MAKING SENSE OF REASONS:

PROSPECTS FOR AN INTERPRETIVIST ACCOUNT OF PRACTICAL REASONS

DAVID BEESLEY, BA, PGDip, MA

Thesis submitted to the University of Nottingham

for the degree of Doctor of Philosophy

2011

To Carly

Abstract

This thesis investigates the prospects for an interpretivist account of practical reasons. The proposed account identifies practical reasons with sets of propositional attitudes from which certain actions follow, given the constraints of interpretable functioning. Following Davidson, these constraints are taken to be enumerated by formal decision theory and formal semantics. Thus the account of practical reasons is framed in terms of what rationally follows from agents' beliefs and desires. The hope is that an account of practical reasons of this kind can explain the existence of practical reasons without invoking irreducible normative properties or relations. This outcome depends upon the availability of a theory of (radical) interpretation which is free from prior normative commitments. It is argued that a non-normative reading of Davidson's theory of radical interpretation is available, such that the account of practical reasons can meet this requirement. Although the proposed account of practical reasons does not admit of the possibility of categorical reasons for action, the ensuing objection that it fails to allow for the possibility of moral reasons for action is resisted. It is suggested that a plausible account on which moral reasons are hypothetical in kind can be provided. In particular, an account of moral reasons which is framed in terms of the motivations associated with a capacity for empathic affect is advanced. More generally, the aspiration of the thesis is to provide an account of practical reasons framed in terms of the requirements of interpretable functioning which will be regarded as an interesting and credible naturalistic option.

Acknowledgements

First, I would like to thank my supervisors, Christopher Woodard and Stephen Barker, for their invaluable help and encouragement in writing this thesis. I am extremely grateful for their time and input into my work and have greatly enjoyed the discussions that I have had with both of them over the past few years. Many thanks to you both.

I would also like to thank the many other people who have offered me help, guidance and support, or who have discussed my ideas with me. These include: James Andow, Robert Black, Gregory Currie, Andrew Fisher, Alex Gregory, Matthew Kennedy, Anna Ichino, David Ingram, Greg Mason, Charlotte Matheson, Ben McGorrigan, Stefano Predelli, Greg Scorzo, Neil Sinclair, Ben Smart and Matthew Tugby. I apologise to anyone who I have forgotten to mention. I would like to thank the Philosophy Department at the University of Nottingham, and particularly the departmental administrators, Ann Currie, Liz Rawding and Jane Pytches-Walker. I would like to thank Paul Noordhof, whose help and advice with my application for an AHRC doctoral award was invaluable. And I would like to thank David Bain, whose good advice has remained with me since I first undertook to study philosophy.

I would especially like to thank the AHRC for the doctoral award which has funded this project.

I would also like to thank Kelly Heuer, for kindly allowing me to cite her paper 'Hypotheticalism and the Objectivity of Morality' in chapter 5.

I would like to thank my parents for always encouraging me to pursue my goals and for all of their support along the way. I would also like to thank Viv and Eric Bignell for their most generous support during my university education. Finally, I would like to thank Carly Collingwood for all her love and support. This thesis is dedicated to Carly, with love.

Contents

Introduction

This thesis is about practical reasons. It is an attempt to provide a plausible philosophical account of what practical reasons are. The proposed account is interpretivist in kind. That is, it seeks to explain practical reasons in terms of what is necessary for agents to be interpretable. The account is informed by two overarching concerns: (i) maintaining that practical reasons exist; (ii) conforming with a broadly naturalistic worldview. There is a well-known tension between these two concerns. This arises because practical reasons are normative entities; accounting for normative entities within a naturalistic framework has proven to be problematic.

As indicated above, I offer an interpretivist account of what practical reasons are. I develop this approach by claiming that practical reasons are sets of attitudes from which certain actions follow, given the constraints of interpretable mental functioning. This approach is inspired by Davidson's philosophy of language, mind and action. On a Davidsonian approach, an agent is interpretable if it is possible to attribute a set of propositional attitudes to her (and to ascribe a set of meanings to her utterances) in light of her behaviour; the sets of attitudes and meanings ascribed must conform to certain formally specifiable constraints on interpretation, where these are enumerated by formalised decision theory and formal semantics (Davidson, 1980; 1995). Davidson takes these theories to provide a set of systematic constraints on agents' propositional attitudes, and on the meanings of the words and sentences that they utter. Conforming to these constraints is necessary if propositional attitudes and meanings are to be reliably attributable to agents in light of what they say and do.

My proposal is to treat practical reasons as sets of attitudes from which certain actions follow, given the constraints of interpretable functioning. Thus, in crude terms, to have a reason to perform some action is for it to make sense that one does so, given one's beliefs

and desires. This proposal allows the existence of practical reasons to piggyback on the existence of systematic constraints on how agents can function while remaining interpretable. Assuming, of course, that agents *are* interpretable, this seems to be an appealing move. Explaining practical reasons in terms of the existence of systematic constraints on interpretable functioning allows them to be (rightly) located as a central feature of agents' practical lives. In fact, it affords an explanation of why reasons have this central role in our practical reality (i.e. that acting for a reason is a necessary feature of being interpretable) while avoiding the need to treat reasons in primitive terms.

The thesis comprises five chapters. Chapter 1 introduces the problem of accounting for practical reasons in naturalistic terms, explains what is required of any given naturalistic account, and discusses some existing naturalistic approaches. Chapters 2 and 3 develop the proposed account of practical reasons. In chapter 2 I explain the proposed interpretivist account of what practical reasons are. This account is framed in terms of the rationality of having certain intentions to act and, derivatively, of performing certain actions, given ones beliefs and desires. I also provide an account of practical reasons' normative force in terms of agents having a constitutive commitment to function in accordance with the constraints of interpretable functioning. In chapter 3 I discuss the rational principles which underpin the account of practical reasons given in chapter 2. In particular, I discuss Davidson's Principle of Continence as well as a decision-theoretic Principle of Maximisation. I claim that the former takes psychological priority, but that neither principle can be seen to take rational priority. I also argue that both principles are suitable for interpretive purposes.

Having set out my proposed account of practical reasons, chapters 4 and 5 respond to two objections to the proposed view. Chapter 4 discusses two objections to the effect that interpretation is irreducibly normative, and thus cannot be invoked in providing a naturalistic explanation of practical reasons *qua* normative entities. These are the objections that meaning and preference are, respectively, irreducibly normative. I argue that it is at least possible for a Davidsonian to give non-normative accounts of meaning ascription and preference attribution, such that the proposed account of practical reasons remains a plausible naturalistic view.

In chapter 5 I discuss the objection that my account of practical reasons does not admit of the possibility of moral reasons for action. This objection arises because moral reasons have traditionally been taken to be independent of the specific contents of agents' desires (i.e. to be categorical reasons for action). I accept that my view of practical reasons cannot accommodate the existence of categorical reasons for action. However, I argue that a plausible hypotheticalist account of moral reasons, in terms of certain desires relating to our empathic responses to others, is available, such that the objection from moral reasons can be avoided.

Before proceeding with the rest of the thesis, it is important to note that at no point do I claim that interpretivism about practical reasons is true, or even that it is the best or most plausible naturalistic view of practical reasons available. My purpose in this thesis is to consider what an interpretivist account of practical reasons might look like, and to argue that such a view is a plausible, naturalistic contender in explaining practical reasons. I hope to be able to persuade the reader of this claim.

Chapter 1: The Problem of Practical Reasons

> *Naturalism exacts a high price from us: it robs us not only of the objective normative authority of moral imperatives, but of* any *imperative, even imperatives of instrumental reason. Jean Hampton, 'On Instrumental Rationality'.*

1. Introduction

Practical reasons, as commonly conceived, are normative entities. In Scanlon's terms, they are considerations which count in favour of performing an action (Scanlon, 1998: ch.1). That it is raining favours my use of an umbrella; that I want a coffee favours putting the kettle on; that donating money to charity will reduce suffering favours making a donation. Such reasons may be defeated (by other, stronger reasons, for example), but it is nevertheless the normative status of practical reasons which is their hallmark, or so it seems. A practical reason which does not count in favour of some action does not seem to be a practical reason at all.

Perhaps one might deny this claim by suggesting that practical reasons are not normative considerations at all, but merely motivating ones. However, to be a consideration (rather than a mere cause) seems implicitly to involve a normative element (Dancy, 2000: 97; 169). Further, eschewing all normativity, by picking out mere causes of behaviour, leaves nothing to distinguish reasons from other kinds of behavioural causes (such as reflexes, ticks, impulses and the like). Thus it seems that a normative element is required to distinguish practical reasons from other kinds of factors which influence our behaviour.

This causes problems for the naturalist. In the first instance, naturalism might be taken as a metaphysical thesis to the effect that whatever exists must be 'composed of entities that our best scientific theories require' (Prinz, 2007: 2). However, this kind of metaphysical thesis is relatively liberal, being compatible with the existence of normative relations even if

such relations are treated as irreducible. Thus one might claim that the normative supervenes on the natural, and that it is certain natural objects/states of affairs which realise (or instantiate) normative properties, while denying that normative properties are reducible to some natural property kind (Sturgeon, 1988; Wedgwood, 2007).

One objection to these kinds of non-reductive view is that they lack parsimony. Why propose that it is the instantiation of certain normative relations which explains our normative judgements (and associated conduct) when this data can be equally explained in terms of human affective responses to certain situations (Blackburn, 1984; Mackie, 1977). Further, one might argue that explanations of ethical conduct (in the broadest sense of ethical) which invoke irreducible normative relations are crowded out by evolutionary explanations of the cognitive and affective capacities involved in normative judgements (Street, 2006).

