non-experience-implying-desire is a desire, the satisfaction of which need not involve any experience on my part. The clearest example of desires like these, are desires for states of affairs that we might remain ignorant of even after they have come to pass. My desire for my child to be happy might be an example of such a desire: suppose my child is very secretive, very hard to read, or hard to communicate with, such that I am not able to tell whether she is happy or not. Nevertheless, *if* my child is, in fact, happy, then my desire is satisfied, despite my lack of awareness of the fact. Of course, even if I *was* aware of my child's being happy, my desire would still be classified as a non-experience-implying desire: for my desire is not to have the *belief* that my child is happy, but *for my child to be happy*. The former is an experience, whereas the satisfaction of the latter need not involve any mental state on my part: indeed, examples of such desires can include desires about what happens after one's own death.

Besides the above distinction, there is a further distinction to be made: one between sensory and non-sensory desires. Some experience-implying desires we have are sensory: the objects of these experiences are *sensations*. Typical examples of this kind of desire include the desire to go surfing, the desire for chocolate, the desire for sex, the desire for a massage, or the desire for a walk in the country. What is typically being desired in these examples are sensations: the roller-coaster-like feeling of being on that board as one speeds through the barrels, the taste and texture of the chocolate in one's mouth, the pleasures of the physical intimacy and the massage, and the colourful sights and varied sounds of the countryside. But not all desires are desires for sensations. The desire for my children to be happy, the desire for me to be satisfied with my life, or the desire to avoid actions that I believe to be morally wrong are not desired for the sake of any *sensation* that they produce: sensations are not the objects. It is merely states of affairs that are desired.[48] How does this distinction between sensory and non-sensory desires fit into the distinction between experience and non-experience implying desires? Sensory desires, obviously, can only ever

[48] This is not to say that states of affairs are *not* desired in sensory desires: the desire for the taste of chocolate presumably involves my desiring a state of affairs in which I am eating chocolate. What differentiates sensory from non-sensory desires is that the former are desires for a state of affairs *because a certain sensation is involved*, whereas the desires for states of affairs in non-sensory desires do not have this feature.

be experience-implying the desires: the desire for a sensation is the desire for just one kind of experience. Non-sensory desires may be either experience-implying or non-experience implying. My desire for my children to be happy, for instance, is a non-sensory non-experience implying desire. But my desire to be satisfied with my life, or to avoid actions that I believe to be morally wrong, are non-sensory experience-implying desires.

I would like to say that desires of all these kinds: both experience-implying, and non-experience-implying, sensory and non-sensory, are central to welfare. Such a broad view has particular advantages. Firstly, it respects the intuition highlighted by Nozick in his 'experience machine' thought experiment, that our well-being does not just consist in mental states. On a purely hedonistic account of welfare, desires about what happens in the world are rendered irrational, irrelevant to one's welfare; not worth worrying about if one's life is to go well. But this goes against some strongly held intuitions of ours (Arneson, 1999: 121, Sobel 2002: 244); intuitions which are respected by desire satisfaction theories.

However, these intuitions are not without their challenges. At this point it is worth looking at some challenges that confront the desire-satisfaction theory account of well-being. The following passage from Arneson brings to mind the first problem: as he calls it, the problem of 'non-prudential desires'.

> I might listen to a televised plea for famine relief, and form the desire to aid distant starving strangers, without myself thinking (and without its being plausible for anyone else to think) that the fulfilment of this desire would in any way make my life go better. So one needs to restrict somehow the class of basic desires whose fulfilment contributes to well-being. It will not do to stipulate that each agent determines for herself which of her basic desires bear on her well-being. It will not do to stipulate that each agent determines for herself which of her basic desires bear on her well-being. Surely an agent could make a mistake in making this determination, and we need some way of deciding when a mistake occurs… If we say that she divides her basic desires into two piles, those whose satisfaction would contribute to her well-being and the rest, our account is rendered viciously circular, and requires that we already have an idea of well-being independent of desire fulfilment. 1999: 124.

We must be careful to interpret Arneson's objection properly, lest we be distracted by the problematic aspects of the example he raises to make this point: for it is not clear why it is implausible to regard my responding to famine relief as a desire the satisfaction of which does not contribute to my welfare. Surely it cannot be because it is an example of a non-experience-implying desire, for Arneson indicates (ibid: 121) that he views the idea that states of the world (irrespective of an agent's experience of them) impacting an individual's welfare is a strength of any wellbeing account that implies this. Nor, could we say, it is the fact that the desire is altruistic, or involves me giving up possessions that belong to me. If my desire is strong enough, if I really care that the starving people are looked after, and if my money means little to me in comparison, then there is every reason to think that giving to the famine relief appeal *contributes* to my welfare. But let us not get distracted by the details of the example. The point is that it is reasonable to think that there are *some instances* in which we can desire something, the satisfaction of which would constitute a self-sacrifice. And the very notion of self-sacrifice involves the notion of *sacrifice of wellbeing*. Because it is possible for us to have desires that involve self-sacrifice, then desire-satisfactionism cannot be the correct account of welfare.

Arneson regards this 'problem of non-prudential desires' as 'intractable' (ibid, 18), but I think he has overlooked a fairly simple solution to this problem, which renders desire satisfactionism immune from any objections along these lines. The desire-satisfaction theorist can hold that it is possible to have a desire, the satisfaction of which, will involve self-sacrifice, without giving up the notion that well-being consists in the satisfaction of desires. Suppose I desire to save your life, though I know full-well that my effort to do so will cripple me or kill me – it will involve me throwing myself in the path of an oncoming car that is about to hit you. Suppose that your well-being is not that important to me, such that me saving you is not likely to contribute to my wellbeing that much, and suppose further that I know you to be an ungrateful person who will not likely benefit me through the return of good deeds if I survive the ordeal. Nevertheless, I am overcome by a desire to save you, and so I throw myself into the path of the oncoming car. I save you, but I become a paraplegic and am never able to walk again. Now, the desire satisfaction theorist is committed to saying that my satisfying my desire of saving you does contribute to my welfare in some respect. But this does not at all preclude my saving you as being a sacrifice of my well-being, *overall*. Suppose my entire life and the majority of my desires revolve around a love of using my body. Suppose I love to play all sorts of sports, and all sorts of musical instruments, in which I enjoy seeing all the ways I can use my limbs and muscles to make things happen. Now that I am a paraplegic, I will never be able to satisfy those desires again (at least, not in the way that I love and was so passionate about). I am in for a life of frequent, deep, desire frustration and unfulfilment as a result of fulfilling that one desire of saving you. My saving you, we might say, contributed to my welfare +5 points, in virtue of the fact that it satisfied a desire, but also contributed -500 points given all the desire frustration my sacrificial decision brings about. When we describe things in such a way, we can see that the notion of a non-prudential desire is perfectly compatible with the thesis that welfare consists in the satisfaction of desires. Of course, if Arneson means, by 'non-prudential desire', that there can be desires which *in and of themselves* contribute negatively to welfare (to use our numerical analogy: that there can be desires that directly subtract points from welfare, rather than merely entailing the frustration of other desires, which detracts from welfare), this will be discussed later.

Another aspect of desire satisfaction theories that might be challenged is the idea that all non-experience-implying desires really contribute to welfare after all. An example that Parfit raises certainly pushes our intuitions on this point: suppose I meet a stranger on a train, get talking to her, and begin to sympathise with her life aspirations. At the end of the journey, we part ways, and I wish her well (that is, I desire that her life goes well), but I know I will never see her again, and I don't. According to Parfit, it is implausible to think that the satisfaction of a desire like this would really make a difference to my life (1984: 468, 494). In my view, Parfit's example does not show what he thinks it shows. It would certainly make sense to say that the satisfaction of my desire for the stranger's life to go well may contribute to my welfare *very little* – the desire is, after all, probably a very weak and fleeting one. But to say that the satisfaction of this desire does not contribute *anything* to my welfare seems to undermine the very basis on which we think that more 'plausible' examples of non-experience-implying desire satisfaction *do* contribute to our welfare. We cannot say that the satisfaction of the desire that the stranger be well-off is not good for me simply due to its being a non-experience-implying desire. For this would entail that the desire for my spouse to be faithful, or for my colleagues to respect me, do not contribute to my welfare either. But surely we would not want to say *this*. Nor could we say, (as Parfit suggests on the desire satisfactionist's behalf: 494), that the reason this desire does not impact my well-being is because it is not a desire about my own life, a desire the outcome over which I have no control. For this would commit us to saying, on hearing that there has been an earthquake in Tokyo while my dearly loved friend is there, that the satisfaction of my desire that he not have been killed makes no difference to my welfare. Surely we would not accept this. Nor can we say that the satisfaction of this desire contributes nothing to my welfare because it is not strongly felt or long-lasting: all this entitles us to say is that the satisfaction of my desire contributes a little bit to my welfare, but not very much.

But Parfit raises another objection to desire satisfactionism, a more difficult objection, which I will not attempt to resolve until section 2.6. I will simply summarise it at this point, to be resolved further on. Suppose I give you an addictive substance, and thereby make you an addict. Every morning you wake up with an intense desire for this substance, and every day you satisfy this desire by taking it, for the rest of your life. The drug gives you no pleasure whatsoever (and let's also assume that the taking of it does not interfere with any other aspect of life, to put the example in cleanest form). By making you an addict, I have given you lots of extra desires, which are always satisfied, and have thereby made your life go better. But this, Parfit says, is implausible, and on this occasion I agree with him. I do, however, think that the desire satisfaction theory can get around this if some extra attention is paid to the question of what it is to have a desire, and what its relationship to pleasure and pain is. I will discuss this in the following subsections, and eventually address Parfit's worry in 2.6.

Parfit's concern points the way to a more general worry. It is quite intuitive to think that there are desires that are just not good for an individual to satisfy. If, to bring up Williams' gin drinking example from chapter 2 again, I desire to drink the clear liquid on the table that I think is gin (but is actually cleaning fluid) we would think that this is a bad desire for me to satisfy. Similarly, if I desire to buy shares in a company that I falsely believe will make me rich, but in actual fact will ruin me, this is another example of a desire that is plausibly bad for me to satisfy. Many puzzles like these, however, can be solved by simply making a distinction between intrinsic and extrinsic desires. Extrinsic desires are desires for things that (or desires to do things that) will help satisfy a further desire: an intrinsic desire for something, that we want for its own sake. In the gin-drinking example, my desire to drink the clear fluid is merely an extrinsic desire. My intrinsic desire is to quench my thirst and refresh myself. Similarly, the desire to invest money in the particular company and earn lots of money is an extrinsic desire: my intrinsic desire, the desire that I hope this former desire to accomplish, is for all those pleasures that my wealth could buy – my living in a fine house, walking in fine gardens, driving my favourite car, or going on countless skiing holidays. Clearly, extrinsic desires can be evaluated as contributing to welfare and detracting from welfare insofar as they help us fulfil our *intrinsic* desires: the two extrinsic desires are bad for one to satisfy because (at least among other reasons), they frustrate the satisfaction of the intrinsic desires they are meant to fulfil. Desire satisfaction theories, then, have a way of saying that some desires are bad for a person to satisfy, without dropping the thesis that wellbeing consists in desire satisfaction: specifically, intrinsic desire satisfaction.

Nevertheless, a question persists as to whether even *intrinsic* desires can be bad for a person to satisfy. For it is possible for people to have intrinsic desires that are crazy, seemingly irrational, or seemingly harmful. I might have an intrinsic desire to be a grass-counter, or an intrinsic desire to poison myself or damage my friendships. How does the desire satisfaction view guard against the counterintuitive results that the satisfaction of *these* desires contribute positively to well-being? The standard response from desire satisfactionism is to modify itself such that it is not the satisfaction of *our actual* desires that contribute to our well-being, but the satisfaction of desires that a fully informed version of ourselves would have (Arneson 1999, Rosati 1995, Sobel 2002). However, such 'full information' accounts of desire satisfaction have problems of their own. Sobel ibid: 249) points out that the acquisition of information can in and of itself contribute to our welfare: the process of learning about the history of Europe, learning quantum physics, or learning a physical skill could all be plausibly counted as processes that contribute to welfare. But presumably a fully informed version of myself by definition would not need to undergo such learning. With these considerations in mind, Peter Railton's reformulation of the 'full information account' looks attractive: 'an individual's good consists in what he would want himself to want, or to pursue, were he to contemplate his present situation from a standpoint fully and vividly informed about himself and his circumstances, and entirely free of cognitive error or lapses of instrumental rationality.' (Railton 1986: 16). But there are problems even with this formulation. Sobel writes:

> There are conceptual problems in combining all of the experiences that could be one's own such that one can then rank the relative value of the various options that one could experience. Consider trying to experience what each of one's possible first kisses would have been like. What we are expected to

> be able to do is to experience first-hand one way our first kiss might have gone, remember this experience such that we remain fully appreciative of all the valuable aspects of the experience, and then try to fully and accurately experience another way our first kiss would have felt had it been the first kiss that we experienced, and so on for all the ways our first kiss might have gone. But the fact that one has experienced and clearly remembered several such kisses would seem, at least in many people's cases, to cause the experience of some possible first kisses to differ importantly in ways that distort the experience from what it would have been like to actually have it be one's first kiss… Some decisions take on the phenomenology that they do because they radically affect the structure of one's life narrative… The decision to attempt to become a philosopher or a physician would be experienced differently than it would be by us if the decision were made by someone who suspected she would eventually get around to trying both… Someone who exhibited the heroism of Henry V and later cowered in fear at slight challenges will experience a different form of diminishment from the coward who never knew what it was to be brave. The same cowardly acts likely will be experienced differently because of the different narrative that the acts fit into… In demanding that the idealised agent be able to fully experience these aspects of nonidealised agents like us, we demand that she do incompatible things. The idealised agent is to be affected as we can by fundamental life choices that close off some life paths and shape who we are and what we can make of our lives, but at the same time she is to be practically immortal. (Sobel, 2002:250-1)

To summarise: the phenomenology of our experiences contributes something to our well-being. But, the phenomenology of being a non-idealised self is just too different from the phenomenology of an idealised version of oneself for her stances to be too indicative of one's welfare.

In the following section, I will try and provide a way of my own that the desire satisfaction theory might distinguish between 'good' intrinsic desires, and 'bad' intrinsic desires. This way has much to do with understanding the relationship between desire and pleasure that I will present. In the theory of welfare that follows, I will try to preserve the attractive elements of both hedonism and desire satisfactionism.

2.3 Which sensory desires contribute to welfare?

As I have said above, I intend to give both hedonism and desire satisfactionism central roles in my theory of welfare. The task of this section and the next is to answer the question: which desires contribute to welfare, and which do not? Spelling out the relationship between pleasure and desire will help answer this question. It will also be important that the question is answered in such ways so as not to reintroduce the same problems that confront objective list theories: if one is to say that the satisfaction of some desires and not others contributes to welfare, one always runs this risk.

Recall that just as there are sensory and non-sensory desires, there are sensory and non-sensory (or rather, 'attitudinal') pleasures. I will now spell out what I think is a plausible relationship between sensory desire and sensory pleasures. With regard to sensory pleasures, my contention is that what makes them pleasures is that they are in fact desired, and that the initial experiences of such pleasures bring such desires for them into being. Mill hinted at such a hypothesis (according to Sobel 2002: 245-6). If an experience of a sensation counts as a pleasure for an individual, the individual will then develop a desire for subsequent experiences of that sensation. This is not quite the same as saying that pleasure is something an individual wants to *continue* in the moment of experiencing them. Some sensual pleasures might be so intense that the continuation of them for more than a few moments might amount to a form of pain (an orgasm might be an example of this). Similarly, some sensual pleasures deliver diminishing marginal returns the longer they go for (the pleasure experienced from a hot bath might be an example of this: the first five minutes are glorious, but every minute after that is less pleasing, as the body gets acclimatised to the water). What I am claiming is that, if a sensation is a pleasure for an individual, the individual will desire to have the experience again at a later point in time. Similarly, the satisfaction of our desires for sensations is constituted by such pleasurable experiences: experiences which generate further desires, and so on. To be clearest, we can formulate the idea the following way: we desire sensations because we believe (even if just subconsciously) that they will be instances of pleasure, P. And something is an instance of P if the experience of it generates further desires for it.

The plausibility of this account of the relationship between pleasure and desire can be demonstrated when we consider how our sensory desires change over time. When I was a child, I tasted an olive, and disliked the experience intensely: that is to say, I found it unpleasurable. Accordingly, I had no urge to taste an olive again. Sometime in early adulthood, I tried an olive to see if I would like it this time round, and I did. What changed between my first experience and my second, such that the first experience counted as a displeasure and the second experience counted as a pleasure? Surely what changed was the fact that, after the second tasting, I desired to eat more olives (if not immediately, then at some later point in the future). Similarly, as a teenager, I used to enjoy playing Super Mario on the Nintendo 64 – it was a pleasure for me. But one day, I suddenly realised it was not giving me that much pleasure. What does such a realisation amount to if not the realisation: 'I don't want to do this anymore?' Now, in the cases of both the olive eating and the Super Mario playing, there may have been phenomenological differences which also to some extent explain the difference between liking and not liking the sensation in question (maybe the sensation of eating the olive as a child was in and of itself slightly different to the sensation of eating it as an adult, given a change in the configuration of my taste buds). But once again, as was said toward the end of 2.1, we cannot define pleasure on the basis of phenomenology, since my liking the Mario game, and my liking the olive, are themselves very phenomenologically different. But surely what liking the Mario game and liking the olive have in common is their dictation of my desires: that they cause me to desire the experiences of playing/eating again is what makes them both pleasures.

Accordingly, the objects of our sensory desires are sensory pleasures. When we desire a sensation, it is because we believe the sensation to be pleasurable: that is, we believe it is the kind of sensation that, as we experience it, we would want to experience again. This claim also gives us a fairly easy way of determining which sensory desires are good for us to fulfil, and which sensory desires are bad for us to fulfil. Sensory desires are good for us to fulfil if they are based on *true beliefs* about the pleasures they will bring. If I have a desire for sensation X, and sensation X will turn out not to be pleasurable for me, then my desire for sensation X is a bad desire for me to satisfy. One should not think that this doctrine is just another dogmatic objective-list like doctrine, which claims that pleasure is good for an individual whether or not the individual cares about it. For what I am claiming is that the objects of our sensory desires *are in fact sensory pleasures*. So the absence of pleasure from an activity that we think will bring it constitutes a kind of being mistaken about what we want: I thought I wanted sensation X, but when I got it, I discovered that it is not what I wanted. Nor am I claiming that pleasures are the only thing that we in fact desire. I am only claiming that sensory pleasures are the objects of *sensory desires*, not all desires.