However, these are not strictly objections from naturalism itself, even though the explanatory overcrowding objection involves reference to a highly plausible naturalistic theory of human behaviour (evolutionary psychology). If responses to the parsimony and explanatory overcrowding objections are available then (other objections notwithstanding) non-reductivism about the normative, and metaphysical naturalism may both be true. Nevertheless, naturalism is often adopted as more than a metaphysical claim about the composition of reality. Naturalism is commonly adopted as an explanatory thesis too. For instance, one might claim that there must be a 'systematic correspondence' between non-scientific explanations and scientific ones (Prinz, 2007: 2-3). If we are to explain agents' behaviour in terms of their apprehension of normative properties then, the explanatory naturalist might argue, there had better be some respectable explanation of how this kind of story relates to the causal theories of worldly phenomena provided by the empirical sciences.

It is at this point that non-reductivism about the normative faces a challenge. Even if normative relations might be taken to supervene on natural ones, the explanatory naturalist will not be content unless some principled explanation of the presence of this

supervenience relation can be given. What explains the supervenience of normative states of affairs on causal processes at the fundamental physical level?

As it stands this is a challenge to the non-reductivist rather than a decisive objection. However, the purpose of this thesis is not to investigate how the non-reductivist might meet this challenge. Rather, I propose to accept a presumption against non-reductive views of normative properties based on concerns about these views' naturalistic credibility, as well as their credibility given worries about a lack of parsimony and about explanatory overcrowding.

If we suppose that some form of explanatory naturalism is correct, *and* that this is at odds with non-reductive views of normative relations, then it might appear that normative relations cannot be taken to exist at all, and that practical reasons must fall by the wayside with them. Given that normative relations, such as favouring, appear to be directive in a way which cannot be accounted for in merely causal terms, it seems that accepting the existence of such relations must involve accepting the kind of irreducible normative properties which appear problematic from a naturalistic perspective. The naturalist has two options in light of this worry (short of giving up naturalism): (i) offer an account of practical reasons which does not invoke irreducible normative relations; (ii) deny that there are any practical reasons whatsoever. Of these, I take option (i) to be preferable, as it is more compatible with our ordinary views about practical reasons (i.e. that such things exist, and that our actions can be guided by them).

The purpose of this thesis is to investigate one way in which a naturalistic account of practical reasons might be given. The purpose of this chapter is to set the scene for that investigation. Thus, in this chapter, I attempt to clarify some of the central issues which arise in attempting to give a naturalistic account of practical reasons, and to discuss some existing naturalistic approaches. The aim of the chapter is to illustrate the different ways in which a naturalistic account of practical reasons might be developed, and to show what is at stake for any given naturalistic account.

2. Normativity

In section 1, I suggested that practical reasons are commonly conceived of as normative entities. I elucidated this notion by giving some examples in which a reason favours an action. These included the rain favouring my use of an umbrella, and my desire for a coffee favouring that I put the kettle on. Before proceeding with the discussion of practical reasons, it will be useful to make some more explicit remarks about normativity here.

Normativity has been a topic of much recent debate (see Finlay, 2010, for a discussion of recent work on normativity). Much of this debate has focussed on the nature of normative force—the oomph behind normative assignations, so to speak. For example, there have been many attempts to demystify normativity by providing a clear explanation of the nature and existence of normative force. One popular recent approach here has been to explain normativity in terms of the constitutive requirements of agency, such that the normative force of reasons can be explained in terms of there being certain requirements that agents must meet if they are to count as acting at all.[1] This approach can be seen as a response to the arguably unsatisfying, primitivist conceptions of normativity offered by, among others, Parfit (2006; forthcoming) and Scanlon (1998: ch.1). On these views, normativity is identified in terms of primitive normative relations, such as that of 'favouring', where no further explanation of the normative force transmitted by these relations is offered.

Normative force is an important feature of normativity. However, the notion of normativity is not exhausted by the notion of normative force. A fuller conception of normativity is provided by the idea of something's existing within the domain of norms or standards. Norms/standards provide constraints on how things must be. These constraints are not physical but evaluative; they are constraints on how things must be in order to be meet certain evaluative criteria, such as goodness, rightness or correctness. For instance, building

[1] This kind of approach is adopted by Goldman (2010: 66-82 & 181-5), Korsgaard (2008, ch.3; 2009, esp. chs. 2 & 4-7) and Velleman (2000: chs. 1 and 8).

regulations are normative in the sense that they set out certain standards which any newly erected building must meet. Any newly constructed building which does not meet these standards is deficient.

One way of understanding the idea of something's being subject to norms or standards derives from Timothy Schroeder (2003). Schroeder suggests that the application of a norm involves two features: (a) the presence of a *categorisation scheme*, by which certain actions, events, or objects can be classified as either correct or incorrect, right or wrong; and (b) the presence of a normative *force-maker*, by which the normative significance of falling under one category or another is established. Schroeder lists functions bestowed on objects by our intentions, social pressure and natural selection as possible sources of normative force. Other sources of normative force accompany the different views of practical reasons discussed in this chapter (for example, desire in the case of Mark Schroeder's neo-Humeanism; natural functioning in the case of neo-Aristotelianism).

For immediate purposes, what is important is that normativity involves the existence of certain categories which an action, object, or event can fall under, where falling under one category rather than another is normatively significant, given the presence of some normative force-maker which attaches normative significance to these categories.

I take the idea of a categorisation scheme to be relatively clear. The classification of actions, objects and events (among other things) is something that we routinely do. However, the ideas of normative force, and of normative significance, are more obscure. I am not sure that I can offer any clearer rendering of these ideas. This is not because the use of these notions is in any way confused or inappropriate. Rather, it is because they appear to be conceptually primitive. Normative force seems to be a notion which cannot be unpacked in terms of anything more basic. One might simply say that our concept of normative force is of that which backs up any (legitimate) normative assignation—that which makes such assignations significant to us. Or, in slightly different terms, normative force is that which solicits a concern to be, or to do, certain things. For the purposes of this thesis, I assume that the notion of normative force is conceptually primitive. I do not take this to prejudice

attempts to provide a conceptual reduction of it. Rather, I am simply unaware of any candidates for a conceptual reduction of normative force which promise to be successful.

Now, the claim that performing certain actions can be normatively significant, in some way, is obviously somewhat troublesome. It is this claim which seems, on the surface, to be incompatible with naturalism. For it seems impossible to account, in any way, for its being normatively significant that we do certain things without invoking some kind of non-natural phenomena. It is this problem which leads to the worry that, on a naturalistic worldview, there are no reasons for action.

Nevertheless, this worry is indecisive. It remains open for the naturalist to attempt to account for practical reasons, including their normative force, in naturalistic terms. The purpose of this thesis is to investigate one way in which a naturalistic account of practical reasons might be given, having established what is at stake for such a view in this chapter.[2]

[2] There are, of course, arguments which are taken to undermine all naturalistic accounts of normative phenomena. Moore's open question argument against a naturalistic account of goodness is the most famous example (Moore, 1903: 15-16). Roughly, the argument is that if 'good' just means (say) 'being that which we desire to desire', then the question of whether it is good that we desire to desire outcome should not appear to be open. But asking whether it is good that I desire to desire something does seem to be an open question (which Moore assumes it would not if the definition succeeded). So it cannot be that goodness is identical to that which we desire to desire after all. Moore takes this result to generalise to any other supposed naturalistic definition of goodness, while the argument can be seen as applying more broadly to all naturalistic accounts of normative entities.

Popular responses to the open question argument include: claiming that identity relations between normative and non-normative concepts need not be obvious, such that a conceptual question can appear to be open even where two concepts are identical (Smith, 1994: 37-8); accepting that normative *concepts* cannot be analysed in non-normative terms (e.g. that goodness cannot be defined as that which we desire to desire), while denying that this prejudices attempts to identify normative *properties* in terms of non-normative ones (Brink, 2001). I do not discuss the open question argument or other such anti-naturalistic arguments in this thesis, the purpose of which is to explore a particular naturalistic option rather than to give a general defence of the naturalistic approach.

3. Hypothetical Reasons

Until fairly recently it was assumed that worries over the incompatibility of naturalism with the existence of practical reasons only applied to certain kinds of reasons. Thus Foot, Mackie and others claimed that there are no categorical reasons for action, only hypothetical ones (Foot, 1972; Mackie, 1977). A categorical reason is a reason which applies to an agent regardless of the contents of her desires. A hypothetical reason is a reason which is contingent on the contents of her desires.

Based on her failure to find a non-linguistic basis for categorical imperatives, Foot suggested that 'it is uncertain whether the doctrine of the categorical imperative even makes sense' (Foot, 1972: 312). For Foot (the early Foot, at least), categorical imperatives are just hypothetical imperatives dressed up in categorical rhetoric. This is the only sense that she can make of the notion that we just *are* required to do, for instance, what is moral. Behaving morally, like behaving politely, is 'the done thing' but there is no intelligible sense, for Foot, in which it is *to-be-done*. It is only if we care about moral ends (as Foot supposes that we do) that we have reasons to do what is morally right. Roughly, 'if you want to promote other people's wellbeing, behave morally' is the kind of hypothetical imperative which generates moral reasons, on Foot's account. So, although Foot suspects that the notion of a categorical imperative is unintelligible, the notion of a hypothetical imperative, and its associated reason-giving force, is taken to be unproblematic.

Mackie also attacked the existence of categorical imperatives. He suggested that the existence of any objectively and categorically prescriptive entities to issue in such imperatives would be unacceptably 'queer' (Mackie, 1977: 38-42). As such, Mackie dismissed any belief in moral truth as in radical error. Nevertheless, Mackie did not question the existence of hypothetical imperatives. His target was the categorical prescriptions of morality, which he claimed could not be objective due to the threat of metaphysical and epistemological queerness. The objectivity of hypothetical imperatives, such as those of instrumental reason, was not taken to be similarly queer.

Recently, attention has been drawn to Foot's and Mackie's oversight. The issue, it has been pointed out, is not with categoricity but with normativity. It is the suggestion that performing some action is imperative (whether categorically *or* hypothetically so) which is problematic as, in either case, a normative relation is posited which stands in need of explanation. Thus Hampton writes that 'the contingency of the directives of a hypothetical imperative on a certain desire does not, by itself, explain why we *ought* to follow the directive' (Hampton, 1996: 93). That an imperative (or, indeed, a reason) would not exist without a certain desire does not explain why that imperative or reason has normative force.

Seen in this light, it seems somewhat strange to think that desiring could have anything to do with the existence of normative force. Desires are psychological states which feature in the explanation of actions. Roughly, they select certain ends that an agent will attempt to realise, other things being equal; desires are psychologically directive. But the selection of an end by a desire does not, in itself, seem to involve that end's having any kind of normative significance. That desires are psychologically directive does not entail that, where some desire is instantiated, there is any normative force which attaches to the pursuit of the desired end. To assume the contrary is to commit an is/ought fallacy: to assume that having an end entails that one ought to pursue it.

However, things are more complicated in that there is a principle of instrumental reason which, one might suppose, plays a crucial role in the move from desiring an end to its being normatively significant that one takes the means to that end. Roughly, the instrumental principle directs agents to take the (necessary) means to their preferred ends.[3] This principle is not entailed by the mere fact of desiring. However, assuming that it is a genuine principle of practical rationality, and that rationality is normative, wherever an agent desires

[3] This is a rough statement of the instrumental principle. The exact formulation of this principle is somewhat controversial, with philosophers disagreeing about what it specifically requires (see Broome, 1999; Dreier, 1997). These finer details are unimportant for present purposes.

some end, pursuing the means to that end will be normatively significant. Perhaps, then, it is this principle of instrumental rationality which generates the normative force of hypothetical imperatives and hypothetical reasons.

An assumption of this kind might explain some philosophers' tendency to treat hypothetical imperatives and reasons as benign. Nevertheless, adverting to the instrumental principle as a source of normativity merely relocates the problem, shifting it from the normative status of desiring to the normative status of instrumental rationality. As Korsgaard points out, 'philosophers have, for the most part, been silent on the question of the normative foundation of this requirement [the instrumental principle]' (Korsgaard, 1997: 215). Korsgaard is rightly concerned to account for its normative foundation. It cannot be taken for granted, as to do so would be to simply assume that instrumental rationality plays the required role in underpinning the normative force of hypothetical imperatives and reasons.

In fact, things are worse for those who take hypothetical reasons and imperatives to be unproblematically derivable from the instrumental principle. For this principle to even apply, it must be treated as a categorical principle of rationality. Thus, according to Dreier:

> M/E [the instrumental principle] has a kind of ground level normative status. I think it counts as a categorical imperative, too. Of course, the particular reasons that M/E generates are all hypothetical reasons. But M/E itself is not hypothetical. Its demands must be met by you, in so far as you are rational, no matter what desires you happen to have (Dreier, 1997: 96).

The idea here is that the instrumental principle provides the normative force behind hypothetical imperatives and reasons only because it is itself categorically prescriptive—only because it prescribes to anyone, regardless of her desires, that she takes the (necessary) means to her preferred ends. Without such categoricity, it is difficult to even make sense of what the instrumental principle might be (if you *want* to take the means to your ends, then take the means to your ends?!).

So, avoiding worries about practical reasons' normativity by retreating to a hypothetical account of practical reasons does not work. Hypothetical reasons either depend upon the categorically prescriptive instrumental principle for their normative force, or else this force must be entailed, somehow, by the mere fact of desiring. The first option leads straight back to the categorical prescriptions which we were trying to avoid. The second option involves committing an is/ought fallacy, as having an end in no way entails that one ought to pursue it. Given that the normative force of hypothetical reasons is as problematic as that of categorical reasons, an account of practical reasons of any kind (hypothetical or categorical) must be able to provide some explanation of their normative force.

4. Practical Reasons and Mind-Independence

So far I have done the following: introduced the problem of accounting for practical reasons in naturalistic terms; explained that normativity involves the application of a categorisation scheme which is normatively significant in some sense; explained why a retreat to hypothetical reasons does not avoid the problem of having to explain the normativity of practical reasons, given that the normativity of hypothetical reasons stands in need of as much explanation as that of categorical reasons.

In this section I discuss the issue of mind-independence. Specifically, I distinguish between two issues of mind-independence. The first is whether practical reasons apply to agents independently of the contents of their existing desires or not (whether practical reasons are categorical or hypothetical). The second is whether the normative force of practical reasons is independent of agents' existing attitudes or not. Plausibly, it is the conflation of these two issues which has lead some philosophers to treat hypothetical reasons as ontologically benign (by assuming that the normative force of hypothetical reasons is explained by their connection to agents' desires). However, once these issues are clearly separated, it can be seen that an explanation of the normative force of practical reasons (whether categorical or hypothetical) must be given rather than assumed.

Having distinguished between the two varieties of mind-independence, I discuss how they combine and briefly canvass some existing views which fall under each option.
A practical reason is categorical if that reason applies to an agent independently of the contents of her existing desires. For example, if Judy has a reason to tell the truth about having murdered Punch, regardless of what she desires, then this reason is categorical. A practical reason is hypothetical if that reason is contingent on the contents of the desires of the agent to whom that reason applies. For example, if Laurel has a reason to trip Hardy because he desires to see him fall over then this reason is hypothetical.

The view that practical reasons are (at least sometimes) categorical is typically associated with a Kantian view of practical rationality. According to this view, what agents have a reason to do is determined by the categorical imperative: (roughly) perform only those actions for which you can will that your maxim is a universal law. On Kantian views, at least some such maxims are taken to be desire-independent (moral maxims against lying, for example), such that the reasons that agents have are at least sometimes categorical. Categorical reasons for action can also exist on other views, such as consequentialist moral views, according to which agents' reasons to promote the good are often independent of their having any desire to do so.

The view that practical reasons are always hypothetical is typically associated with a Humean view of practical rationality, according to which 'reason is, and ought only to be the slave of the passions' (Hume, 1969: 462). Some maintain that Hume himself was a sceptic about practical rationality (Hampton, 1998: 142-9; Millgram, 1995). Nevertheless, the view that all practical reasons are contingent on the contents of agents' desires is generally considered to be Humean in spirit, even if it was not Hume's own view.

Given the explanation of normativity offered in section 2, the issue of whether practical reasons are hypothetical or (at least sometimes) categorical can be seen to concern the application of the categorisation scheme for practical reasons. In that practical reasons are normative entities, they depend upon the operation of some (normatively significant) categorisation scheme, where this scheme determines which objects/facts/state of affairs

count as reasons for which agents to perform which actions. Thus it is the categorisation scheme for practical reasons which establishes that some object/fact/state of affairs, *x*, is a reason for some agent, *a*, to perform some action, *φ*. Depending on the nature of this scheme, something's being a reason for action will either be dependent upon, or independent of, the contents of some agent's existing desires.

For example, on an prototypical Kantian account, *x* is a reason for *a* to *φ* iff *a* can will that a maxim according to which she *φ*'s, given *x*, is a universal law. On this categorisation scheme, practical reasons are at least sometimes categorical in that some of the maxims that it is possible to will to be a universal law do not (according to the standard Kantian approach, at least) involve reference to the contents of an agent's desires. By contrast, on a prototypical Humean account, *x* is a reason for *a* to *φ* iff *a* has some desire such that her *φ*'ing, given *x*, would promote the satisfaction of that desire (or else it is believed that it would promote it). On this categorisation scheme, agents' practical reasons (plausibly) depend on the contents of their desires, given that this determines the actions which would count as promoting their satisfaction.[4]

In addition to practical reasons' categoricity or hypotheticality, there is the further question of whether practical reasons' normative force is mind-independent or not. On a mind-independent (MI) view, the normative force of practical reasons is independent of any agents' existing attitudes. For example, the normative force of practical reasons might stem from its being a constitutive requirement of agency that one reasons in certain ways (Goldman, 2010: 181-5; Korsgaard, 2009 chs. 2 & 4-7; Velleman, 2000: chs. 1 and 8). On such views, the normative force of practical reasons is mind-independent in that, regardless of whether anyone cares about her reasons for action or not, she is bound to pay attention to

[4] For present purposes I have assumed, *contra* Mark Schroeder's suggestion (2007: ch.6), that there are no actions which would promote the satisfaction of any possible desire. This is a simplifying assumption, which allows me to avoid discussing certain nuances regarding different Humean approaches to practical reasons at this point.

them, given that it is a constraint on her being an agent she tends to respond to her practical reasons in certain ways.

According to a mind-dependent (MD) view of normative force, the normative force of practical reasons is dependent on some agent's existing attitudes. For example, the normative force of practical reasons might stem from one's having certain desires whose objects would be promoted by responding to certain entities (one's reasons) in certain ways. On this kind of view, it is a desire for some end which attaches normative significance to one's pursuing it. This kind of view seems to arise from Mark Schroeder's (2007) account of practical reasons.

From the above discussion, it can be seen that the mind-(in)dependence of practical reasons comes in two forms: the categorisation scheme for practical reasons applying independently of the contents of agents' existing desires; the normative force-maker for practical reasons being independent of agents' existing attitudes. This yields four available positions, so far as mind-(in)dependence is concerned. These are:

1. MI-normative categoricalism (the view that the practical reasons that agents have are (at least sometimes) independent of the contents of their existing desires, and that the normative force of these reasons is independent of any agent's existing attitudes).

2. MI-normative hypotheticalism (the view that the practical reasons that agents have are always dependent on the contents of their existing desires, but that the normative force of these reasons is independent of any agent's existing attitudes)

3. MD-normative categoricalism (the view that the practical reasons that agents have are (at least sometimes) independent of the contents of their existing desires, but that the normative force of these reasons is dependent on their existing attitudes).

4. MD-normative hypotheticalism (the view that the practical reasons that agents have are always dependent on the contents of their existing desires, and that the normative force of these reasons is dependent of the their existing attitudes too).