There is a second way that the fulfilment of sensory desires can turn out to make us worse off overall. Suppose we have a con-attitude, or attitudinal displeasure, toward the fulfilment of a sensory desire. Given that, as I will claim, the experience of attitudinal as well as sensory pleasure contributes to welfare (and that attitudinal displeasure as well as sensory displeasure detracts from it), the fulfilment of some sensory desires can make one worse off due to the extent of attitudinal displeasure one experiences in conjunction with it. This is the case, whether a con-attitude to the sensory pleasure was pleasant before the fulfilment of the desire for it, or if one's attitudinal displeasure only 'hits one' at the time pleasure is being experienced. Either way, attitudinal displeasure to the fulfilment of a sensory desire can render some sensory desires as bad for us to satisfy overall.

The most easily recognisable example of sensory desires that are bad for us to fulfil are desires for certain sexual experiences. Suppose Jane finds Brad very attractive – he's tall, dark, muscular, handsome, has a deep, sonorous voice. But suppose Jane also believes that Brad is a jerk. Although he can turn on the charm when he wants, Jane knows that, underneath it all, he has no respect for women and has a bad character in general. Nevertheless, Jane ends up fulfilling her desire to sleep with Brad, and experiences all the sensory pleasure that comes with it. But the morning after, she's filled with regret. She has intense attitudinal displeasure toward the fact that she had sex with someone who has no respect for her, or commitment to her. She suffers a temporary loss of respect for herself too. An example like this is one example of the fulfilment of one's sensory desires contributing negatively to one's welfare overall.

At this point one may ask why, exactly, attitudinal pleasure is so central to welfare. This question will not be fully answered until next section, in which I outline the relationship between non-sensory desires and attitudinal pleasure. But there is something that is worth pointing out now. Often, the experience of attitudinal displeasure at a sensory pleasure can impact the sensory pleasure itself, and make it less pleasurable. Take Jane's situation once again. Jane's attitudinal displeasure at sleeping with Brad is very likely to hamper the sensory pleasure involved, even if some discernible sensory pleasure remains. The least we could plausibly say is that Jane would experience *more* sensory pleasure if she had no attitudinal reservations about sleeping with Brad. This phenomenon of attitudes spreading themselves onto sensory experiences is familiar also with regard to the experience of eating food. If one eats chocolate 'guilt free', so to speak, with no negative attitudes toward the calories one is ingesting, it tastes 'better' than if one is eating it with the self-loathing that attends the knowledge that one is incapable of sticking to one's diet. One possible explanation for the fact that sensory pleasures can be somewhat muted by con-attitudes, and enhanced by pro-attitudes, is that such attitudes themselves have sensory feels (this is a possibility that Mason, 2007, asks Feldman to consider). Accordingly, if one believes that sensory pleasure is central to welfare, one ought to believe that attitudinal pleasure is central to welfare too, at least indirectly. But I will give a more thorough treatment of attitudinal pleasure in the next section.

In short, the satisfaction sensory desires contribute to welfare if: 1) they are based on no false beliefs, and that the satisfaction of them is therefore pleasurable, and 2) there is not significant outweighing attitudinal displeasure that attends the satisfaction of them.

2.4 Which non-sensory desires contribute to welfare?

Non-sensory desires can include experience-implying and non-experience implying desires. My suggested story for how such desires develop is quite similar to the story of how sensory desires develop. Some non-sensory desires may develop as a result of an experience of attitudinal pleasure in the same way that sensory desires develop as a result of an experience of sensory pleasure. Imagine a seven year old who has just won his first school running race. Seeing the crowd of parents and peers cheering for him, he experiences a rush of attitudinal pleasure toward the fact that he won the race that generates desires to win more races in the future – maybe even a desire to be a famous runner. Attitudinal pleasure and displeasure, I contend, generate non-sensory desires. This is the case even when attitudinal pleasure is experienced at imagined facts (this would have to be the case, given that it is possible for us to have non-sensory desires for states of affairs that we have never experienced). I will give a couple of examples of how both an attitudinal displeasure and an attitudinal pleasure can generate non-sensory desires. Suppose, for the first time, I see a horrible news story about war in the Middle East. I imagine thousands of people dying, being separated from loved ones, and I experience an intense sadness and attitudinal displeasure at this state of affairs. This generates in me a desire for wars and violence to cease to exist: a desire for world peace. Take another example. As a young teenager, I witness a loving relationship between my older cousin and her boyfriend. The way they talk to each other, care for each other, touch each other and smile at each other causes me to imagine what it would be like if I were in love. I take a great deal of attitudinal pleasure toward the imagined state of affairs of myself being in love, and being so important to somebody else. Accordingly, I develop a (at least partly non-sensory) desire to be in love. Attitudinal pleasures of a variety of types, then, are the building blocks of our non-sensory desires.

Unlike sensory desires and sensory pleasures, attitudinal pleasures are not necessarily the objects of non-sensory desires.[49] Think of many examples of non-sensory desires: the desire to make my parents proud, the desire to avoid morally wrong actions, or the desire to get a university degree. None of these desires have attitudinal pleasure as their *object*. However, if these desires are authentic, attitudinal pleasure will be experienced on the perceived fulfilment of such desires: if I really, authentically, desire to please my parents, then the belief that I have in fact pleased my parents is bound to bring attitudinal pleasure. Similarly so with the other examples.

We have, then, a way of distinguishing between non-sensory desires that are good for us to fulfil, and non-sensory desires that are not good for us to fulfil. If the fulfilment (or perceived fulfilment, or imagined future fulfilment) of a non-sensory desire is not attended by some attitudinal pleasure, then the fulfilment of such a desire does not contribute to one's welfare. The reader might wonder what the basis of such an assertion is. If the attitudinal pleasure is not even *the object* of a non-sensory desire, why must it be present on the perceived fulfilment of it? As I said above, the presence of attitudinal pleasure (or, the fact that attitudinal pleasure *would* be experienced *if* a desire were fulfilled) on the satisfaction of a non-sensory desire is what indicates that the desire is *authentic*. Recall, part of the worry about objective list theories is that it might render the individual's good as something that is alien to her. It strikes me that the same may be true of some desires. Some desires may fail to be *genuine* or authentic in the sense that an individual may be alienated from them. Consider the following example. As a child, I am brow-beaten into wanting to be a medical doctor when I grow up. I am told that it is the only way to make my family proud, that it is the most respectable profession, and that I should care about such things. My desire to be a medical doctor is, in a sense, quite strong: it dictates what I do, and how I organise my life. Nevertheless, I feel no attitudinal pleasure at the thought of

[49] I say 'not necessarily', because there are some non-sensory desires that *do* have as their object attitudinal pleasure: to be satisfied with one's life, for instance.

being a doctor. If I were to become a doctor, I would gain no pleasure. I may even resent the fact that I have this oppressive desire to become a doctor. In other words, my desire to be a doctor is not an authentic desire, it is not authentically *mine* – it is something that has been imposed from outside, and something that the rest of me is somewhat alienated from. I think it is perfectly reasonable to suggest that such a desire would not be good for me to fulfil. Non-sensory desires are only good for us to fulfil if they are *authentic*, and the way to tell whether a sensory desire is authentic is (among other things) if attitudinal pleasure would be experienced on the perceived fulfilment of them.

To be most clear, this is true even of non-experience implying desire. Suppose I want my children to be happy after I die. Although (assuming I have no post-mortem existence) I won't be around to see whether the satisfaction of this desire takes place, we can still say that, if my desire is authentic, I *would* feel attitudinal pleasure at my children's being happy *were* I to somehow perceive this. Furthermore, we can say that my *imagining* my children being happy after my death would bring me attitudinal pleasure if my desire really was authentic. Once again, the test of whether *any* non-sensory desire (experience-implying or non-experience implying) is authentic is whether the perceived satisfaction of the desire (whether actually perceived, counterfactually perceived, or imagined) brings attitudinal pleasure. And only authentic non-sensory desires count toward well-being.

2.5 Overall well-being

To finish off my hybrid hedonist and desire satisfaction account, I will summarise the different elements, and then give a description of what, according to the theory, the most well-off kind of life consists in. The best kind of life an individual can have is a life in which the greatest number of their intrinsic desires, both sensory and their authentic non-sensory desires, are fulfilled. This satisfaction will necessarily involve a large amount of attitudinal and sensory pleasure in life.

At this point, a further qualification must be made. Does my theory imply that a person who has experienced less pleasure, and therefore has fewer desires created to be satisfied, is better off than a person for whom many pleasures have created many desires which remain unsatisfied? The answer to this question is, 'no'. If the answer were 'yes', this would imply that deliberately closing people off from possible sensory or attitudinal pleasures would tend to be good for them, and this is a counterintuitive result.

Generally speaking, exposing people to more sensations or situations increases the amount of potential desires they have, and it counts against a person's welfare to keep their experience of pleasure restricted such that certain desires never develop. I think we can see the plausibility of this if we consider the plight of women in backward, misogynistic societies, or before second wave feminism gained for them many of the freedoms, opportunities, and privileges we have today. Many women in the pre-second-wave feminism world claimed to be happy, a statement we might interpret as 'I have everything I want'. But given the hampered development of a broad set of desires arising from such constricted conditions, and such a small vision of one's own potential, such desire-fulfilment can hardly be decisive for one's wellbeing.

Rather, what we should say is: the greatest satisfaction of *potential* intrinsic desires is what welfare consists in: and one's potential desires are those desires that one would have if exposed to certain pleasurable experiences.

I must be careful at this point, for I ought not be being taken to say that every possible sensation ought to be indulged for the sake of the possibility of creating a desire. For the experience of some pleasures, particularly sensory ones *count against* well-being overall. This may happen in a variety of ways. Firstly, the experience of a sensory pleasure may cause attitudinal displeasure. Secondly the experience of a sensory pleasure may cut one off from being able to experience other pleasures, whether sensory or attitudinal, later on. Since I have already gone into examples about how attitudinal displeasure may be taken toward the experiences of certain pleasures, let me give examples of how sensory pleasures might cut one off from other pleasures. Suppose I want to try a party drug, ecstasy. I want to experience the sensation and see if it is pleasurable. But something terrible may well happen to me if I do this – I may cause myself permanent physical damage, which would in turn cut me off from other pleasures. Suppose I want to see what it would be like to have a broad range of sexual partners while I am young, and experience the pleasures associated with this. Though I experience something pleasurable in this, it is plausible to think that I am cutting myself off from other pleasures further down the track – perhaps, once I find someone that I truly love, the sexual intimacy with them will be somewhat cheapened by my adventurous past, and I may not get the deep satisfaction from a monogamous relationship that I seek.

So then, there might be said to be two different ways in which welfare is enhanced, which potentially pull in opposite directions from each other and create a degree of tension in the seeking of welfare. On the one hand, a life full of the whole range of pleasures is the best kind of life that someone can live. But to experience a maximally large set of satisfied desires, one will most likely have to choose to forego some desire satisfactions, and some pleasure experiences for the kinds of reasons given in the examples above. At the same time, however, one's exposure to pleasures must not be too limited, lest desires for lots of things never develop in the first place, and a life resembles the example of the life lived by a woman in a pre-feminist culture.

With these two principles in mind, we can say that the best kind of life for an individual is a life in which the greatest possible set of desire satisfaction (sensory and non-sensory) is had. Such a set can be shrunk to less than 'the greatest possible' as a result of two mistakes individuals can general make: one mistake is to simply not be acquainted with enough pleasures to have enough desires to be satisfied. Another mistake is to seize on every possible opportunity for pleasure, but experience draw-backs and desire frustration on the whole. An individual's adopting some mean between these two extremes will give them the best chance of living a life with the greatest possible satisfaction of sensory and non-sensory desires, and the attendant experiences of pleasure.

2.6 Objections and further questions

In this section, I will answer a number of objections and questions that confront the theory of welfare I have advocated.

1) You say that what all sensory pleasures have in common is that they generate a desire to experience them again. But what about pleasures which, by definition only happen once, like one's first kiss? One cannot desire to experience one's first kiss again despite the fact that many first kisses are regarded as pleasurable.

Firstly, I do not think it entirely certain that it is *impossible* to have desires about 'first tastes' of things. It is very common to want to relive one's first kiss again, exactly how it was, exactly how we were at the time, even if the desire is impossible to satisfy. I think this reply is perfectly sound, but if one doesn't want to accept it, there are other things that might be said that maintains my thesis that what all sensory pleasures have in common is a certain relationship to desires: namely a desire for their repetition.

Here are other ways one's first kiss might generate desires in such a way such that the thesis that something is pleasurable if it generates desires stays intact. A pleasurable first kiss might give rise to the desire to have more kisses, whereas a non-pleasurable first kiss might not give rise to any such desires, and leave one wondering 'what's all the fuss about?' (if one does have desires for more kisses after an unpleasant first kiss, these will not be desires for the exact sensation, but rather a desire to find out just what the fuss is about in the hopes that one's second kiss will be better). One might argue that this reply slightly misses the point: the point is not that one's first kiss is pleasurable because it is a *kiss* but because there is some extra element of pleasure added to it by virtue of the fact that it is one's *first* kiss (call this the *first-ness* pleasure). And it is this first-ness pleasure that cannot be related to desires, because it cannot be repeated.

But I think this is incorrect. I do not think all desires must necessarily be future-related, and there are certainly ways of making sense of the first-ness pleasure's relationship to desires. If one's first kiss was a pleasurable first kiss, one might say that it generates the desire for it not to have been otherwise. Perhaps one can also say that, when first-ness pleasure is experienced, it gives rise to a tendency to want to bring the sensation of the first-ness to mind in memory (rather than to suppress the memory, as one might with a sloppy or unpleasurable first kiss). In any event, the fact that some pleasures, or some elements of pleasures cannot be repeated, does not seem to raise an insurmountable obstacle to defining sensory pleasure primarily in terms of a propensity to generate desires, either for the repetition of the sensation, or at least for things closely related to it.

2) Can your view deal with Parfit's addiction example?

I have said above that it is good for a person's potential desires to be brought into existence. But this raises again Parfit's addiction case: plausibly, there are some potential desires that are *not good for a person to have*, such as the desire to take a drug every morning which gives no pleasure.

When we focus on what is probably going on in Partit's example, we see that his addiction case poses no problem to my view. If I have a desire to take the drug every morning, and yet I experience no pleasure when the desire is fulfilled, one must wonder what the phenomenology of this desire involves. It is certainly not an anticipation of pleasure. But my having the desire would probably involve some form of pain – for what would give me the urge to take the drug if not my suffering and agitation that the chemical relieves, (even if such relief does not constitute positive pleasure)? What is going on in Parfit's example, strictly speaking, is likely then to be *my desire to relieve a particular pain* – the taking of the drug is just an extrinsic desire, and the relief of the habitual pain is the intrinsic desire. We see then that Parfit's addiction poses no problem: for it is only the satisfaction of *intrinsic* desires that makes one's life go well. In this case, my intrinsic desire is to *not have the pain that my drug taking must relieve*. In this way, my desire satisfaction theory (and any desire satisfaction theory) is entirely consistent with the sensible claim that I would have been better off if my addiction had never formed.

3) Can your view deal with 'crazy' desires?

Suppose I desire to be a grass-counter. Is there any way that my theory can support the view that the satisfaction of desires like these do not contribute to welfare?

The main way my view can yield this result is to appeal again to the satisfaction of potential desires. Intuitively, what might strike us as unhealthy about grass-counting is that anyone with a fair enough experience of a variety of pleasures, anyone with enough potential desires awakened, would have a desire to count grass being very low, maybe nonexistent. I think this is what we tend to think about the grass counter: if she has a desire to do *that* then she just must not have experienced much of what life has to offer: there are potential desires she has that have not been awakened, and it would be good for her to have them awakened then satisfied.

It this is not the case, and if our grass counter has experienced a vast variety of sensations, and still desires to count grass above everything else, then I'm not sure how persuasive the intuition is that counting grass is a desire that wouldn't contribute to welfare. I, for instance, have experienced many and varied sensations, and still find that the desires that dominate quite strongly might be regarded by others as the silliest: eating icecream in bed, salsa dancing, and controlling an army of aliens on a computer screen. If we're prepared to say that the satisfaction of these desires can contribute to my well-being on the basis that I find them entertaining and pleasurable, on what possible grounds could we say that an enlightened grass-counter's pleasures don't contribute to hers?

4) You say that pleasure always attends the perceived satisfaction of desires. But how is this reconcilable with the fact that we loathe the satisfaction of some desires, and desire some things that are unpleasurable?

In my view, puzzles like these can be almost entirely solved by being precise about the *objects* of our desires. One activity, for instance, the consumption of child pornography, may constitute the satisfaction of one desire, and at the same time the frustration of another. One might desire the sensations involved in the consumption of child pornography, but at the same time desire that *one not be the kind of person who consumes child pornography*. Accordingly, one will feel a mixture of pleasure and displeasure: sensory pleasure (the satisfaction of the sensory desire), along with a loathing or attitudinal displeasure (representing the frustration of a non-sensory desire to not be a child pornography consumer) at the fact that one is getting one's pleasure from such a source. Thus the example of loathed pleasures constitutes no counter-example to the two theses: that sensory desire satisfaction is always constituted by sensory pleasure, and that the satisfaction or frustration of authentic non-sensory desires will be attended by attitudinal pleasure or displeasure, respectively.

The same goes for examples of 'desiring unpleasurable things'. Suppose I desire to look after a sick elderly relative this weekend, and that I know many aspects of it will not constitute pleasurable experiences in the slightest: the elderly relative is obnoxious, has forgotten who I am as a result of Alzheimer's, and thinks I want to kill her. As such, she will hurl verbal and physical abuse at me every time I help her to the toilet and wipe drool off her face. Once again, this is no counterexample once we consider carefully that there are multiple objects of desire, not one, in this case. I certainly may not desire the *activity* of helping an abusive elderly relative for the *sake of it* – I take no sensory pleasure in wiping drool off ungrateful faces and being abused. I do, however, have a desire that I am the sort of person who helps my elderly relatives when they are in need. Accordingly, if this desire is genuine, I can expect to feel a pro-attitude and glimmer of attitudinal pleasure at points during this otherwise painful weekend. Accordingly the thesis that objects of experience-implying desire always consist of pleasures remains intact.

5) You say that the satisfaction of a non-sensory desire will always result in attitudinal pleasure if it is authentic. But surely there are counter-examples to this. Suppose I have a desire to make my father proud. Once the desire is satisfied, I suddenly feel a loss of purpose in life because I have achieved what I wanted, rather than pleasure. But surely, we cannot say that this necessarily renders my desire to make my father proud inauthentic?

Once again, I believe that this challenge can be answered, once again, by refocusing our attention on what the *object* of the desire in this example really is. For there are at least two ways we could interpret the desire to 'make my father proud'. One way we could interpret the desire is as a desire to *have it be the case that my father is proud of me*. The other way we could interpret the desire is as a desire to *enjoy the process of doing whatever it takes to make my father proud of me*. Now, suppose our desire is the former. In this case, then, it would be very likely that one would experience a lot of attitudinal pleasure. If one authentically desires the state of affairs that my father is proud of me, then surely, it will result in attitudinal pleasure on being fulfilled. But if my desire is the latter: merely the desire to engage in a *process* of making my father proud, then it is entirely natural that I would experience attitudinal pleasure during the process and the endeavours, but an anti-climactic feeling once it is over. Either way we interpret the desire, the thesis that the satisfaction of authentic, non-sensory desires will result in attitudinal pleasure upon perceived satisfaction remains intact.