MI-normative categoricalism is most strongly associated with a Kantian view of practical reasons, such as that proposed by Korsgaard (2008: ch.7; 2009 esp. chs. 2 & 4-7). On such views what agents have a reason to do (as determined by the categorical imperative) is at least sometimes independent of their existing desires. Meanwhile, practical reasons' normative force comes from the (mind-independent) normative status of rationality. Korsgaard, for instance, regards this as a constitutive constraint on choice. Apart from Kantians, MI-normative categoricalism can also be associated with some consequentialist moral theories, and with the primitivist conceptions of practical reasons offered by, for example, Parfit (2006; forthcoming) and Scanlon (1998: ch.1).

MI-normative hypotheticalism is a fairly popular view amongst contemporary Humeans, some of whom have suggested that for agents to have any reason to pursue the objects of their desires, there must be some categorical requirement for them to take the means to their desired ends (Beardman, 2006; Dreier, 1997). However, the practical reasons that the means/end principle generates are hypothetical, in that the means that one will have reasons to take will (plausibly) depend upon one's particular ends.

MD-normative categoricalism is an uncommon view. However, an example of this type of view seems to be Mark Schroeder's account of practical reasons (which he, perhaps confusingly, labels 'hypotheticalism'—Schroeder, 2007). On Schroeder's view, practical reasons are facts/states of affairs which (partially) explain why performing some action will promote the satisfaction of a desire. This view is Humean in spirit, in that it attaches practical reasons to desires. However, Schroeder controversially claims that there can be some facts/states of affairs which (partially) explain why performing some action will promote the satisfaction of any possible desire. Where such facts/states of affairs obtain,

there will be a reason for any agent to perform some action, regardless of the contents of her particular desires (i.e. a categorical reason for action). Nevertheless, the normative force of these categorical reasons is, on Schroeder's account, still explained by the fact that acting on them will promote the satisfaction of at least one of an agent's particular desires (Schroeder, 2007: 79-82).

MD-normative hypotheticalism is perhaps the most Humean of the available views, in that it treats both reasons' application and their normative force as mind-dependent. Thus, perhaps surprisingly, it is difficult to situate any particular account of practical reasons within this camp. I expect that this is partly because the question of explaining practical reasons' normative force has only recently been raised under that specific guise. When it was raised, this question was quickly directed as a challenge to Humeans to explain the normativity of hypothetical reasons, which seemed to have been simply assumed up to that point (Hampton, 1996; Korsgaard, 1997). Attempts to explain this normativity have generally been mind-independent—picking out the categorically prescriptive instrumental principle (Dreier, 1997); or, suggesting that desire-satisfaction is a constitutive aim of action (Goldman, 2010), for example.

From the above discussion, it should be clear that the issue of whether practical reasons are categorical or hypothetical is distinct from the issue of whether their normative force is mind-dependent or mind-independent. So far as offering a naturalistic account of practical reasons is concerned, it is the latter issue which seems to be more important. This is because the main challenge for the naturalist is to offer an account of practical reasons on which their normative force is compatible with a scientific worldview, where ideally such an account will retain the mind-independent normative force of practical reasons.

5. Naturalism and Reduction

So, the key challenge is to give an account of practical reasons which affords a naturalistic explanation of their normative force. One way of meeting this challenge is to give a

reductive account of practical reasons' normative force. Reductions come in two kinds: ontological and conceptual.

Ontological reductions involve taking some target property, P (which is perhaps somewhat mysterious on first consideration) and showing that it is identical to some other (perhaps less mysterious) property, Q. For instance, Kripke's familiar example of water being identical to H_2O involves a case of ontological reduction (Kripke, 1980: 128-9). Conceptual reductions involve taking some target concept, C (which is perhaps somewhat elusive on first consideration) and showing that it is identical to some other (perhaps more familiar concept), D. For instance, it has been suggested that Bentham tried to define rightness in terms of conduciveness to general happiness (Bentham, 1988: ch. I § I note, and §§ IX & X; Moore, 1903: ch.1).

Conceptual and ontological reductions do not necessarily go together. For instance, the above example of water being ontologically reducible to H_2O is a case of ontological reduction without conceptual reduction. Our concept of water, as Kripke points out, is that of an odourless, colourless liquid which is non-toxic to humans, etcetera. Our concept of H_2O is that of a simple molecule made up of two hydrogen atoms and one oxygen atom. These concepts are not identical, even though water (the stuff) is H_2O.

So far as this thesis is concerned, it is ontological reduction which matters. So long as any properties which are attributed to an entity, according to some theory, are compatible with a scientific worldview, that theory is naturalistic. It does not matter, so far as naturalism is concerned, whether the concepts that we use to refer to such properties are also reducible to naturalistic concepts. For naturalism, how we think and talk about something can be as non-scientific as we like, so long as the existence of whatever we ultimately turn out to be referring to can be naturalistically accounted for.

Non-reductivism about practical reasons, the view that practical reasons cannot be explained without employing irreducible normative concepts, *and* that practical reasons' existence involves the instantiation of irreducible normative properties, is, I have supposed

incompatible with a naturalistic approach. As such, non-reductive views of practical reasons are set aside for the purposes of this thesis.

This leaves two available reductive approaches. The first is to treat practical reasons' normative force as ontologically reducible to some natural property kind, while claiming that the notion of normative force evades conceptual reduction (non-reductive naturalism—where 'non-reductive' means 'non-conceptually-reductive'). The second is to treat practical reasons' normative force as ontologically reducible to some natural property kind, and the notion of normative force as reducible to some naturalistic concept (reductive naturalism). For my purposes, it is the issue of ontological reduction which is of primary importance. The conceptual irreducibility of the normative is compatible with naturalism, such that questions concerning it are an issue of detail within any naturalistic approach, rather than definitive of a view's naturalistic credentials.

One way of avoiding the issue of reduction altogether is to deny that there are any practical reasons. For example, one might adopt an error theory about practical reasons, disposing of practical reasons entirely by treating all positive statements about practical reasons as false. Since Mackie's famous attack on moral realism, error theory has been a popular view in metaethics, with fictionalist explanations of moral discourse enjoying enduring attention in the literature (Mackie, 1977; Joyce, 2001; Kalderon, 2005; Nolan, Restall and West, 2005). However, error theories of practical reasons in general are not similarly popular. Such a view would involve a general error theoretic account of normative properties. In a review of recent work on normativity, Finlay sets aside such views commenting that 'error theory about normativity as such is virtually unheard of' (Finlay, 2010: 334).[5]

Although generalised error theories of the normative are unpopular, they might prove to be the last resort for the naturalist if no other account of normative phenomena can be found. If the naturalist is to escape the conclusion that all of our positive claims about practical reasons are in error, she must find some way to account for them which is compatible with

[5] Streumer (forthcoming) does situate himself in this camp, however.

her naturalistic outlook (or else abandon her naturalistic prejudices). Before countenancing an error theory of practical reasons, I prefer to take up this search. Human beings are practical creatures, who habitually make reason-judgements and talk about practical reasons in conducting their everyday lives. I prefer not to find this activity in radical error.

Another way of avoiding the issue of reduction altogether is to offer a non-cognitivist account of normative language. On non-cognitivist views, the purpose of normative discourse (including talk about practical reasons) is not to state beliefs about normative facts but to express affective states and/or to issue prescriptions.[6] As such, non-cognitivism appears to dispense with practical reasons altogether, given that practical reasons are entities which instantiate normative properties, and of which we can predicate such properties by asserting normative propositions. However, this might be too quick, as the non-cognitivist can maintain that there are entities which 'instantiate normative properties', while offering a non-cognitivist account of what this involves, say, in terms of the role that such idioms might play in the expression of an affective response towards some state of affairs. She can also maintain that we do predicate normative properties of certain entities by asserting normative propositions, while giving a non-cognitivist explanation of what it is to assert a normative proposition. So long as the non-cognitivist's account of normative language use is compatible with the normal linguistic constraints (logical, grammatical and so on) which apply to the statement of propositions, she is free to maintain that moral language is 'propositional', even though it does not express normative beliefs (Blackburn, 1984: esp. ch. 7; 1993: ch. 9; Sinclair, 2007).

Nevertheless, I maintain that non-cognitivism still disposes of practical reasons in the prototypically normative sense of the term. This is because, on a non-cognitivist account, 'normativity' and 'normative properties' are to be understood in terms of our uses of normative language, rather than in terms of relations which hold between agents and their natural environment independently of such language. Thus although something can 'instantiate a normative property', in the sense that this idiom can be understood to play a

[6] Prescriptions, here, are directives issued by one person to another.

role within a non-cognitivist account of normative discourse, nothing can instantiate a normative property in a sense of property instantiation which is independent of that discourse. This places non-cognitivism at odds with any attempt to characterise practical reasons as entities of which instantiate normative properties in a prototypical sense of property instantiation. Thus I set aside non-cognitivism, which disposes of practical reasons in the sense of the term with which I am concerned.

A third way of avoiding the issue of reduction altogether is to adopt a constructivist approach to practical reasons. This kind of approach involves claiming that the role of practical reasoning is not to discover normative truths which exist independently of our practical reasoning processes (e.g. to discover what practical reasons we have). Rather, (correct) practical reasoning is a process of constructing normative truths, including truths about our practical reasons. The normative force of practical reasons, on this approach, derives from the correctness of a practical reasoning process which picks out certain entities as practical reasons. That it can be correct to reason in favour of responding to some entity in some way explains the normative significance of adopting such a response.[7]

If practical reasoning is not directed at uncovering normative truths, but at constructing them, then on a constructivist approach there are (supposedly) no troubling normative entities or relations for the naturalist to accommodate. The issue of whether normativity can be reductively explained or not simply does not arise, as normativity is seen to be entirely dependent on the process of normative reflection. Normative reflection is, plausibly, a natural process, insofar as it involves agents having certain brain states. But, for the constructivist, normative truths (including truths about normative force) are explained by the content and structure of a (correct) normative reasoning process and not by facts about its physical realisation. Normative truth is a matter of what we (correctly) do with normative concepts.

[7] Clearly, an account of correctness in reasoning is owed by the constructivist. This issue is discussed in what follows.