6) But doesn't putting so much stock in desire satisfaction still render your theory of well-being implausible? After all, desires can be so vulnerable to manipulation: I could conceivably hypnotise you to have different desires to the ones you have

> already, or make you swallow a pill that would completely change your desires. Why, then, should well-being depend so completely on desires?

There are two parts to the answer to this question. Firstly, suppose you were to change my deepest intrinsic desires by some process of hypnosis, or a pill, and suppose you did this so effectively that this change in desires really was genuine or authentic. Once again, I see no problem with a bite-the-bullet response: in changing my desires so thoroughly, you have changed *who I am*, you have in one sense turned me into a different person or a different creature. And accordingly, you have altered the structure of my well-being and what my well-being consists in. The plausibility of this response is bought out if we consider another, more extreme example. Suppose you miraculously transformed me into a dog. In doing so, you have thoroughly changed the desires I have (and have changed even the set of desires I am potentially *capable* of having). You have turned me into a different creature, and thus have transformed the structure of my well-being.

Such a response is completely compatible with the intuition (that is perhaps driving the question) that a kind of *harm* may have been done as well. Suppose, at time *t1*, I have one set of desires, A, and at the time after you change me through hypnosis, *t2*, I have a completely different set of desires, B. Suppose also, that at *t1* I had a strong desire *not to be changed*, and especially a desire not to have a set of desires like B. In changing my welfare structure from *t2* onwards, then, it is fair to say that you have harmed my former self, the self of *t1*. Given the extremity of the change, we may view the change you forced on my consciousness as a kind of death to the person I was at *t1* (and inflicting a non-consensual death on someone is certainly a harm). But note that the possibility of *harmfully changing someone's desires* is completely compatible with the desire satisfaction theory: the change was harmful to my former self precisely because I did not *want* it. But it is equally plausible to add that, now that I *have* the new set of desires, that my new self is no longer being harmed – I am simply a different creature and have been 'changed into something else'.

3. Universal Sources of Human Wellbeing

3.1 Introduction

This final section will be relatively brief. I will present what are some fairly orthodox beliefs from positive psychology: that the most important human sources of welfare are relational. In the terminology of my theory of welfare, the greatest potential set of desire satisfaction for humans invariably consists in satisfied desires for strong relationships and friendships in which a significance to, or intimacy with, others is recognised. While there may be more superficial differences between the potential set of desires that people have, these relational desires, in which there are some fairly deep similarities between human beings, are plausibly much more significant.

I write this section partially in response to another potential objection to the theory of welfare that I have presented. The theory I have presented leaves quite open the possibility that, if one has enough malicious or immoral desires that are consistent with each other and well informed, then such desires are good for one to fulfil. Indeed, my theory does leave open this possibility in principle. But work in positive psychology supports a fairly robust empirical claim that no *human* persons are likely to have a large set of consistent desires that involve harming people. Given that human beings are instinctively empathetic, appear to be motivated by desires for significance and respect in the eyes of others, and experience positive affect (a mixture, most likely, of attitudinal and sensory pleasure) when they behave compassionately, or see others better off as a result of compassionate actions, it can be reasonably concluded that purely malicious actions are very unlikely to result in more well-being overall. Certainly, there will be occasions in which treating others badly is instrumental to a further end, in which welfare may be enhanced. Suppose for instance, I need to steal from someone else in order to keep my family and loved ones alive: this could plausibly count as such an example. But the idea that *purely malicious actions* – malicious actions done just for the sake of it – bring more desire satisfaction or pleasure than compassionate actions, looks unlikely. At the close of the last section I said that the best possible life consists in the satisfaction of a maximally large set of consistent possible desires. In this section, I will survey some literature from positive psychology that suggests that such a life for human beings will be one that prioritises close relationships, kind or considerate actions toward others and a sense of significance gained from such things. A maximally large set of desire satisfaction is, for a human being, unlikely to be one which is constituted by a high number of malicious desires satisfied.

Recall that the purpose of this chapter was to present a theory of welfare that was consistent with Williams' reasons internalism, but which yielded the result that there are some things that can be said to be absolutely good for all (or almost all) human beings regardless of culture, age, gender, or any other particularity. The first part of this goal has been accomplished. The second part of this goal is much easier, and only requires this final section.

A few remarks are in order from the outset. Like me, positive psychologists assume there to be a strong link between pleasure and desire: if an experience is experienced as pleasurable or correlated with positive affect, it generates a desire for more of the same: desires are built out of pleasures. So, though positive psychologists differ in their terminology, there is a general assumption in common with my theory: that if something is a pleasure for an individual, this is indicative also of what their desires are (Baumeister & Leasry 1995: 498). The terminological differences between positive psychologists and my own are not necessarily significant. Although positive psychologists us, the term 'positive affect' in the place of 'pleasure', it is very likely, given how positive affect is described, that it is a form of pleasure, covering emotions such as 'happiness', 'gladness', 'cheerfulness', 'excitement', 'tranquility', 'tenderness' and the like. Such emotions may constitute instances of attitudinal pleasure: one can be 'happy that…' 'glad that…' or 'excited by…' certain propositions. Alternatively, positive affect probably forms a component of many sensory pleasures. It is rare to partake in a pleasurable sensory experience like an exhilarating surf or a warm bath without some emotional component attending the mental state.

The fact, then, that positive affect is highly correlated with fulfilling relationships and compassionate behaviour, as we shall see, probably indicates that sensory and attitudinal pleasure is highly correlated with fulfilling relationships and compassionate behaviour as well. No doubt, there are many sources of sensory pleasure that are not mentioned in the research I am about to survey (nice food, and hot baths for instance). But it is reasonably self-evident that things like nice food and hot baths lead to pleasurable experiences, and so this perhaps does not need much of a mention. What may not be so self-evident, in an increasingly individualistic culture, is the importance of what might be called ‘relational goods’: their propensity to be pleasurable and thus to be indicative of what some of our deepest, (perhaps not completely transparent to us) authentic intrinsic desires are.

3.2 Relational, universal goods

In this section I will briefly summarise some findings from positive psychology. The studies and articles that I will be summarising are largely studies of how interpersonal relationships and compassionate or considerate behaviour generate positive affect, and thus are indicative of what desire satisfaction, or potential desire satisfaction, consists in.

Baumeister and Leary’s famous paper (1995) proposed that all human beings experience a drive for belonging: ‘First, there is a need for frequent, affectively pleasant interactions with a few other people, and, second, these interactions must take place in the context of a temporally stable and enduring framework of affective concern for each other’s welfare.’ Ibid: 497. The authors plausibly proposed that a broad range of motivations, including the motivations to accomplish achievement and to experience approval and intimacy are derivative of this desire to belong:

> [People] prefer achievements that are validated, recognised, and valued by other people over solitary achievements, so there may be a substantial interpersonal component behind the need for achievement. And the needs for approval and intimacy are undoubtedly linked to the fact that approval is a prerequisite for forming and maintaining social bonds, and intimacy is a defining characteristic of close relationships. Ibid: 498.

The authors go on to argue that the desire for the creation of lasting social bonds is exhibited by the joy and positive perceptions associated with events such as marriage, in which two people ‘promise to fulfil each other’s belongingness needs’ (ibid: 506), and around childbirth. Although parenthood can cause certain increases in the frequency of negative affect due to stress and marital strain (ibid), it is plausible that the positive perception of childrearing that persists even in the midst of this (ibid) indicates a satisfaction of a non-sensory-desire in the form of an attitudinal pleasure toward the fact that one has children.

While positive affect attends the forming of social bonds, negative affect attends the breaking and absence of such bonds. Three emotions mentioned: anxiety, jealousy, and loneliness, are forms of negative affect, and all occur when close bonds are absent or interfered with. Anxiety can be triggered by being separated from loved ones for long periods of time, imagined future loneliness, and by memories of past rejections (ibid: 506). Jealousy, and particularly sexual jealousy, has been found to be universal across cultures, despite prior claims by anthropologists that sexual jealousy was a culturally-conditioned emotion (ibid). And loneliness, an emotion that arises particularly from a lack of intimate contact with others, was shown to affect people with wide circles of acquaintances but few close relationships. Thus, mere social contact was not found to be enough to stave off negative affect, but belonging in particular.

Other work adds further detail to Baumeister and Leary's proposal that human beings have fundamental desires for belonging. It has been noted that the fulfilment of desires for intimacy in particular is required for the experience of positive affect (Diener & Diener McGavran, 2008) note the importance of attachment figures in the lives of infants, and the predictive power of positive affect later in life, as well as the enthusiasm and capability for problem solving (Matas, Arend, & Sroufe, 1978). This research reinforces an intuitive observation that most of us are capable of making: that satisfied desires or positive affect in one component of life can open one up to the experience of more positive affect in other areas. The satisfaction of relational desires may be essential, then, to the satisfaction of seemingly non-relational, task oriented desires such as the desire for achievement, as has already been noted (Baumeister & Leary 1995, Diener & Diener McGavran, 2008). To tie these findings back to the terminology of my theory of welfare: exposure to certain pleasures (those related to feelings of intimacy and security) seem to both generate other desires in other areas of life, and indicate that there are certain hard-wired desires for intimacy there in the first place.

The research around marriage relationships allow for the possibility that particular relational choices, such as entering committed relationships and remaining committed, are generally better or worse for humans:

> The majority of the people worldwide develop a close relationship with a spouse in marriage, and it is well-documented that marriage is related to greater subjective well-being (Diener et al., 1999). Most

> research shows that married people are happier than never-married, divorced, separated, or widowed individuals… In a cross-sectional study Kamp Dush and Amato (2005) showed that being in a romantic relationship (either with a spouse, a cohabiting partner, or a steady dating partner) was associated with greater subjective well-being, compared to the single individuals not dating or individuals dating multiple people. Moreover, the greater the commitment in the relationship, the greater the association between the romantic relationship and subjective well-being. Even after controlling for relationship quality, married individuals still had the highest levels of subjective well-being.' (Diener & Diener McGavran, 363).

Diener & McGavran go on to cite research about divorce: those who divorce experience a significant dip in positive affect, which lessens in severity over the five years following the divorce, but which tends not to adapt back to the levels of happiness experienced during the marriage (ibid: 365). Once again, this confirms that much pleasure and corresponding desire satisfaction for human beings is to be found in close, committed relationships.

While the above studies place an emphasis on a need for intimacy, other studies examine the need for significance: the need to matter, or be important and influential to someone or something. Demir et. al. (2011) found in their study that a perception of one's being significant in the eyes of others, one's making a difference to another's life, mediated the association between the quality of friendship and happiness. But another paper that draws on this need is Myers's famous 'Religion and Human Flourishing' (2008). A survey that Myers sights which captures an obvious link between religiosity and attitudinal pleasure is Gallup's (1984) 'Religion in America' which found that those whose religious convictions were strongest; those who agreed most strongly with statements such as "God loves me even though I may not always please him" were twice as likely as those least committed to such convictions to report being 'very happy'. (Myers 2008: 324) Myers cites such beliefs as sources of durable self-esteem: one finds one's self-worth in the fact that one is loved by one's creator unconditionally. This durable self-esteem hypothesis is one of the most plausible suggestions on the potential of a person's faith to induce positive affect: for even when the aspect of the social belonging that religion secures is controlled for, religious socially-connected subjects score higher on the positive affect tests than their non-religious counterparts. In other words, the beliefs themselves are grounds for attitudinal pleasure regarding one's own significance or importance (ibid: 327). The point of this is to highlight the ubiquity of the human desire to be significant: for one's life to matter in the grand scheme of things, and to others. It is in this significance to others that we derive much of the meaning in our lives.

One last study is worth a mention. Mongrain et. al (2011) surveyed the existing literature on compassionate behaviour to test a hypothesis. Roughly speaking, the study sought to see whether personality differences impacted the level of increase in positive affect upon practicing compassionate behaviour over time. Two personality traits were looked at in the subjects: some who were driven to seek relationships out of a fear of abandonment and anxiety, and others who were more cautious, reserved, suspicious of others, and apparently more self-reliant. Contrary to the authors' expectations, the reported increase in positive affect was the similar across the sample of 'anxious' and 'avoidant' subjects, suggesting that compassionate behaviour in and of itself is a source of positive affect (ibid: 965-6, 974-5).

In conclusion, the following can be said. While we may have a tendency, in an individualistic society, to think that what is good for a person differs radically from person to person, and that each person can fiercely determine what is good for oneself, the dominant opinion that emerges from positive psychology is that this is not the case. There are deep, systematic similarities between what people deeply desire and are motivated by, and in what sources positive affect or attitudinal pleasure is found. These are all related to the quality of interpersonal relationships, and our behaviour toward other people. Although there are basic non-relational physical desires which could be said to constitute needs: the needs for enough food, sleep, and shelter, beyond this it is the quality of our relationships with others that largely determines our well-being. Sure enough, we are all familiar with the fact that the pleasures we take in gourmet food, great music, a massage, a good surf, or an engaging Sudoku puzzle do contribute to welfare, their importance to our welfare should not be overstated at the expense of our pursuit of relational goods.

The significance of these findings for the project overall will be discussed in chapter five. These findings are, in summary, that there appear to be facts about human psychology: our desire for relationships, significance in the eyes of others, and intimacy, which suggest that the best-off life for probably any human will be a life in which self-centred or unkind behaviour is minimised, and in which other-person-centred behaviour plays a significant part. A human life in which the maximal potential set of desires are satisfied, as a matter of contingent fact, will involve a life full of friendships, kindness, achievements appreciated by others, and intimacy. On the research looked at in this last section, I make what I think is a fairly modest claim about the sources of welfare. A life in which the greatest potential set of desire satisfaction occurs is likely to consist in satisfied desires for intimacy and significance to others, which is brought about most effectively by participation in loving relationships, challenging experiences, and compassionate behaviour. A life dominated by malice or selfishness, on the other hand, is likely to be far lower in positive affect, and leave many potential desires unsatisfied.

5:
Ideal Observer Sentimentalism

1. Ideal Observer Theories

1.1 Ideal observer theory

Recall the ‘moral consideration claim’ (MoCC) from chapter 3:

> **The Moral Consideration Claim:** A makes a positive moral judgment about X iff A judges that there is a positive relation between X and the wellbeing of others. A makes a negative moral judgment about X iff A judges that there is a negative relation between X and the wellbeing of others. A makes a neutral moral judgment about X iff A judges that there is neither a positive or negative relation between X and the wellbeing of others.

MoCC is a doctrine about what it takes to make a normative judgment a moral judgment. Chapters 4 and 5 are dedicated to the task of illuminating what it takes to make a *true* moral judgment. Chapter 4 was dedicated to giving a theory of what ‘wellbeing’ consists in. This chapter will spell out what is meant by a ‘positive’ or ‘negative’ relation between something morally evaluated, and the well-being of others. In chapter 3, this positive relation was given a stipulative definition: ‘any relationship which a plausible normative ethical theory might take to be relevant’. This included a relationship of promotion (of well-being), or an instance of following a rule that could be said to promote well-being if followed, or an embodying of a respect for the rights of others, or having sprung from a motivation of care for others and their wellbeing. In this chapter, I will identify a feature that all such things could be said to have in common.

I will argue that a positive or negative relation between something morally evaluated, and the wellbeing of others, is determined by the responses of a certain kind of ideal observer. In its most basic form, the view is as follows:

> **Basic:** R is a positive relation between X and the wellbeing of others if Ideal Observer would have response Y to X. R is a negative relation between X and the wellbeing of others if Ideal Observer would have response Z to X.

In this chapter, I will fill out two details in turn: what the ideal observer's characteristics are, and what the responses are: the answers to these questions will distinguish the view from other ideal observer views given. But firstly, some introductory remarks on what purposes these two elements in ideal observer theories serve, and where the appeal of ideal observer theories lies.

The thought behind the ideal observer views is simple. When we make moral judgments, we normally think that our judgments have a greater chance of being true if we exhibit certain characteristics when we make them. Examples of such characteristics may include the following: our being impartial, our being knowledgeable of empirical facts concerning what we are judging, our being sympathetic or empathetic to every party involved (if we are attempting to resolve a conflict of interest with our moral judgment), and to be perfect in our inference-making, lest the moral arguments on which our judgments rest be invalid. If we agree that our possessing such a set of characteristics improves our chances of making true moral judgments, the argument goes, then our moral judgments would be infallible if we possessed such characteristics *perfectly*. And this is the role that the concept of the ideal observer plays: a being who possesses those qualities or characteristics that we think of as necessary for having perfect insight into moral truth. We can see this thought present in those who defend ideal observer theories. Firth writes: 'In analysing ethical statements, for example, we must try to determine the characteristics of an ideal observer by examining the procedures which we actually regard, implicitly or explicitly, as the rational ones for *deciding* ethical questions… in practice we are likely to rate moral judges by reference to their similarity to an ideal observer.' (1952: 332-333).

Sometimes, an ideal observer theory is taken to be an analysis of moral judgments (Firth: 1952, Smith, 1989), while other ideal observer theories are taken only to indicate the truthmakers or standards of correctness of moral judgments (Carson, 1981). My ideal observer theory falls into the second camp. Chapter 3 gave us our analysis of moral judgments in MoCC, which said that moral judgments must concern positive relations to well-being. I am proposing that what all positive and negative relations have in common with each other is determined by a certain response from an ideal observer. But one need not believe this to qualify as a competent maker of moral judgments and possessor of moral concepts. To repeat the analogy used in chapter three: a three year old competent user of the concept of 'water' does not need to know that what the rain and the stuff in the bathtub has in common is that it is H2O. But the fact that the rain and the substance in the bathtub is H2O is what makes his judgments about water true: 'water is falling from the sky'… 'there is water in the bathtub.'

1.2 Avoiding Uninformativeness

As I have said above, there are two elements of the ideal observer theory to be clarified in this chapter: the characteristics of the ideal observer, and the response the ideal observer has to morally evaluable things. In both these tasks, there is a danger to be aware of: the danger of rendering the theory uninformative.

When outlining the ideal observer's characteristics, one must be careful not to include any characteristics that involve moral terms: such as 'morally astute', for instance. This would render the ideal observer theory uninformative, given that such terms essentially mean: 'makes true moral judgments'. But moral truth is precisely what we are trying to give an account of. Firth, again, makes this point at a number of places in his famous paper (1952: 321, 326, 334). Accordingly the characteristics of the ideal observer that I will list will not make use of any moral terms, and as in Firth's theory, the term 'ideal' simply indicates that our agent possesses certain characteristics to a certain degree (ibid: 321).

One thing Firth remains uncommitted about in his theory is what kind of responses the ideal observer would have. However, (and especially in my theory), it will be important that we avoid uninformativeness again, and avoid falling into the trap of saying that the ideal observer's responses are 'moral judgments'. For according to our theory, the ideal observer’s moral judgments would be judgments about the positive or negative relations between what is judged, and the well-being of others. And we are trying to use the concept of the ideal observer’s responses to clarify what these positive and negative relations *are*. It is obvious, then, that we must not use any moral terms in our explication of the ideal observer’s responses just as we must not use any moral terms in the explication of the ideal observer’s features.