So, the ontological side of the constructivist's equation is compatible with naturalism (she can accept that all that exists is natural, including our reasoning processes *qua* physically instantiated events). But, for the constructivist, the ontological side of the equation does not, in any way, explain normative truth. Normative truth is explained by what we (correctly) do with normative concepts—by the conceptual side of the equation. It is the correct application of normative concepts which makes certain normative propositions (including propositions about practical reasons) true, regardless of any ontological dependencies which exist between instances of practical reasoning *qua* physical processes and the physical world. As such, the constructivist has nothing normative to give an ontological reduction of in order to retain a compatibility between constructivism and a naturalistic worldview. Normativity is explained by the structure and content of correct practical reasoning; this is neutral with respect to the ontological concerns of naturalism.

It should be obvious from the above exposition that one of the biggest challenges for the constructivist is to offer an account of what makes a process of practical reasoning correct. Korsgaard, whose account of practical reasons I introduced in section 4, above, takes a constructivist line in explaining the normativity of practical reasons (Korsgaard, 1996: § 1.4.4; 2008: ch.10). According to Korsgaard's account, the normativity of practical reasons is explained by the correctness of a process of practical reasoning by which an agent takes herself to have certain reasons for action. For Korsgaard, correctness in reasoning is in turn explained by the constitutive constraints which apply to engaging in a practical reasoning process at all (where these are, equally, the constitutive constraints on action, as action is a form of behaviour which is appropriately connected to practical reasoning). These constraints, as outlined above, are taken to include Kantian principles of practical rationality.

So, Korsgaard proposes that in order to reason practically at all, an agent must (generally) reason in certain ways. Practical reasoning which is carried out in these ways is correct. The reasons which follow from a correct process of practical reasoning have normative force because they derive from such a process (i.e. because they are the reasons which our being able to reason practically at all requires that we recognise).

This kind of agent-constitutive explanation of correctness in reasoning, and thereby of normative force, has a certain appeal.[8] However, it is problematic as a feature of a constructivist account of practical reasons' normativity. This is because some explanation of why the constitutive requirements of practical reasoning make a process of reasoning 'correct' is needed. Providing such an explanation seems to involve stepping outside of a constructivist framework; to involve invoking something independent of any process of practical reasoning in explaining why reasoning in accordance with the constitutive requirements of practical reasoning is the correct thing to do (this type of point is made by Hussain and Shah, 2006).

So, for instance, it might be claimed that what makes reasoning in accordance with the constitutive requirements of practical reasoning correct just is that *correct practical reasoning* is identical to *reasoning conducted in accordance with the constitutive norms of practical reasoning*. In this case the view seems to be ontologically reductive. Correctness in reasoning is being explained by a proposed identity relation between the property of being *correct practical reasoning* and the property of being *reasoning conducted in accordance with the constitutive requirements of practical reasoning*. This identity relation, if it obtains, is independent and prior to any process of practical reasoning. As such, it is not a process of practical reasoning which explains the normativity of practical reasons. Rather, it is the constitutive requirements that apply to something's being a process of practical reasoning. So, this version of the agent-constitutive view of correctness in reasoning (and, therefore, of normative force) is not constructivist, but ontologically reductive.

Alternatively, it might be claimed that what makes reasoning in accordance with the constitutive norms of practical reasoning correct is the existence of some *sui generis* normative prescription in favour of reasoning in accordance with these norms (i.e. a prescription in favour of being an agent). In this case, the view seems to be non-naturalistic. Correctness in reasoning is being explained by a proposed prescription in favour of

[8] I pursue a similar explanation of normative force in chapter 2, albeit one which is framed in terms of the constitutive requirements of interpretability rather than of agency.

reasoning in accordance with the constitutive norms of practical reasoning. This prescription, if it applies, is independent and prior to any reasoning process, such that it is not a process of practical reasoning which explains why it applies. So this version of the agent-constitutive view of correctness in reasoning (and, therefore, of normative force) is not constructivist either, but non-naturalistic.

Absent an explanation of why reasoning in accordance with the constitutive norms of practical reasoning is correct which is framed in terms of some reasoning process itself, it seems that making good on Korsgaard's account of correctness in reasoning involves situating her view outside of the constructivist camp.

Are there any alternative accounts of correctness in reasoning which are more thoroughly constructivist? One proposal, from Street (2008), is that reasoning is correct only by reference to a further process of reasoning, which is itself subject to standards of correctness. On this view there is no ultimate explanation of correctness in reasoning. Correct reasoning is reasoning that we (correctly) reason to be correct. This understanding of correct reasoning is somewhat unpalatable (given its tendency towards regression) but it is more successfully constructivist than Korsgaard's approach. It is agents' reasoning processes which explain the correctness of reasoning and thus the normativity of practical reasons.

However, this view of correctness does not seem to take into account the existence of constitutive constraints on our reasoning processes, such as the kind that Korsgaard invokes. It is plausible to think that there are constitutive constraints on the kinds of practical reasoning that we can apply. Parallels between theoretical reasoning and practical reasoning have been drawn here. For example, it has been suggested that just as following certain rules of logical inference is constitutive of theoretical reasoning, so the application of a means-end principle (say) is constitutive of practical reasoning (Dreier, 1997; Railton, 1997). It is hard to characterise what practical reasoning is without positing certain constraints of this kind on what it *is* to be a process of practical reasoning.

Korsgaard's suggested constraints on practical reasoning are more stringent than the mere means-end principle cited above. However, a circular explanation of correctness in reasoning of the kind proposed by Street does not allow that constraints on practical reasoning of any kind whatsoever, even instrumental ones, can play a role in the explanation of correctness in reasoning. Given that an account of correctness in reasoning is available in terms of these kinds of constraints (even if it involves abandoning constructivism), the suggestion that correct reasoning can only be explained by reference to a further reasoning process seems unwarranted. Some degree of correctness, at least, seems to derive from the constraints which apply to something's being a process of reasoning in the first place.

Further, it is not entirely clear why a circular account of correct reasoning does not turn out to be a form of eliminativist conventionalism about normativity. Given that it treats normativity as being ultimately groundless, it seems that the circular view of correct reasoning disposes of normative force and replaces it with reasoning convention. If this is the case then accounting for practical reasons' normativity along circular constructivist lines is a non-option.

For these reasons, I consider circular constructivism to be an unappealing way to avoid the issue of reduction. Assuming that a non-circular form of constructivism, such as Korsgaard's, must invoke some non-constructivist explanation of correctness in reasoning, I set aside constructivism as way of providing a naturalistic account of practical reasons' normative force. This leaves ontological reduction as the only remaining avenue for the naturalist to take if she wishes to maintain that there are practical reasons in a prototypically normative sense while retaining her naturalistic worldview.

6. Ontologically Reductive Accounts of Practical Reasons

In section 4 I explained that the key challenge for the naturalist is to give an account of practical reasons on which their normative force is compatible with a naturalistic worldview. Ideally, such an account will treat practical reasons' normative force as mind-independent. However, an account which treats practical reasons' normative force as mind-dependent

may need to be adopted, if an account of normative force as mind-independent turns out to be unavailable to the naturalist.

In section 5 I introduced ontological reductivism as one way to give a naturalistic account of practical reasons' normative force. I also discussed some alternative approaches, including error theory, non-cognitivism and constructivism, but set each of these aside. The first two were rejected for eliminating practical reasons entirely (at least in the prototypically normative sense with which I am concerned). The third was rejected for either depending on some non-contructivist approach (Korsgaard's agent-constitutive constructivism) or else providing an unsatisfactory (and perhaps eliminativist) treatment of normativity (Street's circular constructivism).

This leaves us with the option of reducing practical reasons' normative force to some naturalistically acceptable property. In this section I discuss two approaches to practical reasons which afford an ontological reduction of their normative force. On the first (neo-Aristotelian Naturalism) practical reasons have mind-independent normative force. On the second (Mark Schroeder's neo-Humeanism) practical reasons' normative force is mind-dependent.

These views are discussed as instructive examples of the kind of approach available to the naturalist about practical reasons. I offer some criticism of the views, by way of illustrating my motivations for developing an alternative account of practical reasons, but I do not take these criticisms to be decisive. Nor do I take the discussion to be exhaustive of the naturalistic options available.[9] The purpose is to establish a context for the account of

[9] In chapter 2, section 5, I discuss my own account of practical reasons' normative force. This account draws on the agent-constitutive approaches to normativity given by Goldman (2010: 66-82 & 181-5), Korsgaard (1996: esp. §1.4.4 and §3.3.1; 2008, ch.3; 2009, esp. chs. 2 & 4-7) and Velleman (2000: chs. 1 and 8) each of whose views of practical reasons has naturalistic credentials. All of these views ground practical reasons in the constitutive requirements of agency, and they are thereby very similar to my own approach, which grounds practical reasons in the constitutive requirements of interpretable functioning. Thus rather than discuss these views as examples of an agent-constitutive

practical reasons to be provided in this thesis, by canvassing some existing naturalistic approaches to practical reasons and highlighting some of the issues which arise in relation to these.

6.1. Neo-Aristotelian Naturalism

One popular naturalistic approach has been to offer an account of normative properties and relations in terms of human beings' natural mode of functioning. This kind of Aristotelian project has been pursued by the later Foot (2001), Hursthouse (1999), Nussbaum (1995) and Thomson (1997), among others.

The general idea here is that there are ways of functioning that are distinctive of different kinds of natural creatures, where functioning in these ways is conducive to the survival, maintenance and reproduction of such creatures. To use Foot's example, hunting in packs is part of wolves' way of surviving. It is part of the natural functioning of wolves to hunt in packs. Thus we can say that *'wolves hunt in packs', in the sense that wolves which function* in the way that they are naturally suited to function hunt in packs.

From this kind of classificatory claim, it is then suggested that creatures which do not function according to their natural kind are not 'as they should be', in the sense that they are not functioning in the way which they are naturally suited to function. A wolf which hunts alone, or which free-rides on the hunting activities of its pack, is not functioning in the way that it is naturally suited to function. This is a 'defect'.

This line of thought leads to the idea that we can understand practical reasons in terms of the requirements of functioning in the way that we are naturally suited to function. Where

---

approach to practical reasons in the current chapter (over and above their appearance in the discussions of normativity and, in Korsgaard's case, the discussion of constructivism) I postpone their discussion until chapter 2, where I explain my focus on the constitutive requirements of interpretable functioning rather than the constitutive requirements of agency.

human beings are concerned, natural functioning includes being rational, as well as other more animal traits. Humans are normatively required to behave rationally in the sense that if they do not, they fail to function in the way that they are naturally suited to function. Thus it is the natural functioning of human beings which, on neo-Aristotelian accounts, provides the normative force of practical reasons. Because human beings' mode of natural functioning is independent of any agent's existing attitudes, this normative force is mind-independent.

I have left many details out of this quick sketch of the neo-Aristotelian naturalist's position.[10] However, it should be clear what the general neo-Aristotelian approach to practical reasons involves. One worry with this approach is that using the term 'reason' to denote that something is in accordance with our natural mode of functioning might not seem to be properly normative. Thus one might wonder whether there is any normative force which backs up the requirements of functioning after our natural kind. Sure, a human who acts irrationally may be labeled 'defective', or 'not as she should be', but these terms need to import more than the idea of 'not functioning in a certain way' if they are to count as truly normative. What we need is some account of why functioning after the manner of our species is *to-be-done*, not just something that we are equipped to do and which we can fail to do.

One response here is to suggest that the objector is asking for too much. The objection involves asking for an explanation of the normative force of natural functioning when, for the neo-Aristotelian naturalist, normative force might simply reduce to something's being a part of our natural mode of functioning. This kind of response is canvassed by Finlay, who remarks that:

> Some worry, however, that neo-Aristotelianism cannot meet *normative* challenges. Why think that we have any *reasons* to avoid being defective members of our kind? Some Aristotelians have adopted the same response as the quietists. If to have a

[10] See Lenman (2008) for more, or see Foot (2001) or Hursthouse (1999) for full treatments.

> reason to do A is nothing other than (e.g.) for it to be the case that we would be defective human beings if we did not do A, then there is no coherent challenge here (Foot, 2001, Thomson, 2008). (Finlay, 2010: 339).

Finlay is right that this response avoids the charge that normativity is missing on a neo-Aristotelian account. On such an account, normativity reduces to being necessary to accord with our natural mode of functioning. No further explanation of normativity is required.

However, if the neo-Aristotelian is offering a reductive account of normativity, the important question is then whether this account is any good. That is, does the proposed reduction capture the relevant features of normativity? One problem here is that the neo-Aristotelian's account does not seem to allow for the existence of reasons to act in ways which go against our natural functioning. Ultimately, neo-Aristotelian accounts of natural functioning are cashed out in terms of functioning in some manner which is conducive to the natural, biological ends of survival, maintenance and reproduction. However, there can be reasons for us to act in ways which are not conducive to these ends. For example, I can have a reason to smoke cigarettes (e.g. that smoking them involves a certain degree of pleasure) even though smoking militates against all of the ends which our natural functioning aims towards. This, and other similar examples (such as the existence of pragmatic and/or emotional reasons for abortion; the existence of reasons for suicide; the existence of reasons to take intoxicating substances; the existence of certain reasons to be reproductively sterilised; the existence of reasons to use contraception; the existence of reasons for self-sacrifice; and so on) suggests that neo-Aristotelian naturalism is out of step with our ordinary concept of practical reasons.[11] Sometimes we have reasons to act in ways which are not conducive to nature's ends.

The neo-Aristotelian can reply to this objection by suggesting that it involves a misunderstanding of her view. Particularly, the objection ignores the special role of rationality within that view. Rationality is a particular mode of functioning which, it is

[11] None of these suggested reasons are taken to be indefeasible, of course.

supposed, is generally conducive to humans' survival, maintenance and reproduction. Thus on the neo-Aristotelian approach, the categorisation scheme for practical reasons is provided by rationality, rather than by conduciveness to survival, maintenance and reproduction itself. Conduciveness to our natural ends of survival, maintenance and reproduction then gives rationality its normative force, but it is rationality which determines which reasons we have rather than these specific ends in themselves. Although rationality is generally conducive to the achievement of these ends, it does not necessarily involve acting ways which will promote them. So, one can have reasons to act in ways which are not conducive to survival, maintenance and reproduction because rationality, our natural mode of functioning, sometimes supports acting such ways.

For this reply to succeed though, it needs to be shown that rationality *is* generally conducive to survival, maintenance and reproduction. One way of doing this would be to suggest that rationality involves a commitment to certain substantive aims; in particular, the aims of survival, maintenance and reproduction. However, aside from the problem of justifying a substantivist account of rationality which picks out these particular aims, this approach seems to lead straight back to the original objection. This was that neo-Aristotelianism seems to rule out our having reasons to act in ways which are not conducive to survival, maintenance and reproduction, where we intuitively have such reasons (at least in certain circumstances). If rationality has the aims of survival, maintenance and reproduction built into it then significant work needs to be done in showing how it can nevertheless be rational to act in ways which conflict with these aims.

A second way of showing that rationality is conducive to survival, maintenance, and reproduction is to claim that, as a matter of fact, these are the primary aims that we (generally) have. For any creature who has these primary aims (which plausibly we do, given our biology), even a capacity for instrumental rationality will be conducive to survival, maintenance and reproduction, as it will help us to act in ways which promote these ends. Thus even if, given a merely instrumental account of rationality, we can have reasons to act in ways which are not conducive to survival, maintenance and reproduction (because we sometimes have other aims), our capacity for instrumental rationality will be generally

conducive to our survival, maintenance and reproduction, given that these are our primary aims.

Supposing that this suggestion is right, the objection that neo-Aristotelianism cannot account for reasons to act in ways which go against survival, maintenance and reproduction can be overcome. Such reasons can exist on an instrumental account of rationality, while a capacity for instrumental reason can be treated as generally conducive to survival, maintenance and reproduction, given that these are our primary aims.

However, at this point a different objection arises. This is that, in cases where we have reasons to act in ways which are not conducive to survival, maintenance and reproduction, it is still supposed to be instrumental rationality's general conduciveness to these aims which explains the normative force of such reasons. This seems to be wrong. For instance, imagine a case in which someone (Emily, say) is most concerned with the emancipation of women in her country. Suppose that Emily only cares about her survival, maintenance and reproduction insofar as they are conducive to the achievement of this aim. Further, suppose that Emily has an opportunity to martyr herself, and that this is the best way for her to draw attention to the injustices against women which take place in her country. On an instrumental account of rationality, Emily's strongest reason for action will favour taking this opportunity, despite its going against her survival, maintenance and reproductive prospects. Nevertheless, the normative force of this reason will still derive, on a neo-Aristotelian account, from the general conduciveness of instrumental rationality to the survival, maintenance, and reproduction of creatures who have these as primary aims.

This is highly counter-intuitive. The proposed example involves a person of whom the following is true: she does not really have the primary aims of survival, maintenance and reproduction (her primary aim is the emancipation of women in her country); she strongly identifies with an end whose means to achievement, on this occasion, goes against these aims; and she has a strong reason for performing an action which will terminate her ability to survive, maintain herself and reproduce. It seems very peculiar to think that the normative force of this reason stems from the general conduciveness of instrumental

rationality to survival, maintenance, and reproduction in creatures who have these as primary ends, given that Emily does not and that her strongest reason favours acting in a way which is obstructive of them.

More generally, one might wonder why conduciveness to survival, maintenance and reproduction would be considered to be the right reduction base for normative force in the first place. It seems highly counter-intuitive to think that the normative force of rationality simply boils down to its general conduciveness to these ends. Many people, it has been suggested, are not entirely concerned with these ends. Thus, at the very least, the biological reduction of normative force posited by the neo-Aristotelian naturalist is intuitively inaccurate. It does not always match up with our common perception of when something has normative significance. If biological 'imperatives' do not even seem to be imperative to many people, at least not in all circumstances, it is doubtful whether the notion of normative force can be successfully cashed-out in terms of them.

Although this worry is not a decisive refutation of neo-Aristotelianism, I take it to illustrate a particular way in which that view is unappealing. I think that neo-Aristotelianism is right to pick out rationality as the categorisation scheme for practical reasons. I also think it is right in claiming that rationality has normative force because it is connected to a distinctive mode of functioning that human beings naturally inhabit (that normative force is to be explained in terms of what some natural mode of functioning involves). However, I think that the neo-Aristotelian's focus on humans' natural functioning, as directed at the biological ends of survival, maintenance and reproduction, is wrong. These aims do not always seem to be of primary importance to us, such that rationality's general conduciveness to their achievement does not seem to be capable of providing the normative force behind many of the practical reasons that we have.

6.2. Schroeder's neo-Humean Account of Practical Reasons

Schroeder offers an account of practical reasons which is framed in terms of facts/states of affairs which (partially) explain why certain actions will promote the object(s) of our desires. Hence:

> **Reason** For *R* to be a reason for *X* to do *A* is for there to be some *p* such that *X* has a desire whose object is *p*, and the truth of *R* is part of what explains why *X*'s doing *A* promotes *p* (Schroeder, 2007: 59).

This is an ontologically reductive view; no *sui generis* normative properties are posited in the account of reasons. Reasons are facts/states of affairs which (partly) explain why acting in a certain way will promote the object of some desire that an agent has. For example, that it is raining is a reason for me to carry an umbrella in the sense that the rain (partly) explains why carrying an umbrella will help me to avoid getting wet (which I desire).