As I spell out my ideal observer theory, the reader will see that it resembles Firth’s theories in some ways, but also differs in a number of respects. When outlining the characteristics of the ideal observer, I will take into account some of the objections to Firth’s version made by Brandt (1955, 1974) and Carson (1981), and construct my theory in a way that I believe avoids these objections and retains the best elements of Firth’s theory. When giving an account of the ideal observer’s responses, I will borrow elements from Michael Slote’s (2010) sentimentalist account of moral approval and disapproval. Accordingly, I call my theory of the truthmakers of moral judgments ‘ideal observer sentimentalism’.

In putting forward this theory, I do not propose that it is the only plausible theory of the truthmakers of moral judgments. However, it is a theory that fits nicely with the other claims that I have made, and it does have certain advantages that I will highlight toward the end of the chapter.

2. The Ideal Observer’s Characteristics

2.1 Omniscience

One characteristic that almost every ideal observer theory insists on is some degree of empirical or non-moral knowledge possessed by the ideal observer. Knowledge of a given empirical fact can make the difference between my making a positive moral judgment and a negative moral judgment, and we usually take knowledge of enough facts to be one of the features we must possess if we are to be accurate moral judgers. Accordingly, *knowledge about empirical facts* is a requirement discussed in most ideal observer theories.

Firth insists that the ideal observer would be *omniscient about all empirical facts*. My theory basically resembles Firth's in this respect, though I prefer to say that my ideal observer is omniscient about all non-moral, non-normative, or purely descriptive facts. Such omniscience does not only include those that can be investigated by the sciences, but, importantly, an *omnipercipience* (Firth, 1952: 335): an awareness of the feelings and sensations of all conscious beings, and a familiarity with their perspective, their desires, and their history. Another set of facts that my ideal observer's omnipotence ranges over is *knowledge about what is in each interest-bearer's well-being*. According to the theory of welfare given in chapter 4, this means that the ideal observer knows what each person's intrinsic desires and sources of pleasure are, and also what each person's *potential* desires are (which desires people would acquire if exposed to certain sensations). So the ideal observer is not only perceptive of each individual's positive or negative affect at the present point in time, but perceptive of what each individual's affect *would* be at future times given certain situations. The need for the ideal observer to be omniscient of such things should be obvious, since it has been established that welfare is morality's central concern.

There is an argument regarding the characteristic of omniscience, discussed by Taliaferro (1988) in which he compares the merits of Firth's ideal observer theory and Carson's 1981 theory. Carson objects that the ideal observer need only be aware of all *relevant* facts, rather than all of the descriptive facts that there are. Carson writes: 'The trouble… is that the ideal observer's omniscience is incompatible with his humanity. Human beings are not capable of knowing everything. It is not intelligible to ask how someone would react to various things if he were omniscient.' (1981: 57). As I will argue later, Carson's insistence on the ideal observer's humanity is unnecessary and problematic. But even if this were not the case, one could still question Carson's 'relevance restriction' by his own lights: it is unlikely that humans very often know every *relevant* fact to ethical judgment making, anyway.

Besides all this, however, is the problem that Carson's 'relevance restriction' introduces a circularity or uninformativeness to the ideal observer theory. It can certainly be granted that some empirical facts are morally relevant, or relevant to the making of a moral judgment, and some are not. But, given that such relevance is a moral concept, and that *judgments* about which facts make a moral difference and which do not, are a species of moral judgments, the 'relevance restriction' cannot be utilised to illuminate what a true moral judgment is. Carson attempts to avoid this charge by simply defining a relevant piece of knowledge as one that would make a *difference* to an ideal observer's judgment (ibid: 58). But, as Firth himself says, this would involve another form of circularity with regard to the concept of the 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 of any part of it. It follows, therefore, that in order to explain what we mean by "[morally] relevant knowledge", we should have to employ the very concept of an ideal observer which we are attempting to define. Firth 1952: 334

To summarise: our understanding of what is morally relevant or not concerns moral truth, which we are using the ideal observer to give an account of. We cannot, then, put a 'moral relevance' restriction on what the ideal observer is omniscient of. Accordingly, I prefer to say that the ideal observer is omniscient of all non-moral facts (this is certainly a much simpler move, and one which, I hope it will become clear, has no significant downsides). The question of which of these facts make a difference to the ideal observer's responses will have to do with some of the ideal observer's other characteristics: his empathy, benevolence, or desires.

2.2 Empathy-driven desire

When we make moral judgments, we tend to think that our judgments will be more accurate if they are purged of bias: if they are *impartial*. Suppose I made the moral judgment that Mary ought not to ϕ in a particular circumstance C. Suppose I think that Mary's ϕing would cause unfair suffering. Suppose John then found himself in the same circumstance, C. Suppose everything about John's situation is identical to Mary's, but I made the judgment that John was allowed to ϕ in C simply because he is John and not Mary. I would be exhibiting an unacceptable partiality in my moral judgments that we would think an accurate moral judge free of. Accordingly, another of the ideal observer's characteristics majored on by Firth is that of *impartiality*. The ideal observer has no *particular interests*, which Firth defines as interests that 'cannot be defined without the use of proper names' (ibid: 338), or of indexicals. This would involve interests like the following: my privileging *Australian* citizens because they are Australian citizens, my permitting *John* to ϕ because he is *John*, my privileging Sam because he is *my* friend, or privileging *myself* simply because I am *me*.

Firth's impartiality requirement has been seen to be problematic meeting a challenge put best by Brandt (1979):

> Indeed it is not clear that a purely disinterested being would support a moral system at all. Clearly he would lack one quite good reason normal people have for supporting a moral code: that they themselves would like to be able to rely on promises, to be safe from assault, and to be able to command the assistance of others in times of distress. For a disinterested person would not support a moral system for reasons of personal interest. And it is possible he would not support a moral system for its contribution to human welfare generally; for he might be indifferent to human welfare… Matters are somewhat altered if we add 'benevolence' to the conception of an ideal observer. Then, at least, we could know that an ideal observer would support some kind of moral system for the sake of the welfare of sentient beings. But for more precise inferences to the content of the code that would be supported, we need to know more precisely how benevolent he is to be. 1979:227

Although such a criticism is not entirely fair to level at Firth, given that Firth does, in fact, go on to suggest that one could non-circularly attribute characteristics such as love and compassion to the ideal observer (1952: 341). But Brandt's point is nevertheless well made: one cannot get reactions that we would think of as ethically significant out of an omniscient observer who has no interests *whatsoever*.[50] Accordingly, I will now spell out my take on the ideal observer's 'love and compassion'.

My ideal observer has a desire that determines all other desires: a desire for all interest-bearers to be maximally well-off. Under the terminology from the previous chapter, this may be construed as a sensory desire, or non-sensory desire, or both. The desires of the ideal observer for the well-being of other interest-bearers springs from *empathy* – a tendency to be hurt when another is hurting, joyful when another is joyful, excited when another is excited, and so on: an empathy that is made even more extraordinary given his omnopercipience. Thus the desire that the ideal observer has for interest-bearers to be

[50] Of course, this charge by Brandt may be unfair as well – for Firth's claim is that the ideal observer has no *particularistic* interests, rather than no interests at all. However, there is an apt concern behind Brandt's criticism: if the ideal observer is defined merely by his *lack* of interests, rather than a detailed positive account (which, lacking one, constitutes a weakness in Firth's theory), it is doubtful whether his reactions would be ethically relevant.

well-off could be said to be a sensory desire (if we assume that being omnipercipient of other's mind would generate a sensory feel), given that the ideal observer will enjoy the experience of the positive affect that constitutes the well-being of said interest-bearers. But, at the same time, the ideal observer could be said to be someone who takes attitudinal pleasure in the *fact that* interest-bearers are well off, so it may be a non-sensory desire that the ideal observer has. In any case, this ideal observer is a being with a desire set and propensity for pleasures that is directly satisfied via the satisfaction of *other* desire sets and propensities for pleasure (which is, according to chapter 5, what welfare consists in) – a completely empathetic, other-person-centred being.

Two things are worth saying about the ideal observer's desires. Given that the ideal observer's desire for the well-being of interest bearers is driven by empathy, this creates the basis for chapter 3's claim that only sentient beings count as interest bearers of the kind that morality is concerned with. Chapter 3 left off claiming that plants and non-sentient beings were excluded from the category of the kind of 'interest-bearer' we are interested in. This ideal observer theory removes the arbitrariness of this cut-off line, and provides a basis for our intuitions on this point. True moral judgments can only be concerned with the interests of sentient beings, given that their truth conditions are constituted by the responses of an ideal observer who is driven *out of empathy* for a concern for 'interest bearers'. Now, the only kinds of beings that can be empathised with, that the ideal observer's omnipercipience can extend to, are beings who are at least sentient. Here, then, we have a basis for sentience as being the cut-off line for what could be considered an interest-bearer of moral import.[51] One cannot properly be said to be able to empathise with a plant or non-sentient animal (absent a false projection of a consciousness), thus plants and non-sentient animals do not count as 'interest-bearers' in the sense we are interested in. Of course, this does not mean that the ideal observer has no concern whatsoever for the parts of the natural environment made up by non-sentient organisms.

[51] One qualification should be made here. In debates about the morality of abortion, personhood is assigned to embryos by pro-lifers such as John Finnis (2011) and Don Marquis (2011) on the basis that consciousness is a property that *will be* possessed by not-yet-sentient embryos after the passing of some un-interfered-with development. My ideal observer view can allow for (but does not insist on) the possibility that embryos might count as 'interest bearers' in the moral sense on the basis of this future, even though they are not currently sentient. It is not out of the question that the ideal observer that has been described here may feel an empathy or sadness on the part of the creature that could have been, in response to an abortion. Similarly, the ideal observer may feel sadness on behalf of future generations who may suffer from the irresponsible actions of currently existing interest bearers, as I will point out in a moment.

The ideal observer may be concerned about the treatment of the non-sentient natural environment, but only insofar as this affects the well-being of sentient interest-bearers.

Another point that must be made clear is the interpretation of the ideal observer's desire 'that all interest-bearers be well-off'. At least two interpretations of this desire are possible: a de dicto or a de re interpretation. A de dicto interpretation of the desire with regard to currently existing interest-bearers would obviously be inappropriate: this could entail that, if the actual world was populated by interest bearers who were all suffering, the ideal observer would desire that they go out of existence and be replaced by a population of more well-off beings. Such a desire would hardly generate responses that we could take to be constitutive of any plausible moral code. It is, then, a de re interpretation of the desire with regard to currently existing interest-bearers that the ideal observer would possess: a desire that the *actual* currently existing interest bearers be maximally well-off. But with regards to the well-being of future interests and whether our interpretation of this should be de re, a familiar puzzle of the non-identity problem confronts us. Would the ideal observer have any responses that would ground moral judgments to the effect that we ought to make the world a hospitable place for future generations? The difficulty with answering 'yes' to such a question is, of course, that we could not be sure what kind of well-being-orientated concern on the ideal observer's part would give rise to this verdict. The ideal observer cannot be concerned about the well-being of agents who, *but for* the existence of negligent present actions, would not exist. The ideal observer cannot say that current actions make any future persons better off, for current actions make a difference to which future persons even exist. I think, however, all this puzzle shows is that we could not take a de re reading of the ideal observer's desire for the well-being of future interest-bearers. But there is nothing to prevent us from taking a de dicto reading. Our ideal observer wants the future to be populated with as well-off beings as possible, regardless of who they are. Such a desire can generate first order moral judgments, the truth of which we would want to be at least plausible: that we ought to make the world a healthy place for the sake of future generations. In summary then, the ideal observer's desire that all interest-bearers be well-off is de re with regard to currently existing interest bearers, but de dicto with regard to future interest-bearers. What I have said about these desires, of course, are general comments: there may be exceptions to these respective de dicto and de re rules should our intuitions find any, but these need not be specified for my purposes, which are modest with respect to first-order conclusions.

More needs to be said about the ideal agent's desire 'that all interest bearers are maximally well off'. At first glance, one might object that the ideal observer *could not* have such a desire: for not every interest bearer *can* be maximally well off, as the following scenarios remind us. A car crash has taken place, there are two victims mortally wounded, and the paramedics only have time to save one. Both victims cannot be maximally well-off. A natural disaster strikes, displacing a whole community of people. Many people in the community lose loved ones, and their loss is so traumatic that many of them will be scarred for life. Whatever well-being they may recover, these people will never be as well off as they could be. The point is simple: in a world like the one in which we live, in which nature is indifferent to human suffering, and in which trade-offs between the welfare of individuals must sometimes be made, it is not realistic to expect (short of a miraculous transformation of the world) that all interest bearers will be maximally well-off.

However, I do not think these considerations warrant any objection to the idea that the ideal observer would nevertheless have such a desire for all interest bearers to be maximally well-off. Many people who are morally saintly, or who we regard as accurate moral judges, are so because they are driven by a deep desire to see people better off, and a deep desire to see the suffering of the world ameliorated. The same desires that drive charity workers to eliminate hunger in starving communities would plausibly generate the same desire that the world was a more hospitable place. The most compassionate people often find themselves desiring unrealistic things: that the world were a place where no one ever got hurt, where no natural disasters took place. Now, recall, if the ideal observer's responses are to be ethically relevant, he must have certain characteristics to a certain degree. Among other things he must have compassion and empathy more saintly than the greatest known moral saints. Accordingly, there seems to be no reason to think that the ideal observer, given his great compassion, would not possess the unrealistic desire that the world be different to what it is such that every interest bearer in it could maximally flourish. The ideal observer's empathy, we imagine, would lead him to feel sad when horrible trade-offs between the wellbeing of some interest bearers and others takes place, even if such sadness coincides with a desire that a certain trade off does take place, for the sake of damage minimization. The same empathy would lead him to be grieved when natural disasters occur. If he is so sad and grieved, it makes perfect sense to say that this is because he desires that things like this do not happen: he desires that the world be a place where interest-bearers are maximally well-off.[52]

The more significant question is what other desires this desire that all interest bearers be maximally well-off would entail. Some might jump to the conclusion that this would entail that the ideal observer has consequentialist desires that wellbeing be maximised at every point. Accordingly, one might worry that this ideal observer theory yields unpalatable first-order ethical views: that the doctor ought to cut up one healthy patient for the sake of five, for instance. But I think this could be easily disputed, and I do not think such an 'entailment' between the former desire and the latter is obvious. There are many people who desire that the world be a place in which everybody is maximally well-off who would *not* desire that the doctor in the famous scenario cut up the patient (myself included!). The desire for a world in which everyone is maximally well off – an idealistic desire shared by many ethically-minded people – can generate a whole range of desires and corresponding ethical intuitions. In some people, the 'idealistic desire' generates a desire that everyone follow a set of rules which are likely to maximise welfare if followed. In others, the idealistic desire generates such a repulsion at the idea of harming innocents that the following of all possible deontological side-constraints are strictly followed. In others, the idealistic desire generates a desire for people's rights to be respected, and a conviction that such a respect for rights must form the basis of the world's collective well-being. The point is simple: the ideal observer's desire that every interest bearer be maximally well off is not a desire that is tied to any normative theory in particular – at least not *obviously* or uncontroversially so. Different proponents of different normative ethical theories may want to argue that the desire for interest bearers to be well off *ought* to generate desires that are in line with their

[52] I may be accused of making what some see to be a controversial assumption here: namely that we are maximally well-off if we never experience suffering. Though I am sympathetic to such an assumption, my view does not rely on it. We might imagine that, if it is the case that some amount of suffering is good for interest-bearers, the ideal agent would desire that such suffering is undergone. However, this is still compatible with the idea that the ideal observer is sad about the agents' suffering, resulting from his desiring that it not have to be the case that agents experience suffering in order to be maximally well off. The only reason for thinking he would not have a desire of this latter kind would be some reason to think it is *analytic* that maximal well-being requires suffering (and thus that the omniscient ideal observer's desire for maximal flourishing of interest bearers would not be able to be dissociated from the desire that they experience suffering). But the analyticity of such a truth would be difficult to demonstrate.

particular ethical theory (for example, the hard-line consequentialist may *want* to argue that, if one really desires that interest bearers flourish, then it *makes sense* for us to desire that the doctor cut up the healthy patient). But such argument between normative ethical theories is exactly what a good metaethical theory should allow for. What I deny is that ideal observer sentimentalism is *obviously* committed to any normative ethical theory. Indeed it is not clear that debates about which ethical theory is best supported by the ideal observer's basic desire could be easily resolved.[53]

One final, brief point about the ideal observer's desires is worth making. I am imagining that the ideal observer does not have any other desires or concerns besides his empathy-driven desire for the well-being of all interest bearers (and any other desires entailed by that desire). The ideal observer has no aesthetic desires, no interests in other non-moral affairs, except insofar as they impact on the well-being of interest bearers. The only reason I make this point is because it can be doubted (for example, Brandt 1974) that an ideal observer's concerns would always be morally relevant, if they were compromised by non-moral concerns. We can circumvent this problem simply by eliminating such concerns from the ideal observer's desire-set.

A question that is often raised about ideal observer theories is whether they must entail a version of moral relativism (Brandt 1955, Carson 1981, Taliaferro 1988, Smith 1989, 1994). Given that the ideal observer is a *type* of individual, not a particular individual, questions arise as to whether different ideal observers might differ or disagree in their responses. If idealised you and idealised me responded to the same situation, might we not respond differently despite our possession of both omniscience and benevolence as described? If this were the case, would this mean that moral judgments have relativistic truth conditions much like those envisioned by Harman (1975)? Alternatively, we might instead say that there is no moral fact of the matter at all. Whichever way we might want to put the idea, both ideas are ones that I would like to avoid. With this in mind, I have attempted to outline

[53] In order to resolve such a debate, one would have to venture into territory that non-cognitivists are familiar with: the idea of a 'logic of attitudes', or a logic of other desire-like states (see Schroeder 2010 for a discussion of non-cognitivist attempts to employ such a thing as an attempted solution to the Frege-Geach problem). One would have to see whether the ideal observer's desire could in any sense rationally 'entail' other desires. If one could show that this is the case, the way would be open to arguing that the ideal observer would have desires for particular courses of action to be taken in particular situations, and that such desires perhaps vindicate one ethical theory on every occasion.

the ideal observer's characteristics such that ideal observers *would not* differ in their responses, and hence would block relativism and one possible source of ethical indeterminacy. I believe the ideal observer I have constructed will do this: such that whether you, or whether I, come to possess the attributes that the ideal observer possesses, our responses would be the same. I will defend this claim in section 4.

2.3 Rationality

One last set of characteristics worth a mention is the ideal observer's 'rationality'. This is a fraught term, used in various ways. Smith sees rationality as a trait exhibited when one is 'cool, calm, and collected' (1989, 1994), and a characteristic which, if shared by more than one person, will tend to result in convergence of opinion. Presupposing no external reasons, I simply define the ideal observer's 'rationality' as involving an infallibility when it comes to making inferences, constructing arguments, or deducing conclusions from premises.

It should be clear why this component of rationality is required. We often make our moral judgments on the basis of lengthy processes of reasoning requiring inference making and valid argument making. Accordingly, if we make a faulty inference in our moral reasoning, the accuracy of our moral judgment is in doubt. An ideal observer, one whose responses underwrite moral truth, would have to be one free of such fallibility. Perhaps, as has been suggested, there may be such a thing as a logic of attitudes and desire-like states (see Schroeder 2010, Blackburn 2006a, 2006b) which dictates that some desires or emotional responses are contained in, or somehow entailed by, other desires or emotional responses: perhaps the ideal observer's empathy-driven desire for maximal well-being entails other more specific responses in a way analogous to the entailments of some propositions from others. The ideal observer's full rationality, presumably, would involve the ability to follow this logic of attitudes (perhaps instantaneously), and have his responses governed accordingly.

Whether the ideal observer's rationality can be grouped under the umbrella of 'omnipotence' (much like 'omnnipercipience' can), does not really matter. What is important is that the ideal observer's infallibility as an inference maker is made explicit.

3. The Ideal Observer’s Responses

3.1 Judgments and responses

So far, we have outlined the ideal observer’s characteristics. The ideal observer is omniscient about all non-moral facts (and, importantly, is omnipercipient and omniscient about every individual’s welfare). The ideal observer is also maximally empathetic, possessing a desire that every interest bearer be maximally well-off, and is a flawless logician. It is this kind of ideal observer’s responses which determine the truth of moral judgments.

But, as has been pointed out at numerous points, there are different kinds of moral judgments. Thus, the ideal observer’s responses will vary according to which kind of true judgment is in question. I offer the following schema of four different kinds of judgments:

> **Positive/rightness:** A’s judgment about the rightness of X is true iff Ideal Observer would disapprove of a failure to do X.
>
> **Negative/wrongness:** A’s judgment about the wrongness of X is true iff Ideal Observer would disapprove of the doing of X.
>
> **Neutral/permissible:** A’s judgment about the permissibility of X is true iff Ideal Observer would neither approve nor disapprove of the doing of X.
>
> **Positive/goodness/supererogatoriness:** A’s judgment that X is supererogatory is true iff Ideal Observer would approve of the doing of X.

From these four types of moral judgments, we have what we need to say what a positive, negative, or neutral relation between something being judged, and the well-being of others, is. A positive relation between the thing judged, and the well-being of others, is constituted by the fact that our ideal observer would have either an approval of the thing in question, or a disapproval at its failure to be done, as indicated by our two types of positive judgments. So positive relations such as the promoting of interests, an action’s being a member of a set of rules which promotes welfare, or an action’s in some sense ‘embodying’ a respect for the well-being of persons or being motivated by such respect, can all count as relations that our ideal observer might disapprove of for our failing to realise, and in some cases, approve of our realising. Similarly, what all negative relations between an action and the well-being of others have in common is that our ideal observer would disapprove of those actions. And with regards to 'neutral' relations, a lack of either approval or disapproval is the common feature, though there is more to be said about this.

A note of clarification is worth making at this point lest the above words be misunderstood. I said above that certain positive relations all have in common is the approval (or counterfactual disapproval) they draw from our ideal observer. But I do not mean to say that any of these relations that we *generally* think of as positive are *always* positive relations. Take, for instance, the promotion of the interests of the doctor's five patients through the killing of the one innocent patient. Now, though most of us generally regard the promotion of interests as a *positive* relation between actions and the wellbeing of others, many of us would regard *this* particular promotion of interests as more of a negative relation to the wellbeing of others, given that we think the doctor's action wrong. So when I say that various types of relations between actions and the well-being of others are *positive* relations, I should be taken to be saying 'in general'. There are situations, like the one above, in which otherwise normally positive relations between actions and welfare are negative.[54] So when I say that what all types of positive relations have in common is the response of approval/counterfactual disapproval from the ideal agent, I do not mean to imply that these relations are *always* positive, nor that they always engender this response in the ideal observer. In our doctor's scenario, (if our judgment about the wrongness of the doctor's act is true), our ideal observer's reaction would be one of disapproval.

The question to answer now is: what does the ideal observer's disapproval or approval consist in? As I said above, we must be careful that our theory does not become uninformative at this point. Accordingly, we cannot construe the ideal observer's approval and disapproval as consisting of judgments about the morality of the action in question. In section 3.2, I will spell out Michael Slote's sentimentalist account of moral approval and disapproval, and then say that something like it constitutes the ideal observer's responses. In section 3.3, I will combine what was said from 3.1 and 3.2, providing a full description of

[54] Similarly, there are situations in which normally negative relations to well-being, such as the deliberate infliction of harm, may constitute positive relations if the harm inflicted is an instance of retributive punishment. See the note on this in chapter 3.

the ideal observer's responses in each instance, bringing out the points of plausibility of this account.

3.2 Slote's account of moral approval and disapproval

In the early chapters of 'Moral Sentimentalism', (2010) Michael Slote develops an account of moral approval and disapproval. Moral judgments are constituted by moral approval and disapproval, which in turn are characterised by a specific type of empathic reaction (ibid, 27-8). From his account of moral judgments, Slote develops an account of the reference fixing of moral terms, or, the truthmakers of moral judgments (ibid). I will not go into details about Slote's truthmaker account, but I will spell out his account of moral judgments. To be clear from the outset, I reject Slote's account of moral judgments, as it rests on claims I rejected in giving my own account in chapter 3.[55] But Slote's account of moral approval and disapproval is the account that I wish to give of the ideal observer's responses, because I believe it yields some plausible intuitive results. I now turn to outlining Slote's account.

Slote discusses how his view differs from, and improves on, earlier sentimentalist accounts of moral judgment, particularly that of David Hume (1969 [1740]). The empathic reaction or feeling Hume thought of as constituting moral approval and disapproval was an empathic reaction to the recipient of actions. We know what it is like to experience pleasure as the result of other people's benevolent actions. So when we see others being benefited by people, we empathise, experiencing a pleasure at the recipient's pleasure. Such a feeling, Hume thought, constituted the basis of moral approval. Similarly, when we see the harmful result of another's action, we empathise with the pain caused to the recipient of the action, and this constitutes our moral disapproval (Slote 2010: 30-33)

But Hume's account of moral judgment, as Slote sees it, has some problems. If moral approval or disapproval is an emotion sparked by empathy with the harm or benefit being done to the recipients of actions, why do we judge a *person* who kills another person as immoral, while refraining from judging, say, a boulder that kills another person? A related question: why do we (generally) judge deliberate acts of killing as morally worse than accidental killings, even though the harm done is the same? Why do we tend to judge the

[55] In particular, Slote believes that the moral approval and disapproval – an affective response – is a necessary component of moral judgment (2010: chap 3). This is exactly the sort of thing that I (and Nichols) regard as conceptually dissociable from the recognition of a normative system of standards, as I discussed in chapter 3.

murder of a person we know and dearly love as no morally worse than the murder of a stranger in a distant country, whose harm we may not empathise with as strongly? And why is moral approval or disapproval so phenomenologically different from prudential approval or disapproval? Whether the recipient of my beneficial action is me, or somebody else, your approval should not differ in feel if Hume's account is correct. But it surely does.

Slote's account of moral judgments improves on all of these weaknesses of Hume's by making one simple adjustment: the empathy that forms the basis of our moral approval or disapproval is empathy with the *agent who is acting*, rather than the *recipient* of the action. When an agent acts for the benefit of another, they are exhibiting the same sort of empathy for others affected by beneficial actions that Hume describes. Slote says that it is *this* empathy for the recipient of their actions that the agent feels, which *we* then empathise with, that forms the basis of our moral approval and disapproval of actions. Slote says the following about this 'agential empathy':

> When we empathise with agential empathy, what we are doing is very different from what the agent is doing. The empathically concerned agent wants and seeks to do what is helpful to some person or persons… But when we feel empathy with empathically concerned agents (as agents), we empathise with *them*, not with the people they are empathising with or focussed on. We empathise, in other words, *with what they as (potential) agents are feeling and or/desiring,* and such empathy is, I believe the core basis of moral approval and disapproval… I want to say that such reflective feeling, such *empathy with empathy*, also constitutes moral approval, and possibly admiration as well, for agents and/or their actions. Ibid 34-5.

Call the empathy with those on the receiving end of beneficial or harmful actions *receiver* empathy (what Hume took to be the basis of moral approval and disapproval). When agents (deliberately) act in helpful or beneficial ways toward agents they themselves display receiver empathy.[56] When agents act in callous, harmful or self-centred ways toward others, they display a lack of receiver empathy. Agential empathy is the empathy with not only an agent's receiver empathy, but with their lack of receiver empathy.

Although receiver empathy is *not* agential empathy, our ability to experience agential empathy does depend on our ability to experience receiver empathy, according to Slote

[56] Of course, this may not necessarily be so: one might benefit another out of a rather cool sense of duty. But it is clear that such empathy *often* accompanies beneficial actions, and such an empathy is what Slote's 'agential empathy' is parasitic on.

(ibid 35-6). For if we lack receiver empathy, or are unempathetic, we will not ‘pick up on’ either the empathy or lack of receiver empathy in agents. But if we ourselves are receiver-empathetic, then the receiver-empathy (or a lack of it) of others (agents) will be on our radar, and we can empathise with them (ibid). This agential empathy will feel different, depending on what we pick up on. If we empathise with (receiver) empathy in the agent we are empathising with, we will experience a positive emotion, as Slote describes it, a ‘warmth’ or ‘tenderness’. If we empathise with a lack of (receiver) empathy in the agent we are empathising with, we will experience a negative emotion, a feeling ‘left cold’ or ‘chilled’ by the lack of empathy picked up on (ibid). These respective feelings of ‘warmth’ or ‘coldness’ constitutes moral approval and disapproval, and is also Slote’s shorthand way of describing the phenomenological difference between the two.

Though the ‘warmth’ Slote describes is commonly seen as a positive emotion, while the ‘coldness’ is seen as a negative one, Slote attempts to go further in his account of why ‘approval’ so described is positive, and ‘disapproval’ so described is negative. The answer Slote settles on is that the feelings motivate: the positive or warm feeling motivates us *toward* the actions of the agent we are empathising with, and the negative or cold feelings motivate us *away* from the actions of the agent we are empathising with. This Slote says, explains the inherently motivating power of moral judgments, which he believes to be a feature (ibid 46-7).

Here ends my summary of Slote’s account of the phenomenon of approval and disapproval that he believes to be constitutive of moral judgment. Although, as I have said, I reject the account *as an account of moral judgment*, I believe Slote’s account of the kind of moral approval and disapproval felt by those who take morality seriously (moralists) feel. I differ from him in saying that moral approval or disapproval is a necessary component of moral judgment: only those who take morality seriously will feel the kind of moral approval and disapproval Slote speaks of, but moral judgments may be made independently of this, and thus can be made by amoralists. Putting these differences aside, and given that I think Slote’s moral approval or disapproval is a good account of the feeling that attends the moral judgment making of empathetic moralists, I think it also makes a good account of what a maximally empathetic ideal observer’s responses would be.

3.3 Slote's approval and disapproval in my ideal observer

Let us now return to the four-judgment schema outlined in 3.3, applying Slote's account to each response.

> **Positive/rightness:** A's judgment about the rightness of X is true iff Ideal Observer would disapprove of a failure to do X.

An action will be right or morally obligatory if the agential empathy our ideal observer feels would result in an experience of 'coldness', were the agent in question to fail to perform the action. One might ask why *disapproval*, rather than approval, is the basic emotion referred to concerning this kind of judgment. Would not the ideal observer *approve* of the doing of obligatory actions?

I prefer to answer 'no' to this question (though not much hangs on this insistence). We do not tend to feel the warm approval Slote describes, I contend, when agents do things that they are obligated to do: things that they 'should be doing anyway'. We do not feel approval at employees for turning up to work – it is what they are meant to be doing anyway. We do not feel approval at people for refraining to steal or murder – again because we regard these as 'those things one are supposed to refrain from doing anyway'. It is no big feat to fulfil one's moral obligations: doing so does not make one morally good, but only entails that one is not morally bad (at least in respect to the particular obligations). All this is consistent with the claim that, *breaking* one's moral obligations brings a feeling of disapproval from moralists. But keeping one's moral obligations, typically, does not engender approval. Of course, we feel approval when we see people doing nice things, when we see people going 'out of their way' to help others, or making a demanding sacrifice for others. But such actions are typically seen as supererogatory (as the phrases 'going out of one's way' or 'going above and beyond' suggest). In minimally empathetic moralists like us moral approval is not inspired by the doing of obligatory actions because we believe the *not* performing them to be a violation, and a keeping of them to be 'the bare minimum'. It is likely, then, that a maximally empathetic agent like the ideal observer would react *more* seriously to actions we regard as violations of obligations, and thus be equally likely to lack positive approval for the doing actions we typically regard as moral obligations; equally likely to regard the keeping of them as a 'bare minimum'.

Of course, there are questions that can be asked about this. Isn't it possible that we – or indeed the ideal observer – would respond with approval to the performing of moral obligations in cases where the performing of moral obligations is particularly demanding? If we were alive during World War II, and we, like many at the time, believed it was a young man's duty to go and fight, would we not nevertheless admire him or approve of him for doing so, given that his life is at risk? If we witnessed a parent give birth to a child that was severely disabled and demanding, would we not approve of the parent for keeping the child and raising it? Also, the idea of demanding obligations puts to the test the idea that we would necessarily *disapprove* of those who failed to meet them. Would one *really* disapprove of the soldier who deserted, or the parent who abandoned her severely handicapped child? I think we must concede that we do tend to have a different emotional reaction to the upholding or flouting of obligations that we consider to be demanding, than those we do not. But we need not concede that this different emotional reaction constitutes a difference regarding our approval or disapproval. We would not approve of the parent for choosing not to abandon her child in a moment of weakness, just as we would not approve of someone choosing not to steal. But we may feel more pity, more sympathy, for the parent due to the fact that this particularly demanding obligation has landed on her. And we can feel such an emotion without this resembling anything like the 'warmth' that Slote describes as constituting moral approval. We may certainly feel ourselves wanting to help the mother of the child, and maybe lend a hand, due to the demandingness of her obligation. But the idea that we would feel that 'warmth' of moral approval simply at the fact that she is looking after one she is obligated to look after is not, I do not think, what we would typically feel. Nor, I think, would we refrain from moral disapproval at the flouting of a demanding duty. Take the example of the deserting soldier. If we really believed that he had an obligation to fight – if we really believed that the world was in danger if he, and others like him, did not fight, then I believe we would still experience an emotion with a phenomenological feel that constitutes disapproval. No doubt, the nature of our disapproval might be a little different to what we might feel at the flouting of an ordinary obligation. While, at the flouting of an ordinary obligation, we might feel a chill or coldness, we might feel a more pained or confused combination of emotions at the soldier's desertion. While we might feel a little cold, we also, again, may feel some sympathy: if the moral obligation to fight had fallen on *us*, we might have done the same thing and deserted. In reaction to the soldier's desertion, we may not even feel the 'chill' that Slote describes: we may experience more of a depressive or sinking feeling, one that often characterises disappointment in someone or something. Such a feeling may well attend the thought 'his obligation was demanding, but it is a shame he did not uphold it'. Such a feeling of *disappointment,* though different from what Slote describes, can be said to be a form of moral disapproval (perhaps Slote's insistence that moral disapproval consists in a 'feeling left cold' is too narrow). Thus, I think that a lack of approval of doing an action that is correctly thought by us to be right, and a disapproval of a failure to do such an action, are the right responses to attribute to our ideal observer.

Ultimately, if the reader remains unconvinced about what has just been said, this contention is not central to the theory, and one that I need not insist on. Perhaps, if better arguments favour this, the positive/rightness conditional can be appropriately altered to include a clause that, if a judgment that one has an obligation to ϕ is true, but if ϕing is demanding, the ideal observer would both disapprove of one failing to ϕ, and approve of one's ϕing. I leave it up to the reader to decide whether such a clause is necessary.

> **Negative/wrongness:** A's judgment about the wrongness of X is true iff Ideal Observer would disapprove of the doing of X.

As has already been implied, an action will be morally wrong or prohibited if the ideal observer's empathy with the agent performing the action results in a 'coldness' or negative feeling. At this point it is worth addressing a question: what makes the ideal observer's experience of 'warmth' a positive feeling, and his 'coldness' a negative feeling? Slote steers away from characterizing this as a matter of one feeling being pleasant and the other unpleasant, and instead opts for a motivational account. Even though our ideal observer is strictly an *observer* and thus cannot perform any actions, in the world he observes, we could still give a motivational account (a counter-factual one): if the ideal observer *were* able to act in the real world, his agential empathy would motivate him toward, or away from certain actions.

> **Neutral/permissible:** A's judgment about the permissibility of X is true iff Ideal Observer would neither approve nor disapprove of the doing of X.

This needs some qualification. Morally permissible actions that I may do to benefit myself would no doubt please the ideal observer, given that he is maximally empathetic and desires that every agent, including me, be well off. And morally permissible actions that I might do that harm me would distress the ideal observer for the same reason. However, his emotional reaction is still distinct from the kind of moral approval and disapproval characterised by agential empathy. Recall, when the ideal observer approves of an agent's action in Slote's sense, it is because the ideal observer is empathising with an empathy for others displayed by the agent performing the action. But in prudentially beneficial (morally permissible) actions, no such empathy with empathy is present. Thus the emotion that the ideal observer has as he sees a prudentially good action is, (as it typically is for us) quite phenomenologically different from the approval he has of morally good actions. So when the above statement says that the ideal agent neither approves nor disapproves of morally permissible actions, this is not to say that he has no positive or negative reaction at all: he just won't have the *specific kind* of positive or negative reaction that moral approval or disapproval is constituted by.

As we can see, this is only one way in which Slote's account helps my ideal observer sentimentalism get good intuitive results as a theory of truthmakers. On the account given, it will turn out that actions that only benefit oneself are prudentially, but not morally good: an intuitive result that we should want, given the conceptual distinctions between morality and prudence drawn in chapter 3. But the other advantages of Slote's agential empathy account carry over as well: it will turn out that it is morally wrong for a person to kill another person, but not morally wrong (though sad) when a boulder does. Similarly, it will turn out that accidental killings are (unless the agent can be accused of negligence) not as morally bad as deliberate killings.

> **Positive/goodness/supererogatoriness:** A's judgment that X is supererogatory is true iff Ideal Observer would approve of the doing of X.

As has already been said, the ideal observer will experience a positive emotion involving the distinctive 'warmth' Slote mentions as being at the heart of typical approval sparked by agential empathy. Although, once again, the ideal observer will be pleased on seeing an agent greatly benefit herself, the ideal observer will experience a different kind of positive emotion at seeing an agent greatly benefit another – especially if she does this at great cost to herself. For the agent, being maximally empathetic and well-wishing of all agents, will be saddened by the cost to the agent, but will at the same time be 'warmed' in picking up on her love for others that she has in common with him.