Schroeder's is an economical view of reasons. No new entities or relations are posited by the account, and no particular view of the goal of human functioning is endorsed. Reasons are simply identified with facts/states of affairs which stand in a certain kinds of explanatory relation to the satisfaction of our desires. The account also seems to be well motivated, in the sense that the facts and explanatory relations of interest are those which pertain to the satisfaction of our desires – something which we are practically interested in. Thus although the account eliminates mind-independent normative relations, it seems to replace these with something suitably practical: explanatory relations of significance to our everyday practical successes and failures. Schroeder also avoids the worry that his account is unfaithful to our ordinary notion of a practical reason by stipulating that he is not offering a reduction of the concept of a reason for action, but only of the property of being a practical reason (ibid: ch.4).

Nevertheless, I am not persuaded by Schroeder's account of reasons for action. One reason for this is that, on his account, normative force is (it seems) grounded in agents' desires.

That some state of affairs (partially) explains why some action will help to promote the satisfaction of a desire is normatively significant, it seems, only in the sense that the agent *has* that desire and is therefore concerned with its satisfaction. Schroeder does not discuss the notion of normative force in particular, but this seems to be the view of it which comes out of his account of practical reasons. He does discuss normativity, which he contends is best accounted for in terms of reasons (ibid: 79-81). It is this contention which suggests that, on Schroeder's view, normative force is about the having of certain desires, where the objects of these desires will be promoted by acting on our practical reasons. If normativity is accounted for in terms of reasons, and reasons are accounted for in terms of explaining what will promote the objects of our desires, then the normative force of reasons seems to boil down to the having of certain desires to whose satisfaction our practical reasons relate.

I find this view unappealing, because it dispenses with practical reasons' mind-independent normative force. The normativity of practical reasons, on Schroeder's view, derives from agents having certain desires whose satisfaction will be promoted by their acting in certain ways, given certain facts/states of affairs (their practical reasons). Although this may be a promising avenue to take if no mind-independent account of normative force can be given, I consider a mind-independent explanation of the normative force of practical reasons to be preferable, if one is available. Practical reasons appear to matter over and above our having any particular concern which they relate to.

A second worry with Schroeder's account concerns our interest in other people's reasons for action. Sometimes we do not share the desires that other people have, and sometimes we have desires which run in the opposite direction to theirs. In these cases we do not want what other people want, and we do not want them to do what will promote the satisfaction of their desires. Nevertheless, we do seem to be concerned with their acting in accordance with their reasons for action.

For example, suppose that I am at a birthday party where there is one piece of cake left and that John, who is also at the party, would like to eat it. John is considerate enough to ask if anybody else would like the last piece of cake. I would like to eat it but, being aware that

John would like to eat it too, I am polite enough to say that I have already eaten enough cake. Nobody else shows an interest in the cake.

Although I have not registered my interest in eating the final slice of cake, I still do not want John to take it from the salver, or to eat it. That would eliminate any possibility of my having it. However, were John to simply leave the cake on the plate, I would think something was wrong. I would think that John *should* have taken the cake, for he wants to eat it and in order to eat it he must take it from the plate. Thus, as well as having a negative interest in John's taking the cake, I also seem to have a positive interest in his doing so (at least in the sense that I will be quite perplexed if he does not).

Suppose that Schroeder's account of reasons is right; reasons are facts/states of affairs which (partially) explain why acting in a certain way will promote the object of a desire. One question is, why would I be concerned with John's acting in accordance with his reason if (in this case) it relates to his doing something which will promote the object of a desire that, as it happens, I do not want him to satisfy? If I do not want John's desire to be satisfied, why do I nevertheless want him to do the things which will help to promote the satisfaction of this desire?

Schroeder can answer this question. He can suggest that other people's tendency to act on the basis of their practical reasons is likely to have a significant influence on my ability to satisfy my own desires. If other people were entirely unpredictable to me, then I would struggle to satisfy any desires which involved coordinating my behaviour with theirs. Reliably being able to predict what other people will do requires that they behave consistently, and this partly involves their responding to their reasons for action in particular ways. So, as well as wanting John to leave the cake, so that I can have it, I want him to take the cake because I want him to act intelligibly such that he falls in with my general desire that people are predictable.

This is a reasonable answer to the question of why, on Schroeder's account, we might take an interest in other people's practical reasons, even when we do not want the desire that

they relate to to be satisfied. However, I think that the explanation of our concern with other people's reasons for action can be made to run deeper. The above suggestion was that our concern for other people's reasons is explained by an interest in their predictability. Although we are concerned with being able to predict what other people will do, predictability is just one member of a larger set of concerns which we have with the practical lives of others. We are also concerned with evaluating others, with identifying with them, and with forming relationships with them (among other things). Each of these concerns is enabled by our ability to interpret others—our ability to attribute specific attitudes to them in the light of their behaviour. Attributing such attitudes allows us to appraise others' choices, to identify with their concerns, and to develop personal relationships with them (as well as predicting what they will do).

Schroeder's account of practical reasons (in terms of the promotion of desire satisfaction) is compatible with our having each of these concerns. It is also compatible with these concerns explaining our interest in others' practical reasons, as characterised by his account. If others act in interpretable ways then this promotes the satisfaction of many of our other-involving desires. That others' acting interpretably will promote the satisfaction of these desires explains our concern for other people to act in accordance with their (overall) balance of reasons.

However, as well as asking what interests of ours are promoted by others' acting in accordance with their reasons for action, we can also ask why it is that their so-acting promotes these interests. It is at this point that I think Schroeder's account is unappealing. This is because, given Schroeder's account of reasons, interpretability is an incidental outcome of people's acting in accordance with their overall balance of reasons.[12] For Schroeder, acting for a reason is about the promotion of desire satisfaction; promoting desire satisfaction just happens to make people interpretable, such that our interests in other people's lives can be sustained. However, given Schroeder's view, when we take an

[12] I ignore Schroeder's controversial account of reasons' weight, and its implications for deliberation and interpretability, here. See chapter 5, section 2 for discussion.

interest in others acting in accordance with their reasons for action we are not necessarily interested in their acting on these reasons *per se*. Rather, we are (I have suggested) interested in something which just happens to follow from their doing so: namely, that it makes them interpretable to us.

My suggestion is to treat interpretability as an essential, rather than an incidental, feature of people's acting in accordance with their practical reasons. That is, the suggestion is to regard acting in accordance with one's (overall) balance of reasons as essentially a matter of acting in an interpretable way. On this approach others' acting in accordance with their practical reasons does not just happen to bear on the satisfaction of any desires that we might have regarding them as persons; acting interpretably necessarily bears on the satisfaction of such desires. Thus on this view when we take an interest in others acting in accordance with their reasons for action, what we are interested in *is* their acting on such reasons viz. their acting in an interpretable way.

This suggestion is not taken to undermine Schroeder's account of practical reasons. Rather, my intention is to expose a particular way in which I find his account superficial, while indicating the direction in which I think a less superficial account of our concerns with others' practical reasons lies. As such, the concerns that I have raised are not taken to be fatal to Schroeder's view which, for all intents and purposes, is taken to be a live and credible alternative to the account of practical reasons offered in this thesis. Nevertheless, the above points illustrate my motivation for looking beyond Schroeder's view in attempting, in what follows, to account for practical reasons directly in terms of the requirements of interpretable functioning.

7. Conclusion

In this chapter I have outlined a key challenge for naturalistic accounts of practical reasons. This is to account for practical reasons in such a way that their normative force can be accommodated within a naturalistic worldview. The normative force of practical reasons appears to be independent of agents' existing attitudes. A naturalistic account of practical

reasons will ideally retain this feature. However, if no naturalistic account of practical reasons' normative force as mind-independent can be given, a mind-dependent approach may need to be adopted.

Having introduced and clarified the problem which faces the naturalist, I canvassed some different approaches that the naturalist might take in responding to it. These included ontological reductivism, error theory, non-cognitivism and constructivism. I set aside the final three approaches. These either dispose of practical reasons entirely (at least in the prototypically normative sense of practical reasons with which I am concerned (error theory, non-cognitivism, and perhaps circular constructivism)), fail to afford a convincing account of practical reasons' normative force (circular constructivism), or else fail to be a viable independent approach (non-circular constructivism).

This leaves ontological reductivism as the remaining approach for the naturalist to take. I discussed two ontologically reductive accounts of practical reasons: neo-Aristotelian Naturalism and Mark Schroeder's neo-Humeanism. I suggested that both of these have some appealing features. Neo-Aristotelian Naturalism focuses on rational functioning, and its role as part of our natural mode of human functioning, in accounting for practical reasons. Both of these emphases seem to be broadly correct. However, the biological notion of natural functioning which neo-Aristotelianism employs seems unable to provide an account of practical reasons' normative force which is consistent with our everyday intuitions about why acting on certain practical reasons that we have matters.

Schroeder's neo-Humeanism is appealing in both its economy and in its focus on desire satisfaction, which seems to be a plausible naturalistic candidate for explaining practical reasons. However, as Schroeder's view seems to involve treating practical reasons' normative force as mind-dependent, it does not capture one intuitive feature of practical reasons: that the reasons that we have matter, regardless of any specific concern that we have with adhering to them. Schroeder's account also seems to afford a relatively superficial explanation of our interests in other people's practical reasons. For these reasons I set his view aside.

In the remainder of this thesis I attempt to develop and defend an interpretivist account of practical reasons. The proposed view is ontologically reductive but, I hope, maintains a clear sense in which the normative force of practical reasons is mind-independent. Although it is beyond the bounds of this thesis to address all of the issues which arise concerning our ability to make sense of each other, I hope to at least show that this kind of approach is plausible, and to do something towards its development. To the extent that such an account is plausible, I hope that it will be seen as a worthy naturalistic alternative to existing accounts of practical reasons.

Chapter 2: Outline of an Interpretivist Theory of Practical Reasons

1. Introduction

In this chapter I outline the interpretivist account of practical reasons to be defended in this thesis. This account is grounded in ideas about how agents must function in order to be interpretable – what agents must do in order for us to be able to attribute specific attitudes to them. However, rather than cash out the concept of a reason for action directly in terms of the concept of interpretability, I invoke the notion of rational functioning. This is because interpretability requires both that agents' mental states are (in general) rationally ordered, and that their mental states are appropriately causally connected to their bodily movements. It is the rational side of this equation, rather than the causal side, which seems apt to explain reasons for action.[13] Thus the account of practical reasons is framed in terms of rationality *qua* mode of mental functioning which supports interpretation.