4. Strengths of Ideal Observer Sentimentalism

4.1 Irrelevance Objections

In this section, I wish to outline some advantages of ideal observer sentimentalism. This will involve answering potential objections that confront similar views, and showing such objections to be ineffective at undermining ideal observer sentimentalism. The first kind of objection I wish to look at are what I call 'objections from irrelevance': the idea that the responses of an ideal observer so unlike actual human moral agents would be ethically irrelevant for actual human agents.

A significant critique of the Firthian ideal observer comes from Carson (1981: 57). The ideal observer's *omniscience*, (including his omnipercipience) is incompatible with the ideal observer's being *human*: something Carson believes is an important characteristic. It seems that Carson believes that the ideal observer needs to be human in order for his responses to generate an ethical code appropriate to humans, for the content of moral codes will differ across different types of creatures. He makes this comment:

> Suppose that there were Martians who, aside from not being human, possessed all of the other characteristics of ideal observers. Suppose also that the views and attitudes of the Martian ideal observers differed from those of human ideal observers on numerous matters. From the standpoint of the ideal observer theory there is no reason to prefer the views of the one to the other. Rather we must conclude that the views or attitudes of the Martian ideal observers are correct for the Martians and that the views or attitudes of human ideal observers are true for us. 1981:76

There is quite a sensible element in what Carson is saying here. This point that moral codes vary for radically different creatures with radically different things in their wellbeing is quite plausible. But Carson's observation about the relativity of moral codes to the types of creatures they apply to does not support his claim that the ideal observer needs to be human in order for his responses to be morally relevant for humans. For the *reason* that the content of moral codes differs between different kinds of creatures is because different kinds of creatures' welfare is different. If there was a moral obligation Martians had to stab each other in the stomach, for instance (while such an act would be prohibited for humans), this would presumably be because these Martians were constituted such that stabbing each other contributed to their welfare. Given this, we see that the ideal observer does not need to be *human* in order for his responses to be relevant to humans, only that he must care about, and be knowledgeable about, human welfare – a requirement that my ideal observer sentimentalism meets. This need not involve him being human as such.

One question, of course, raises its head: if facts about welfare determine the ideal observer's responses in this way, then how is the ideal observer to react when two, irreconcilable interests clash? Suppose there came into existence a race of sentient Martians whose welfare was heavily dependent on devouring human beings, in much the same way as quality relationships are essential to human welfare? The well-being of the Martians demands that the lives (and relationships) of humans are shattered, and the well-being of humans demands that the lives and relationships of humans persist. How would a maximally empathetic agent who desires the well-being of all agents respond to this? I think the correct answer here (assuming all things are equal, including the number of humans versus the number of Martians) may well be that the ideal observer would have an irresolvable clash in his desires, and thus there would be ethical indeterminacy with regards to what ought to happen; either ethical indeterminacy or no moral fact of the matter. I do not consider this a weakness of ideal observer sentimentalism that this result occurs: given that morality is a worldly, human system of norms, one would expect the system to somewhat disintegrate when such bizarre, other-worldly hypotheticals are introduced.

A more serious problem, perhaps, is that there might be irreconcilable clashes of interest among humans in the *actual* world, where morality is supposed to apply relatively smoothly. If this were to occur to a great extent, the result could be what Joyce calls 'accidental error theory about morality'[57]: perhaps the very concept of morality presupposes a certain degree of determinacy or lack of indeterminacy, and the possibility of widespread and persistent clashes of interest undermines this presupposition (note that this is a very different error theory argument to the main one addressed in this thesis: the idea that moral judgments imply the presence of external reasons). I will address this threat of 'accidental error theory' later on in this chapter. For now, I will look briefly at another kind of objection that confronts the ideal observer theory.

4.2 Relativity

From the outset of this project, I said that I wished to avoid the kind of moral relativism according to which the correctness or incorrectness of one's moral judgments is purely a matter of consistency with one's own opinions (note that a rejection of this kind of relativism is compatible with an acceptance of a different kind of relativism: the idea that moral codes vary when the welfare of types of creatures to whom they apply vary, an idea that I think is better described by the label 'particularism'). But a challenge that Firth's ideal observer theory meets from Brandt is precisely the challenge that his ideal observer theory leaves open the possibility that different ideal observers – an idealised you or an idealised me – may have different responses. If this is the case, then a relativism according to which my judgments are correct with reference to my opinions (albeit my idealised ones), and that yours are correct with reference to yours. This kind of ideal observer relativism inherits the same problems that simple subjectivism or relativism inherit: namely that there is no such thing as moral disagreement. It is desirable, then, that my theory not inherit this problem.

Fortunately, I do not believe my theory suffers from this kind of relativist threat. Suppose that idealised you and idealised me both came to have the characteristics of the ideal observer as I mentioned. Suppose that we have radically different moral upbringings, radically different interests and personal attachments, and that our knowledge of empirical facts covered very different ground: all these factors account for the differences in our moral opinions. Suppose that we become 'idealised': suppose our minds expand to become aware of all empirical facts, all facts about welfare, all feelings and perspectives of sentient beings. Now this alone may not guarantee convergence in our moral opinions: it may just

[57] See Joyce, 2011 for a discussion of this.

ensure that we are both more ingenious propagandists of our own moral views, attached just as much as before to our own vested interests. But suppose that this knowledge is accompanied by the other feature: a strong desire far above everything else, that all interest bearers be maximally well-off. Suppose also, as was said above, that we lose all other desires: we have no aesthetic desires, no other interests which might clash with our desire to see everyone better off, no personal attachments which outweigh our equal love for everyone. I think it is very hard, now, to say what possible source of difference there could be in our moral opinions. The combination of our omniscience and our benevolence would render any past, incomplete, moral framework we were raised with irrelevant to us. Our vested interests would be gone, as would our false beliefs. The ideal observer, whether an idealised you or idealised me, would have the same responses.

It is important to be specific about what is being suggested here, lest a charge of incoherence arise. It is not being suggested that every actual individual, simultaneously, could become an ideal observer, and that a plausible objectivist moral code could come out of this. If the world were populated with ideal observers who had no desires but for the well-being of others, it is hard to say that the ideal observers would have any interests at all. What is being suggested, rather, is that when we consider different possibilities – a world in which I become the ideal observer, a world in which you become the ideal observer, a world in which John becomes the ideal observer, and so on for every individual – are worlds in which the ideal observer has the same desires (the only difference in desires being that each world has a different individual missing whose interests to care about, of course!).

Of course an idealised you or me would be so different to the actual you or me that it is almost senseless to talk about the ideal observer being a 'you' or 'me' at all. But this ought not bother us. The ideal observer's omniscience, empathy, desires, impartiality, and perfect inference making captures to the full extent every attribute that we regard better moral judges to have. At the conclusion of this section, it will be helpful for me to answer two objections that may have come up through the reading of this and previous chapters.

1) If 'an idealised you or me would be so different to the actual you or me that it is almost senseless to talk about the ideal observer being a 'you' or 'me' at all', then

> why should I care about doing what the ideal observer would have me do? What is the relevance of his desires to mine?

Possibly nothing. Given the potential difference between the ideal observer's motivational set and yours, the theory of practical reason spelled out leaves open the possibility that you may have no reason to act as the ideal observer would have you act. But this is simply equivalent to the claim that you would have no reason to do as morality dictates, a possibility we have already left open. Given the arguments from chapter three that amoralists can make moral judgments too, it is quite possible that you may agree with any moral judgments that come from ideal observer sentimentalism, but be unmotivated to, or think yourself lacking reason to, comply with them. Thus, given our previous arguments against the rationalist conceptual claim, the success of ideal observer sentimentalism does not rest on whether or not the ideal observer's desires would be reason-giving or motivating for you.

On the other hand, it is worth adding that it is highly unlikely that you as a human being would share *none* of the ideal observers concerns, given the likely existence of empathic elements in your motivational set. So, even though the success of ideal observer sentimentalism doesn't depend on its potential to provide you with practical reasons, it will nevertheless probably be the case that the theory does provide you with reasons some or much of the time, due to contingent facts about your motivational set.

2) If your theory of welfare allows that, if we have a malicious motivational set, then performing malicious actions are good for us, doesn't this entail moral relativism?

No. Judgments about what an individual's welfare consists in are distinct from moral judgments. The facts that make judgments about an individual's welfare true will have to do with facts about their motivational set. But facts that make moral judgments true will have to do with facts about positive or negative relations *between* individuals, as spelled out by the ideal observer sentimentalist theory. As I have argued, since there happen to be broad similarities and harmonies between actual individuals' interests, indeterminacy is limited. For relativism to be entailed, it would have to be the case that moral judgments are judgments by our ideal selves who care only about our interest. In this scenario, my taking what is yours could conceivably be 'right for me (and my ideal self)' and 'wrong for you (or your ideal self)'. But this is not the kind of ideal observer theory I have presented.

4.3 Moral agreement and disagreement

An advantage of Ideal Observer Sentimentalism is that it makes good sense of both the degree of moral agreement and disagreement that exists among actual agents. If the responses of our ideal observer determine rightness and wrongness, then the amount of agreement and disagreement that exists is unsurprising.

It is very hard to find any culture in the world in which wanton murder of a community member (murder for the fun or enjoyment of it) is not condemned. Similarly, all cultures in the world have rules regulating the distribution of property and possessions, and punishments for theft, when the violations of those rules take place. Similarly, almost every culture in the world regards self-sacrifice or the helping of others, as virtuous (even if not obligatory). These acts the moral status of which is widely agreed on are perhaps the same acts which it is most uncontroversial what our ideal observer's reactions would be. I think it is quite obvious that the ideal observer as I have described him would disapprove of the theft and wanton killing, and would obviously approve of self-sacrifice.

In the same way, ideal observer sentimentalism makes equally good sense of moral disagreement. There are some moral issues which seem irresolvable, and over which, as a consequence, moral disagreement persists. Debates about the moral status of abortion, retributive punishment, going to war, and cloning, are all examples of such debates. And with regard to the status of all these actions, it is quite unclear to us what the ideal observer's reaction would be, even if there would be a particular reaction.

Furthermore, although there may be a way of working out what the ideal observer's reactions to such actions might be, this may nevertheless very hard for us humans to implement. In order to work out what the ideal observer's reaction would be to a given action, we might try to work out a 'logic-of-attitudes' (or desires) *a la* Blackburn,[58] starting with the ideal observer's empathy-driven-desire for everyone's well-being, and ending up with a specific reaction to a specific action. But such a philosophical task may well be

[58] For a discussion of Blackburn's attempt to provide a logic of attitudes, see Schroeder 2010, and Blackburn 2006a and 2006b

impossible for humans, who, after all, *do not* possess the level of empathy of the ideal observer, and who thus would find it very hard to deduce any logic of attitudes from such an empathy-driven-desire. The point I am making is this: ideal observer sentimentalism implies that there *are moral facts*, even where debates over those facts may be in principle irresolvable. A special strength of ideal observer sentimentalism, then, is that it is immune from a sophisticated version of the argument from disagreement. This version of the argument claims that, if a moral question is *in principle* irresolvable, then there is no fact of the matter.[59] Ideal observer sentimentalism provides a counter-example to this argument: there may be, in principle no way for human arguers to resolve the question of the moral status of, say, abortion in a particular circumstance. But this does not rule out there being *a fact of the matter* about the status of the abortion: for even if humans are (in principle, due to our lack of similarity to the ideal observer), in no position to apply a logic-of-attitudes deduction to the ideal observer's genuine desires, the ideal observer himself (especially with his infallibility with regard to exercises in logic) may well be. The fact that ideal-observer sentimentalism provides a counterexample to this sophisticated version of the argument from disagreement renders it a theory that serves well the purposes of moral success theorists.

4.4 The moral status of mental states

Another advantage of ideal observer sentimentalism worth highlighting is the fact that it gets good intuitive results. We have already seen the way in which it makes sense of the fact that something can be prudentially good, and the fact that deliberate killing is worse than accidental killing. There is another intuition, not yet discussed, that ideal observer sentimentalism allows for: the intuition that certain thoughts or mental states can be morally evaluable.

The prime example of mental states that we might think can be morally good or bad are thoughts. 'Bad' thoughts might include racist or sexist thoughts. An example of a morally good thought may be the thought we might have on witnessing another person do something generous, and thinking to ourselves 'They are admirable for doing that – I should do likewise'. Thoughts are deemed to be morally good or bad in that they reflect aspects of

[59] I take this description of the argument from Brink, 1984:115-7

our character: intuitively, my morally bad thoughts reveal some morally bad aspects of my character, and my morally good thoughts reveal some morally good aspects of my character. If my character is an appropriate target of moral evaluation, then so are my thoughts.

Ideal observer sentimentalism gives a nice, unified account of what makes both actions and mental states (like thoughts) morally good, right, bad or wrong. As has already been said, a thought is bad or wrong if our ideal observer would disapprove of it: if, through agential empathy, he would be ‘chilled’ or ‘left cold’ by an uncompassionate thought as he would an uncompassionate action. A thought is good if our ideal observer would approve of the thought: be ‘warmed’ by our having the thought. Ideal Observer sentimentalism, then, gives quite a simple, elegant and intuitive justification of the wrongness or goodness of thoughts.

What about thoughts that are ‘right’ or ‘obligatory’ as opposed to good? Here, the view might initially look like it meets some difficulties: for one can hardly be criticised for not having a certain thought, even if one ought to. Suppose it would be right for me to have a thought that John is a good person. But suppose, for the whole week ahead of me, I fail to think that John is a good person: there are simply too many other thoughts in my head – I am thinking about my exam, my trip to Vanuatu after my exam, my cat’s strange cough that he has developed and whether I should take him to the vet. Because of my headful of other thoughts, I simply fail to think that John is good. But it seems like the wrong result for ideal observer sentimentalism to entail that I have done something wrong by not thinking this thought.

However I think there is a relatively simple answer to this problem: the ideal observer would not disapprove of me for failing to have the thought that John is good, but *would* disapprove of me thinking something that amounted to a *denial* that John was good (suppose I personally didn’t like John, and my personal dislike of him got in the way of me having an admiration for him that I ought to have). A simpler answer still is that the ideal observer would know whether or not I am *disposed* to think that John is good, and so would evaluate my character on this basis. Once we see that my obligation is to refrain from thinking a certain way about John, rather than to think a certain way, we see that the unity in ideal observer sentimentalism is maintained even here: the ideal observer disapproves of us if we flout our obligations.

I consider it a virtue of ideal observer sentimentalism that it is able to make sense of the wrongness of thoughts in a way that is not only elegant, but which seems to get to the 'real reason' many of us are inclined to think that thoughts are morally evaluable. When we imagine someone thinking a racist thought, or a spiteful thought, we feel repulsed (indeed, 'chilled'). And when we know or imagine that someone has warm sentiments or loving intentions, we feel warmed. Ideal observer sentimentalism thus gives an explanation for the rightness and wrongness of thoughts that is very close to our intuitions on the matter: we judge thoughts of people as bad when our thought of those thoughts repulses or chills us, and as good when our thoughts of those thoughts warms, encourages, or uplifts us.

4.5 Accidental error theory and indeterminacy

In a previous section, a potential problem for ideal observer sentimentalism was raised. Suppose there are two very different creatures: a Martian, whose wellbeing depends entirely on devouring a human being, and a human being, whose welfare obviously depends on this not happening. Suppose all other things are equal: suppose both the human and the Martian are hermits, and therefore that neither has any more loved ones than the other who would indirectly be hurt by either outcome. And suppose the existence of either one of them will make no positive difference to the well-being of others. I said that, in a case like this, there may be genuine ethical indeterminacy: the ideal observer would have an irresolvable clash of desires: that both be maximally well off, even though this is impossible.

A case like this, I suggested above, is no real problem for morality. It should not be surprising that the rules which make up the very human concept of morality 'break down', so to speak, when such other-worldly hypotheticals are introduced. Morality is not really intended to govern the relationships between humans and Martians, but only the interest bearers that humans are familiar with and have interactions with: other humans, animals, and, (as far as theists are concerned) God. But we would have a problem on our hands if something like the human-Martian scenario were replicated among humans on a regular and frequent basis. Let me offer some examples. Suppose we have a psychopath, who wishes to maim and kill a victim. Might we have two irreconcilable interests here: the interest of the psychopath to kill, and the interest of the victim to not be killed? Suppose we have a child pornography addict whose entire life is consumed with this addiction. The porn that is being produced unquestionably damages the children. But if it were equally damaging to the porn addict to not consume, might we have an irreconcilable clash of interests here? The problem is as follows: according to ideal observer sentimentalism, moral norms are generated by the responses of a maximally empathetic observer who desires the maximal satisfaction of interests. But if there are too many irresolvable clashes in interests, like what we see in our Martian case (and given that there are many child porn addicts and psychopaths in the human population, it is worth considering this risk), then there might be a degree of ethical indeterminacy that essentially warrants another form of error theory about morality. Morality, it could plausibly be said, presupposes a large degree of determinacy about its facts (even if there is room for indeterminacy here and there in some dilemmas). But if such determinacy is not to be had, then once again, we find that moral discourse is conceptually committed to something that reality does not vindicate.

Fortunately, I do not think that ideal observer sentimentalism (or many normative ethical frameworks) entails the kind of ethical indeterminacy that would vindicate what Joyce calls ‘accidental error theory’ (2011). This is mainly due to fortunate contingencies about human wellbeing, which were discussed in chapter 4. It is unlikely that the two examples involving the psychopath and the child porn addict resemble the indeterminacy-generating genuine clash of interests that the Martian and the human have. For indeterminacy to result, it must at the very least be the case that the pornography addict’s well-being is *equally* harmed as the child’s well-being in the production of the pornography. Likewise it must be the case that the harm done to the psychopath in denying him a kill is *equal* to the harm done to the victim. Given the general facts about human welfare discussed at the end of chapter 4, it is very unlikely that this is the case.

For a start, any behaviour that revels in the harming of another person can plausibly said to be bad for one: even for the coldest porn addict or psychopath. For, even if a porn addict has totally numbed his empathy such that he feels no guilt about consuming child porn, being in such a dead state with regard to the feelings of other humans would render him cut off from being able to experience a great deal of positive affect if he were more caring, if he were more sensitive. The fact that he *would* experience more positive affect if he were to cease his addiction entails, according to the theory of welfare given, that he has a large number of (potential or actual) desires for more fulfilling interactions with other humans than what his addiction offers (the sexual pleasure gained from the experience notwithstanding). Plausibly, then, the porn addict would be much better off if he ceased his addiction, despite his current state of coldness, and probable opinion to the contrary. It may be somewhat harder to say this of the psychopath, though it may not be completely implausible. The psychopath's coldness similarly cuts him off from a range of positive emotional experiences he could have. So here is one possible answer to our problem: there is unlikely to be a genuine irresolvable clash of interests between humans who engage in extremely destructive behaviour toward other humans, and their victims, because such behaviour is very likely bad for the perpetrator anyway.

But suppose the case could be made that the psychopath's behaviour is *not* bad for him. Suppose the case could be made that psychopaths just do not have the set of potential desires that normal human beings have: psychopaths are *not capable* of experiencing the positive affect that the rest of us feel from loving interactions, thus they have no potential desires for it, in much the same way that a dog does not even have a *potential* desire to watch the storyline of a movie unfold. Without being able to defend this view empirically, I think even this claim is subject to doubt: one must wonder what motivates the psychopath to act out their destructive behaviours, if not some form of dissatisfaction, boredom, or restlessness generated from the emotional alienation from others that characterises their experience of life. If this is right, then it may well do a psychopath a lot of good to curb their behaviours, to try to expand their feelings toward others, and minimize such restlessness.

But suppose I am wrong about this, and suppose that psychopaths are not less well-off human beings than others for behaving as they do. It is still unlikely that we have a clash of interests between the psychopath and his victim resembling that of the Martian and his. Even if we can grant that a psychopath may not be harming himself by cutting up an unwilling victim, it is very unlikely that his well-being *depends* on his doing so, and that he would be harming himself as severely as the victim is harmed if he *did not* kill him.

The summary point is this: it is very unlikely that there will be conflicts of interest that are both extreme and genuinely irresolvable on a frequent basis, among human beings. Given the fact that a large component of human welfare (even, arguably, that of the psychopath's) consists in benefitting others and being in satisfying relationships, this entails some degree of what might be called 'harmony' among all our interests: it benefits you to benefit me, and not just indirectly. Such a contingent harmony among human interests is well-suited to insulating morality against the extent of indeterminacy that would undermine moral success theory.

This is not to deny that there are frequent conflicts of interest between humans: there are conflicts over property, one set of interests against others, as we have in familiar 'trolley' scenarios, and much competition that occurs in ordinary life. But it is plausible to think that such conflicts could be resolved: if not through comparisons of consequences, then through desert-based principles, all of which could plausibly be thought to arise from the reactions of a maximally informed and empathetic ideal observer. The above statements about the harmony of human interests is not intended to paint a rosy or unrealistic picture of the world in which conflicts are never present, or are merely illusory. The only point I wish to emphasise is that the interests between human beings do not diverge so radically that the ideal observer would be left so irresolvably 'torn' on a basis so regular that the extent of ethical indeterminacy would be so great as to warrant error theory (as it would if most clashes of interests resembled that of the Martian's and the human's). Although there are many genuine clashes of interests, I believe it is plausible to think that, for most of them, there is an ethical fact of the matter as to how they ought to be resolved, and that ideal observer sentimentalism is consistent with this assumption.

What about the other two elements: relationships between humans, animals, and God? It has already been said that any harm animals may do to humans or to each other is not really morally evaluable anyway, since they lack agency. And with regards to God, the least that can be said is that many traditions describe God as a being whose interests harmonise with human interests in the way already described. Many traditions have it that God gives us certain guidance about how to live. Following such guidance enhances our existence and wellbeing, and enhances God's as well: by both giving him the worship he as a benevolent creator is entitled to.

In short, morality does not presuppose a degree of determinacy that it cannot deliver on. Because of facts about human welfare (and purported facts about God), there is a relative harmony of interests between moral agents in the actual world. Some may be worried about the contingency of this harmony: if our interests were not so harmonious, (and were more like the Martian's and the human's), then we would not have the moral facts we do. But I do not think this worry about contingency ought to be taken seriously. Yes, it is a contingent fact that our interests are as they are. But it is an equally contingent fact that the concept of morality arose as it did. If interests were not already somewhat harmonious, it would of course be unlikely that the concept of a system of norms which existed for the purpose of protecting that harmony would arise.

4.6 Room for further work

Here ends my account of the truthmakers of moral judgments. In chapter 3 it was said that moral judgments are judgments about the positive or negative relations between the thing morally evaluated by the judgment, and the well-being of others. In this chapter I have given an account of what all positive relations have in common, and what all negative relations have in common. If an action is positively related to the well-being of another, then the ideal observer described in this chapter would have Slote's feeling of approval toward the action. If an action is negatively related to the well-being of another then the ideal observer described would have Slote's feeling of disapproval.

Accordingly, moral judgments are true if the thing judged *really does* bear the positive or negative relation to well-being in question: if an ideal observer *would* have the requisite feeling. Suppose I judge that it would be right for the doctor in the famous scenario to cut up one healthy patient and farm out his organs to five dying patients: I judge the promotion of interests to constitute a *positive* relation to the well-being of others, while others might judge the infringement of the innocent patient's rights it to constitute an overall *negative* relation to well-being. Which judgments are true? According to the theory presented, the ideal observer's reactions determine the answer to this question. If the ideal observer would react to the doctor's cutting up the patient with disapproval, then the action is negatively related to the well-being of others, and my judgment to the contrary is false. If the ideal observer would react to the doctor’s cutting up the patient with approval, then the action is positively related to the well-being of others, and any judgment to the contrary is false.

Ideal observer sentimentalism of course leaves much room for debate about which actions are right and wrong. As I have suggested, a good metaethical theory should do just this, and should not answer too many first-order questions too decisively. Although there are a number of pretheoretical intuitions about morality that ideal observer sentimentalism makes good sense of (such as the distinction between morality and prudence, and the intuition that inanimate objects cannot be morally culpable), there are many first order ethical questions that require much more discussion than what can be offered here, in order to be resolved.

If ideal observer sentimentalism is to be applied to the solving of moral questions, further work would need to be done on what reactions (in any given moral dilemma) the ideal agent’s empathy-driven desire for maximal flourishing would generate. If it is possible to formulate some likely logic of attitudes, responses, or desires of the ideal agent, this may turn out to be a fruitful pursuit for normative ethicists. Of course the idea of such a logic of attitudes is a difficult and complex one to pursue, and would be so particularly in the ideal observer’s case. Nevertheless, the potential for further work and application of ideal observer sentimentalism to ethical debates remains.

6: Conclusion: Morality and Rationality

1. Recap

1.1 Morality and rationality

I began the thesis asking a question about what sense in which morality is normative. There were two senses in which morality could be said to be normative, that were discussed in chapter 1. Morality could be said to be normative in the sense that moral judgments entail judgments about reasons for action, or moral judgments could contain propositions about standards. If morality is normative in the former way, then error theory looms: for moral judgments would have to be judgments about *external* reasons which, as was argued in chapter 2, do not exist.

The arguments of chapter 3, if successful, yield the result that the standard-based view about morality gives us everything we could need from a theory about what moral judgments are. Morality is delimited by content (as crystalized by the moral consideration claim), and the rationalist conceptual claim is superfluous. In chapters 4 and 5 I argued in favour of my preferred views about welfare, and my preferred theory of the truthmakers of moral claims. I hope, then, that what has been written so far amounts to a plausible moral success theory: a success theory that is not reliant on the truth of the rationalist conceptual claim, and thus avoids the threat of error theory spelled out by Richard Joyce and John Mackie, and with which (as we saw in chapter 1) several other philosophers are concerned.

A lot has been said so far about the relationship between morality and rationality. Moral judgments need not imply judgments about reasons on pain of conceptual confusion (nevertheless, we could assume that most users of moral discourse typically use it to make implications about reasons). There is good reason to think that moral behaviour plays a big part in one's own welfare, and therefore that moral behaviour is rational at least much of the time. Moral judgments are standard-based judgments, which are distinct from judgments about reasons. A standard-based judgment only entails a judgment about a reason if the judgment maker endorses the standard in question or takes it to be important. The practical reasons that any given agent has are relative to her motivational set, but the rules of morality are not: morality is inescapable, 'non-evaporable', and its standards are set by the responses of an ideal observer.

Having said this much about the relationship between rationality and morality, one may be left wondering whether such a thesis has done morality justice and 'taken it seriously enough.' A strong intuition behind the rationalist conceptual claim, after all, is that morality is special in a way that other norms are not: that morality has more normative 'oomph' than other normative systems. The rationalist conceptual claim does justice to such an intuition in a straightforward way. But does a standard-based success theory do the same justice?

1.2 Intuitions about morality's special value

It is difficult to know just precisely what is being meant by the idea that morality is supposed to be different from other normative systems. I will attempt now to spell out a few different things that this could mean.

Firstly, what the idea could be getting at is that it is a fact that our feelings of being under moral requirements are phenomenologically different from our feelings of being under other requirements. The requirements of morality feel more weighty, they may exert more motivational pressure than, the requirements of etiquette. Perhaps the intuition in question is that the phenomenology of moral experience deserves an explanation, and that the rationalist conceptual claim is best placed to give that explanation.

Secondly, as was mentioned in chapter 2, perhaps what is desired by the proponents of the rationalist conceptual claim is that the result that he who flouts morality harms himself.

Our natural dislike for felons produces in us something of a desire for cosmic justice: a desire that villains get, by some necessity, what they deserve. It is natural, then, to think that it is desirable that moral criticisms are 'propped up', so to speak, by rational criticisms. If moral criticisms of people aren't bad enough already, surely the truth that morality is by necessity a species of irrationality ensures that they are sufficiently bad! We want our moral criticism in and of itself to sting and punish our targets – and the idea that moral judgments are judgments about reasons might do the trick. The worry, then, might be that the standard-based view does not vindicate any of these desires of ours. If morality is a 'mere' system of standards, then there is not much of a 'sting' in moral criticism: the sting of the statement that he who flouts morality hurts himself.

Now, a question that could obviously be asked is: why think that any of these intuitions about morality ought to be vindicated? The question is not unreasonable. Suppose our awareness of moral requirements does (for most of us) have a different phenomenological feel. Why think that this is something that needs to be explained, let alone justified? And maybe we would like it to be the case that he who flouts morality also hurts himself. But so what? What if we just have to deal with the fact that the existence of moral facts does not entail that the universe is just or turns out as we would like. What if we are demanding that moral facts do something that only the existence of God could do?

While these are all sensible questions, for the purposes of this last, short chapter, I will not pursue them further. What I will do instead is proceed on the basis that these three intuitions are intuitions about morality that it would be desirable for a good metaethical theory to vindicate, and show that the standard-based view (along with what has been said about welfare and the truthmakers of morality) vindicates them. Some intuitions will not be able to be completely vindicated. But in these cases, I will point out that an adherence to the rationalist conceptual claim will not vindicate them either, and will certainly not do a better job than the standard based view, of vindicating them. Over the next few sections I will make some broad brush observations about morality that can be drawn from the conclusions of the previous five chapters, and show that they do, indeed, do morality justice in the ways that our intuitions want.

2. Reasons for Individuals and Groups to be Moral

2.1 Giving and receiving

The idea that acting in the interests of others benefits oneself does not need much defence. We secure our basic needs, such as food, shelter and safety, far better if we cooperate with each other than if we go it alone. Such an idea is at the heart of most evolutionary accounts of how moral intuition and behaviour arises.

But there is another sense, as has been seen in chapters 3 and 4, in which moral behaviour may be beneficial. At the end of chapter 4 it was claimed that loving relationships with others are central to human wellbeing, as was compassionate behaviour. Being focussed on the well-being of other people, and thereby enjoying a sense of significance, meaning, and intimacy, is a major source of wellbeing for arguably any person. Given that moral behaviour (behaviour in accordance with what one believes to be moral) will involve behaving in a way that one sees (by our welfarist definition) to be 'positively related' to the well-being of others, it is very likely that the morally good life: a life in which one takes morality seriously, and endeavours to act in accordance with one's moral beliefs, is an essential (even if not sufficient) component of the best life from a well-being point of view. Given that something's being in our welfare indicates (according to both the theory of practical reason, and the theory of wellbeing that I have given) that something is reason giving, then it will turn out that there are, at least much of the time, great reasons for individuals to be moral.

Besides the findings surveyed in chapter 4, Nichols' observations about moral judgments covered in chapter 3, and Slote's account of moral approval and disapproval in chapter 5, indicate that our empathetic nature grounds many reasons to be altruistic. When most of us internalise a normative system prohibiting harmful behaviour (or, on my broader view, prohibiting behaviour that relates negatively to the wellbeing of others), and commending actions which benefit others (or have an otherwise positive relation to wellbeing), we do so as a result of affective mechanisms. Behaviour related negatively to the wellbeing of others is likely to cause negative affect, and behaviour related positively to the wellbeing of others is likely to cause positive affect. Given that (according to my theory of welfare) such positive and negative affect is likely to indicate the existence of desires that positively-related actions be done, and negatively-related actions be refrained from, it is plausible that one's making of a moral judgment likely indicates that there is an internal reason present for one to act in accordance with it. Once again (and as I made this point in chapter 3), this is not an affirmation of the rationalist conceptual claim. Making a moral judgment need not entail that a judgment about one's reasons are present, but it may nevertheless indicate what one's reasons are.

These facts about morality go quite a long way to vindicating one of our desires about morality: that immoral behaviour is a form of self-harm. Of course, the above does not show that there is a perfect, one-to-one correlation between immorality and self-harm. But, arguably, the rationalist conceptual claim cannot achieve this result either, or cannot do so convincingly. If we are to argue that moral judgments constitute judgments about reasons, we have two choices. We may argue that moral judgments amount to judgments about internal, or external reasons. We have seen that an internalist understanding would amount to Harman's kind of moral relativism, which was rejected in chapter 2. But an externalist version of the rationalist conceptual claim would not suffice either. To claim that one has external reasons to behave morally amounts to the claim that one has reasons to be moral *regardless* of how this might be related to one's self interest. It is not clear, then, how the rationalist conceptual claim on an externalist interpretation would vindicate our desires for immorality to amount to a form of self-harm at all. It seems, then, far more promising to claim that, although there is no conceptual link between the making of a moral judgment and the making of a judgment about reasons, it may be contingently true that we have reason to be moral at least most of the time, or that we have reasons to take moral considerations seriously throughout our lives as a whole.

The fact that we are empathetic, the fact that we care so instinctively about others, explains another feature of morality: namely, why it is that moral requirements typically feel so much weightier than others. The reason that morality is felt or taken to be so much more important than etiquette, fashion, or prudence, is because the well-being of others – morality's prime concern – is felt by most ordinary people to be important too.

There may not always be a perfect correlation between the morality of a given act, and the rationality of it. Indeed, there are instances in which the morally right action and the prudentially right action may diverge. But there is good reason to think that orienting one's life priorities around moral concerns is, in a general sense, quite beneficial.

2.2 Morality as the hand that feeds us

There is another feature that morality has which explains (and indeed vindicates) our feeling that it has a special value. Such an idea is expressed in a contention made by David Copp: that an action is morally right or good if it would be rational for a society for such an action to take place within it (1995: ch 6, ch 8).

Although I see problems with Copp's particular idea (the idea that societies or groups of people could be rational choice makers in the right sense would be something I would take issue with), it is clear that there is an important kernel of truth in it. Morality, personified, is *concerned* with the wellbeing of the agents to whom it is directed. If everybody did as morality dictated, or if everybody, at least, was concerned to act morally, there would be much suffering that would be eliminated. If it were possible to conceive of every interest-bearer being a small part of a greater whole, then this greater whole would have prudential reason to have each of its parts act morally. Once again, I do not believe the aggregation of interest bearers *does* constitute a 'whole'. Nevertheless, this metaphor is an interesting way of bringing out a certain point: not only do single individuals benefit if they are individually morally good, but *the whole aggregation of interest-bearers* would benefit if *every interest bearer is* morally good.

This intuition explains well the phenomenological experience of being under a moral obligation. To flout one's moral obligations is to flout the rules of that system of norms which has everybody's interests at heart, *including one's own interests*. It is little wonder, then, that the experience of flouting a moral obligation often feels like one is biting the hand that feeds us. For the fact that many interests, including our own, are impinged on has to do with the fact that immoral behaviour is performed. If I, then, behave immorally, I experience an acute awareness that I am part of one of the world's biggest problems, a problem that effects everybody including myself. Another way to put the idea is as follows. We all want to be well off. But us all being well off is dependent, to some extent, on our not getting in each other’s way, and our behaving morally. To behave immorally, then, resembles a kind of ‘letting the team down’ (the ‘team’ being the aggregation of all interest bearers) and so, the pressure of the moral obligation bears many similarities to the pressure of not letting down the goal of a team in which one is a part.

Of course, our doing our moral duty does not always *directly* benefit us – sometimes doing our moral duty can be a great sacrifice. But it is common, even in these situations, to feel a pressure to perform demanding moral obligations, that has a basis in the idea that morality looks out for the interests of humanity as a whole, of which we ourselves are a part. Such a thought is powerful enough to induce us to perform demanding moral obligations even in instances where we stand to lose something: we still feel that we ‘owe’ something to morality and to our fellow humans, even when our obligations are personally demanding. This feeling of ‘owing’ something to morality can be well explained by the following reason: because moral behaviour is a precondition for the well-being of humanity of which we are a part, we must ourselves behave morally if this precondition is to be realised, even when, on a particular instance, the burden is heavy.

This aspect may further explain the far reaching of the scope of morality, and the degree to which its ‘inescapability’ is felt. While certain normative systems may only apply to one because one chooses to opt into it (think of the rules of a club, or the rules of a game that one chooses to play), there is no opting into, or opting out of morality. One is an interest bearer whether one likes it or not, and the success rate at which moral rules are followed by moral agents has a bearing on one’s own well-being, whether one likes it or not.

Given all this, there does turn out to be quite a significant relationship between morality and practical rationality. One will be more well off if the world is a more moral place, and worse off if the world is a more immoral place. One cannot control whether the world is a morally good or bad place, but one does have control over *one’s part* in the world being a morally good or bad place. If the part one plays adds moral goodness to the world, then one adds to the realization of a central precondition for one’s own well-being. If the part one plays makes the world a more morally worse place overall, then one subtracts from the realisation of a precondition for one’s own well-being. So, although not all immoral actions may be *directly* irrational, there is an intriguing relationship between morality and practical rationality nonetheless. Perhaps the undermining, through one's actions, of a precondition on which one's wellbeing rests does amount to a kind of practical irrationality: though I will remain agnostic about this and leave this claim as 'perhaps.' What is clear, however, is that all what has been said in this thesis about morality does seem to vindicate many of the intuitions about morality that we have: that violating it is specially different from violating any other kinds of norms, that those who flout morality do, in more way than one, most likely harm themselves, but that, at the same time, they do something *more* serious than harming merely themselves: they undercut or interfere with the desires of *others* to be well off. All this explains perfectly well why moral requirements exert the phenomenological pressure that they do, and why our desire for morality to be propped up by rational support is somewhat vindicated.

3. Final Thoughts

What I aimed to provide in this thesis was an understanding of moral success theory that avoided the error theory threats which revolve around considerations about the relationship between morality and practical rationality. It is my hope that the arguments presented here will help to clarify many debates in metaethics, and will provide new directions for enquiry into the nature of moral truth. If this thesis has achieved its goals, it will have inspired a renewal or sustaining of confidence in moral success theory on the reader's part.

I hope also, that first-order moral debates will find the metaethical framework presented here helpful. Indeed the possibility that moral error theory might be true might render normative or applied ethics a mistaken pursuit. My account of the truthmakers of moral judgments, as I said in chapter 5, leaves a lot of room for others to add to. Working out which actions or moral intuitions would be favoured by the responses of our ideal observer is a process that I hope others will be interested in undertaking and debating, with the prospects of ethical progress on the horizon.

Bibliography:

Adams, R. 1999. *Finite and Infinite Goods*. Oxford University Press.

Arneson, R. 1999. 'Human Flourishing versus Desire Satisfaction,' *Social Philosophy and Policy* 16/1: 113-42.

Ayer, A. 1936. *Language, Truth and Logic*. Penguin Modern Classics.

Baier, A. 2010. 'Is Empathy all we Need?' *Abstracta* Special Issue V: 28-41.

Baier, K. 1995. *The Rational and the Moral Order: the Social Roots of Reason and Morality*. Open Court.

Baumeister, R. & Leary, M. 1995. 'The Need to Belong: Desire for Interpersona Attachments as a Fundamental Human Motivation,' *Psychological Bulletin* 177/3: 497-529.

Bentham, J. 2008 [1789]. 'An introduction to the Principles of Morals and Legislation,' 99-103 in Cahn, S. & Vitrano, C. (eds.), *Happiness: Classic and Contempory Readings in Philosophy*. Oxford University Press.

Björnsson, G. & Olinder, R. 'Internalists Beware – we Might all be Amoralists!' *Australasian Journal of Philosophy*, 91/1: 1-14.

Blackburn, S. 1993. *Essays in Quasi-Realism*. Oxford University Press.

Blackburn, S. 1998: *Ruling Passions*. Oxford University Press.

Blackburn, S. 1999: 'Is Objective Moral Justification Possible on a Quasi-realist Foundation?' *Inquiry* 42/2: 213-28

Blackburn, S. 2006a. 'The Frege-Geach Problem,' 349-59 in Fisher, A. & Kirchin, S. (eds.), *Arguing About Metaethics*. Routledge.

Blackburn, S. 2006b. 'Attitudes and Contents,' 369-85 in Fisher, A. & Kirchin, S. (eds.), *Arguing About Metaethics*. Routledge.

Bloomfield, P. 2003. ‘Is there Moral High Ground?’ *Southern Journal of Philosophy* 41/4: 511-26.

Boyd, R. 1988. ‘How to be a Moral Realist,’ 181-228 in Sayre-McCord, G. (ed.), *Essays on Moral Realism*. Cornell University Press.

Brandt, R. 1955. ‘The Definition of an “Ideal Observer” Theory in Ethics,’ *Philosophy and Phenomenological Research* 15/3: 407-13.

Brandt, R. 1979. *A Theory of the Good and the Right*. Prometheus Books.

Brink, D. 1984. ‘Moral Realism and the Sceptical Arguments from Disagreement and Queerness,’ *Australasian Journal of Philosophy* 62/2: 111-25.

Brink, D. 1989. *Moral Realism and the Foundations of Ethics*. Cambridge University Press.

Brink, D. 1997. ‘Kantian Rationalism: Inescapability, Authority, and Supremacy,’ 255-92 in Cullity, G. & Gaut, B. (eds.), *Ethics and Practical Reason*. Oxford Clarendon Press.

Burton, S. 2006 [1992]. ‘‘Thick’ Concepts Revised,’ 511-515 in Fisher, A. and Kirchin, S. (eds.), *Arguing About Metaethics*. Routledge.

Carson, T. 1981. *The Status of Morality*. D. Reidel Publishing Company.

Carson, T. 2000. *Value and the Good Life*. University of Notre Dame Press.

Cohon, R. 1986. ‘Are External Reasons Impossible?’ *Ethics* 96/3: 545-56

Copp, D. 1995. *Morality, Normativity & Society*. Oxford University Press.

Copp, D. 2001. ‘Realist-Expressivism: A Neglected Option for Moral Realism,’ *Social Philosophy and Policy* 18/2: 1-43.

Copp, D. 2004. ‘Moral Naturalism and Three Grades of Normativity,’ 7-46 in Schaber, P. (ed.), *Normativity and Naturalism*. Transaction Books & Gazelle Books.

Cowley, C. 2005. ‘A New Defence of Williams’s Reasons-Internalism,’ *Philosophical Investigations* 28/4: 346-68.

Cullity, G. 1997. 'Practical Theory,' 101-24 in Cullity, G. & Gaut, B. (eds.), *Ethics and Practical Reason*. Oxford Clarendon Press.

Cuneo, T. 2001. 'Are Moral Qualities Response-Dependent?' *Noûs* 35/4: 569-91.

Cuneo, T. 2007. *The Normative Web*. Oxford University Press.

Dahlsgaard, K., Peterson, C., & Seligman, M. 2005. 'Shared Virtue: The Convergence of Valued Human Strengths Across Culture and History,' *Review of General Psychology* 9/3: 203-13.

Dancy, J. & Hookway, C. 1986. 'Two Conceptions of Moral Realism,' *Proceedings of the Aristotelian Society, Supplementary Volumes* 60: 167-205.

Darwall, S. 1983. *Impartial Reason*. Cornell University Press.

Davis, W. 2008 [1981]. 'Pleasure and Happiness,' 163-172 in Cahn, S. & Vitrano, C. (eds.), *Happiness: Classic and Contemporary Readings in Philosophy*. Oxford University Press.

Davis, W. 2013. 'Implicature,' *The Stanford Encyclopedia of Philosophy* http://plato.stanford.edu/entries/implicature/, accessed 11/02/2013.

Demir, M., Özen, A., Bilyk, N., Tyrell, F. 2011. 'I Matter to My Friend, Therefore I am Happy: Friendship, Mattering, and Happiness,' *Journal of Happiness Studies* 12/6: 983-1005.

Diener, M. & Diener McGavran, M. 2008. 'What Makes People Happy?: A Developmental Approach to the literature on Family Relationships and Well-Being,' 347-75 in Eid, M. & Larsen, R. (eds.), *The Science of Subjective Well-Being*. The Guilford Press.

Driver, J. 2006. *Ethics: the Fundamentals*. Blackwell.

Driver, J. 2013. 'Moral Sense and Sentimentalism,' 358-76 in Crisp, R. (ed.), *The Oxford Handbook of the History of Ethics*. Oxford University Press.

Feldman, F. 2004. *Pleasure and the Good Life*. Oxford Clarendon Press.

Filonowicz, J. 2008. *Fellow Feeling and the Moral Life*. Cambridge University Press.

Finlay, S. 2009. ‘The Obscurity of Internal Reasons,’ *Philosopher’s Imprint* 9/7: 1-21

Finnis, J. 1980. *Natural Law and Natural Rights*. Oxford Clarendon Press.

Finnis, J. 1983. *Fundamentals of Ethics*. Oxford Clarendon Press.

Finnis, J. 2011. ‘Abortion and Health Care Ethics,’ 17-24 in Kuhse, H. & Singer, P. *Bioethics: An Anthology*. Blackwell.

Firth, R. 1952. ‘Ethical Absolutism and the Ideal Observer,’ *Philosophy and Phenomenological Research* 12/3: 317-45.

Foot, P. 1972. ‘Morality as a System of Hypothetical Imperatives,’ *Philosophical Review* 81/3: 305-16.

Gallup, G. 1984. ‘Religion in America,’ *The Gallup Report*, 222.

Garner, R. 2006 [1990]. ‘On the Genuine Queerness of Moral Properties and Facts,’ 96-106 in Fisher, A. & Kirchin, S. (eds.), *Arguing About Metaethics*. Routledge.

Geach, P. 1960. ‘Ascriptivism,’ *Philosophical Review* 69/2: 221-5.

Geach, P. 1965. ‘Assertion,’ *Philosophical Review* 74/4: 449-465.

Gewirth, A. 1978. *Reason and Morality*. University of Chicago Press.

Gibbard, A. 1990. *Wise Choices, Apt Feelings*. Harvard University Press.

Haidt, J. & Joseph, C. 2004. ‘Intuitive Ethics: how Innately Prepared Intuitions Generate Culturally Variable Virtues,’ *Dædalus* 133/4: 55-66.

Harman, G. 1975. ‘Moral Relativism Defended,’ *Philosophical Review* 84/1: 3-22.

Harman, G. 1977. *The Nature of Morality*. Oxford University Press.

Harman, G. 1996. ‘Moral Relativism,’ 1-64 in Harman, G. & Jarvis Thomson, J. (eds.), *Moral Relativism and Moral Objectivity*. Blackwell.

Hare, R. 1964. *The Language of Morals*. Oxford University Press.

Haybron, D. 2003. 'What do we want from a Theory of Happiness?' *Metaphilosophy* 34/3: 305-29.

Heathwood, C. 2006. 'Desire Satisfaction and Hedonism,' *Philosophical Studies* 128/3: 539-63.

Horgan, T. & Timmons, M. 2006a. 'Cognitivist Expressivism,' 255-298 in Horgan, T. & Timmons, M. (eds.), *Metaethics After Moore*. Oxford University Press.

Horgan, T. & Timmons, M. 2006b. 'Expressivism, Yes! Relativism, No!' 73-98 in Shafer-Landau (ed.), *Oxford Studies in Metaethics Volume 1*. Oxford University Press.

Hume, D. 1969 [1740]. *A Treatise of Human Nature*. Penguin Books.

Hurka, T. 1993. *Perfectionism.* Oxford University Press.

Jackson, F. 1998. *From Metaphysics to Ethics*: *A Defence of Conceptual Analysis*. Oxford Clarendon Press.

Jackson, F. & Pettit, P. 1998. 'A Problem for Expressivism,' *Analysis* 58/4: 239-51.

Joyce, R. 2001. *The Myth of Morality*. Cambridge University Press.

Joyce, R. 2006. *The Evolution of Morality*. MIT Press.

Joyce, R. 2011. 'The Accidental Error Theorist,' 153-180 in Shafer-Landau, R. (ed.), *Oxford Studies in Metaethics volume 6*. Oxford University Press.

Kamp Dush, C. & Amato, P. 2005. 'Consequences of Relationship Status and Quality for Subjective Well-being,' *Journal of Social and Personal Relationships* 22/5: 607-27

Kant, I. 1997 [1785]. 'Groundwork of the Metaphysics of Morals' 1-66 in Gregor, M. (ed.), *Cambridge Texts in the History of Philosophy: Groundwork of the Metaphysics of Morals*. Cambridge University Press.

Keller, S. 2004. 'Welfare and the Achievement of Goals,' *Philosophical Studies* 121/1: 27-41.

Keller, S. 2009. 'Welfarism,' *Philosophy Compass* 4/1: 82-95.

Keller, S. 2013. *Partiality*. Princeton University Press.

Kirchin, S. 2010. 'A Tension in the Moral Error Theory,' 167-82 in Joyce, R. & Kirchin, S. (eds.), *A World Without Values*. Springer.

Korsgaard, C. 1986. 'Skepticism about Practical Reason,' *Journal of Philosophy* 83/1: 5-25.

Korsgaard, C. 1996. *The Sources of Normativity*. Cambridge University Press.

Lewis, D. 1989. 'Dispositional Theories of Value: II – David Lewis,' *Proceedings of the Aristotelian Society, Supplementary Volumes*, 63: 113-37.

Mackie, J. 1977. *Ethics: Inventing Right and Wrong*. Penguin Books.

Marquis, D. 2011. 'Why Abortion is Immoral,' 51-62 in Kuhse, H. & Singer, P. (eds.), *Bioethics: An Anthology*. Blackwell.

Mason, E. 2007. 'The Nature of Pleasure: A Critique of Feldman,' *Utilitas* 19/3: 379-387.

Matas, L., Arend, R., & Stroufe, L. 1978. 'Continuity in Adaptation: Quality of attachment and later competence.' *Child Development*, 49: 547-56.

McDowell, J. 1998. *Mind, Value & Reality*. Harvard University Press.

Meyers, D. 2008. 'Religion and Human Flourishing,' 323-43 in Eid M., & Larsen, R. (eds.), *The Science of Subjective Well-Being*. The Guilford Press.

Miller, A. 2003. *An Introduction to Contemporary Metaethics*. Polity Press.

Millgram, E. 1996. 'Williams' Argument Against External Reasons,' *Noûs* 30:2 30/2: 197-220.

Mongrain, M., Chin, J., & Shapira, L. 2011. 'Practicing Compassion Increases Happiness and Self-Esteem,' *Journal of Happiness Studies* 12/6: 963-81

Moore, G. 2006 [1903]. 'The Open Question Argument: the Subject Matter of Ethics,' 31-46 in Fisher, A. & Kirchin, S. (eds.), *Arguing About Metaethics*. Routledge.

Nagel, T. 1970. *The Possibility of Altruism*. Princeton University Press.

Nichols, S. 2004. *Sentimental Rules*. Oxford University Press.

Nozick, R. 2008 [1974]. 'The Experience Machine,' 236-7 in Cahn, S. & Vitrano, C. (eds.) *Happiness: Classic and Contemporary Readings in Philosophy*. Oxford University Press.

Parfit, D. 1984. *Reasons and Persons*. Oxford Clarendon Press.

Parfit, D. 1997. 'Reasons and Motivation: I – Derek Parfit,' *Proceedings of the Aristotelian Society, Supplementary Volumes* 71: 99-130.

Parfit, D. 2010. *On What Matters vol. 1*. Oxford University Press.

Perlmutter, M. 1999. 'Desert and Capital Punishment,' 122-9 in Arthur, J. (ed.), *Morality and Moral Controversies: Fifth Edition.* Prentice Hall.

Pettit, P. 1991. 'Realism and Response-Dependence,' *Mind* 100/4: 587-626.

Pigden, C. 1989. 'Logic and the Autonomy of Ethics,' *Australasian Journal of Philosophy* 67/2: 127-51

Prichard, H. 1912. 'Does Moral Philosophy Rest on a Mistake?' *Mind* 21/1: 21-37.

Prinz, J. 2007. *The Emotional Construction of Morals*. Oxford University Press.

Putnam, H. 2002. *The Collapse of the Fact/Value Dichotomy and other Essays*. Harvard University Press.

Putnam, H. 2004. *Ethics without Ontology*. Harvard University Press.

Rachels, J. 1996 [1986]. 'The Challenge of Cultural Relativism,' 488-94 in Feinberg, J (ed.), *Reason and Responsibility: readings in some basic problems of Philosophy, 9th Edition.* Wadsworth Publishing Company.

Railton, P. 1986. 'Facts and Values,' *Philosophical Topics* 14/2: 5-31.

Railton, P. 1992. 'Some Questions about the Justification of Morality,' *Philosophical Perspectives* 6: 27-53.

Railton, P. 2006 [1986]. 'Moral Realism,' 145-78 in Fisher, A. & Kirchin, S. (eds.), *Arguing About Metaethics*. Routledge.

Rawls, J. 1971. *A Theory of Justice*. Harvard University Press.

Rosati, C. 1995. 'Persons, Perspectives, and Full Information Accounts of the Good,' *Ethics* 105/2: 296-325.

Sayre-McCord, G. 1986. 'The Many Moral Realisms,' *Southern Journal of Philosophy* suppl. Vol. 24: 1-22.

Scanlon, T. 1998. *What we Owe to Each Other*. Belknap Press, Harvard University Press.

Schroeder, M. 2007. *Slaves of the Passions*. Oxford University Press.

Schroeder, M. 2010. *Noncognitivism in Ethics*. Routledge.

Shafer-Landau, R. 2003. *Moral Realism: A Defence*. Oxford University Press.

Shafer-Landau, R. 2005. 'Error Theory and the Possibility of Normative Ethics,' *Philosophical Issues* 15: 107-19.

Singer, P. 1993. *Practical Ethics: Second Edition*. Cambridge University Press.

Sinnot-Armstrong, W. 2006. *Moral Skepticisms*. Oxford University Press.

Slote, M. 1985. *Common-Sense Morality and Consequentialism*. Routledge & Kegan Paul.

Slote, M. 2010. *Moral Sentimentalism*. Oxford University Press.

Smith, M. 1989. 'Dispositional Theories of Value: I –Michael Smith,' *Proceedings of the Aristotelian Society, Supplementary Volumes*, 63: 89-111

Smith, M. 1994. *The Moral Problem*. Blackwell Publishers.

Smith, M. 1995. 'Internal Reasons,' *Philosophy and Phenomenological Research* 55/1:109-31.

Sobel, D. 1999. 'Do the Desires of Rational Agents Converge?' *Analysis* 59/3: 137-47.

Sobel, D. 2002. 'Varieties of Hedonism,' *Journal of Social Philosophy* 33/2: 240-56.

Stevenson, C. 1937. 'The Emotive Meaning of Ethical Terms,' *Mind* 46/1: 14-31.

Stoljar, D. 1993. 'Emotivism and Truth Conditions,' *Philosophical Studies* 70/1: 81-101.

Sturgeon, N. 1988 [1984]. 'Moral Explanations,' 229-255 in Sayre-McCord, G. (ed.), *Essays on Moral Realism*. Cornell University Press.

Sumner, L. 1996. *Welfare Happiness & Ethics*. Oxford University Press.

Taliaferro, C. 1988. 'Relativising the Ideal Observer Theory,' *Philosophy and Phenomenological Research* 49/1: 123-38.

Wallace, J. 1990. 'How to Argue About Practical Reason,' *Mind* 99/3: 355-85.

Warnock, G. 1971. *The Object of Morality*. Methuen.

Wedgwood, R. 1999. 'The Price of Non-Reductive Realism,' *Ethical Theory and Moral Practice* 2/3: 199-215.

Wedgwood, R.2002. 'Practical Reasoning as Figuring Out What is Best: Against Constructivism,' *Topoi* 21/1: 139-52.

Wiggins, D. 1987. *Needs, Values, Truth*. Oxford Clarendon Press.

Williams, B. 1979. 'Internal and External Reasons,' 17-28 in Harrison, R. (ed.), *Rational Action: Studies in Philosophy and Social Science*. Cambridge University Press.

Williams, B. 1981. *Moral Luck*. Cambridge University Press.

Williams, B. 1995 [1989]. 'Internal Reasons and the Obscurity of Blame,' in Williams, B., *Making Sense of Humanity*. Cambridge University Press.

Wolf, S. 1982. 'Moral Saints,' *Journal of Philosophy*, 79/8: 419-39

Wong, D. 1991. 'Relativism' 442-450 in Singer, P. (ed.), *A Companion to Ethics*. Blackwell Publishers.

Wood, E. Forthcoming. 'Björnsson and Olinder on Motivational Internalism,' *Australasian Journal of Philosophy*, forthcoming.

Wright, C. 1988. ‘Moral Values, Projection and Secondary Qualities,’ *Proceedings of the Aristotelian Society, Supplementary Volumes* 62: 1-26.

Wright, C. 1992. *Truth & Objectivity*. Harvard University Press.