The account of reasons that I wish to develop is, given its interpretivist nature, heavily indebted to the work of Davidson, particularly on radical interpretation, rationality and action.[14] Davidson's work has received much critical attention.[15] I do not attempt to appraise

[13] It may be that a clear line between the rational and causal components of agency cannot be drawn. However, what is important for my purposes is that the aspects of interpretable agency which seem apt to explain practical reasons fundamentally involve the presence of certain rational relations, regardless of how clearly these relations can be distinguished from agency's causal aspects.

[14] On radical interpretation see, in particular, Davidson (1973; 1974a; 1975; 1980; 1995). On rationality and action see Davidson (1963; 1969; 1980; 1982; 1986; 1995).

[15] Some criticisms of Davidson's work on radical interpretation include Klein (1986); LePore and Ludwig (2005: esp. part II); McGinn (1986); Nozick (1993: 154-58); Soles (1999). Some criticisms of his

the merits of Davidson's various philosophical theses here. Rather, I use a number of his ideas as a basis for developing the account of practical reasons that I wish to defend. The purpose of the thesis is, in this respect, primarily developmental; I investigate one way in which an interpretivist account of practical reasons might be developed. I defend this account against some specific objections, but I do not defend the general theories of interpretation, rationality and so on from which it is drawn.

Davidson's work on interpretation, practical rationality and agency is generally formulated in terms of propositional attitudes, such as desires, beliefs, all things considered judgements and intentions. However, the decision theoretic approach to practical rationality that Davidson invokes is more standardly formulated in terms of preferences, subjective probabilities and subjective values/utilities (Ramsey, 1926; Jeffrey, 1983). This does not reflect an underlying tension, so much as a preference for folk terminology where its use is appropriate (i.e. outside of formal decision theory itself). I tend to use the folk terminology of desires and beliefs, except where reference to preferences, subjective probabilities and the like seems either necessary or useful. It is assumed that explanations of desire in terms of preference, and of belief in terms of subjective probability, are available.

In the present chapter, my aim is to suggest one way of developing the idea that practical reasons can be explained in terms of the requirements of interpretable functioning. As mentioned above, I make particular reference to rationality as a mode of mental functioning which supports interpretability. Following Davidson, my take on practical rationality is fundamentally Humean (Davidson, 1963). Thus I seek to explain agents' practical reasons in terms of what rationally follows from their existing motivations.

The Humean approach to practical rationality is often referred to as instrumentalism. However, there is much debate over what instrumental rationality actually involves. For example, it is unclear whether instrumental rationality, properly speaking, includes taking

---

work on rationality and action include Baier (1985); Elster (1999); Føllesdal (1985); Lazar (1999); Levi (1999). For discussions of both of these aspects of Davidson's philosophy also see Ludwig (2003).

the 'constitutive means' to one's ends, or just the causal means to their achievement (Korsgaard, 1997: 215-6).[16] It is unclear whether instrumental rationality applies only to the adoption of necessary means, or whether it also applies to the adoption of sufficient and/or contributory means to one's ends (Broome, 2002: §§6-11). And it is unclear which relata instrumental rationality applies to. Thus Broome (1999; 2002) conceives of instrumental rationality as relating intended means to intended ends; Smith (2004) conceives of instrumental rationality as relating non-instrumental desires to means-end beliefs in a particular kind of way; and Dreier (1997) conceives of instrumental rationality as relating actions to desires and means-end beliefs.

As well as issues concerning the subject matter of instrumental rationality, there is also a question over the scope of the normative operators which feature in instrumental norms. For instance, is it that if you have an end, you ought to adopt the means to its achievement (narrow scope ought), or is it that you ought, if you have an end, to adopt the means to its achievement (wide scope ought)? On the former view, instrumental rationality is a matter of adopting the means which fit with one's ends. On the latter view, instrumental rationality is a matter of making sure that the means that one adopts are consistent with the ends that one has (either by adopting the means to one's ends, or by changing one's ends). Broome (2002: §4; 2007) adopts a wide-scope interpretation of instrumental rationality; Mark Schroeder is critical of this interpretation (Schroeder, 2004).

Finally, instrumental conceptions of rationality seem generally to be cashed-out by reference to certain normative notions, such as those of obligation, requirement, and even the having of reasons (Beardman, 2007; Hubin, 1999, Schroeder, 2004). Although I agree that rationality is normative (in my case, due to its constitutive role in interpretable

[16] A constitutive means is an action which (partly) constitutes the realisation of an end. For example, suppose that I desire to eat a curry. Eating a Madras in my local curry house is one way of achieving this. However, eating a Madras does not cause my desired end (that I eat a curry) to come about. Rather, eating a Madras *is* eating a curry, such that doing so is constitutive of the realisation of my desired end. Hubin (1999: 32) draws the same distinction in terms of 'criterial' and 'causal' means.

functioning), I prefer to formulate rational principles in non-normative terms. As will be discussed in this chapter, I invoke rationality as a categorisation scheme for interpretable mental functioning, where this scheme can be formulated independently of its having normative force. This non-normative characterisation of rational principles is essential to my approach, as to rely on rationality as an independently normative notion would undermine using it as part of an attempt to (reductively) explain what practical reasons (qua normative entities) are. This issue is discussed further in section 4 of this chapter.

Due to a range of unwanted connotations and confusions associated with the above issues, I prefer not to specifically invoke the notion of instrumental rationality in my account of practical reasons. Instead of focusing on instrumental rationality as such, I prefer to focus on a maximising, decision theoretic notion of practical rationality. This notion of practical rationality *might* be regarded as fundamentally instrumental in kind, in that it involves the idea that practical rationality is essentially about selecting those actions which one expects to maximise the satisfaction of one's desires. This is in keeping with Railton's characterisation of instrumental rationality, according to which instrumentally rational agents are those who 'take the means appropriate to their ends, relative to what they believe' (Railton, 2006: 269).

However, Broome, (2002: §10) has suggested that decision theory is not a model of instrumental rationality at all. Instrumental reasoning, on his account, starts out with some intended end and involves deliberating about how to get there. Decision theory, by contrast, relies on a general conception of the good (say, the maximal satisfaction of an agent's desires) and determines what it is rational for an agent to do in order to bring about the best outcome, given this conception of the good. This allows for divergence between what is instrumentally rational and what is rational according to decision theory in specific cases (as well as implying a difference in scope).

For instance, suppose (as in Broome's example) that I intend to buy a boat. Here, the instrumentally rational thing for me to do is to take the best means to the achievement of that end (taking a loan from the bank, perhaps). By contrast, the rational thing for me to do,

from a decision-theoretic standpoint, might not even involve my buying a boat; what will maximise overall desire-satisfaction, given my intention to buy a boat, may be for me to take course in seamanship, for example, rather than to try to buy a boat at all.

If Broome is right that instrumental rationality is about reasoning from intended ends to intended means, then I take his point about decision theory not being a model of instrumental rationality to hold. However, if instrumental rationality is conceived of more broadly as concerning the rational relations which hold between actions and desires, then a decision theoretic model of practical rationality will count as a model of instrumental rationality.

For my purposes it is not important whether a decision theoretic notion of practical rationality counts as a model of instrumental rationality or not. What is important is that the maximising principle associated with decision theory seems to be far clearer than the notion of instrumental rationality commonly associated with Humean accounts of practical reasons: (roughly) act so as to maximise preference satisfaction, given your means-end beliefs. Although formal decision theory faces many problems and criticisms, the basic maximising axiom that standard decision theory invokes is at least clearer than the so-called instrumental principle which, as discussed above, has been formulated in a number of substantively different ways.[17] Further, decision theory is more comprehensive than the mere instrumental principle. As well as the maximising axiom, decision theory places formal constraints on agents' preferences as part of a systematic account of the rationality of preference orderings when taken as a whole. For instance, decision theoretic accounts of practical rationality generally treat intransitive preferences as irrational (Nozick, 1993: 140-1). The comprehensive, formal structure of decision theory allows it to play a role in Davidson's theory of interpretation (which is essentially holistic by nature) which a mere instrumental principle of practical rationality could not.

[17] For discussions of the merits and limitations of decision theory, both as a predictive theory and as a theory of practical rationality, see Bermudez (2009); Hollis and Sugden (1993); Hurley (1989: esp. ch. 4).

Although, given the above discussion, I refrain from developing my account of practical reasons in terms of any supposed instrumental principle, I am content that it falls within a broadly instrumental approach, both to practical rationality and to practical reasons. On this approach, practical rationality consists in acting so as to promote one's desired ends, rather than in adopting any particular ends that it is (supposedly) rational for one to have. This approach has been very popular in discussions of reasons and rationality.[18]

However, Humean approaches to rationality have also been widely criticised. For example, it is alleged that they have wildly counter-intuitive implications, such as that there is nothing irrational, say, about having a complete indifference to what happens on future Tuesdays (Parfit, 1987: 134-4).[19] It has also been suggested that no way of accounting for the normativity of instrumental rationality can be given on a Humean account of practical rationality (Korsgaard, 1997: esp. §2). Finally, it has been alleged that a decision theoretic approach to practical rationality does not place sufficient constraints on the attitudes that agents can hold for them to be interpretable as agents (Hurley, 1989: chs. 4-6).

This last criticism is particularly pertinent to my account of practical reasons. I discuss the suggestion that there must be substantive constraints on preference, if agents are to be interpretable, in chapter 4, as well as the implications of this claim in terms of my account of practical reasons' reductive pretensions. To preempt a little, I accept that there must be substantive constraints on agents' desires for them to be interpretable, but I deny that these constraints must involve the existence of anything irreducibly normative. This leaves it open, so far as the present worry is concerned, for me to claim that practical rationality is not about discovering what one ought to desire. Rather, practical rationality consists purely in the pursuit of maximal desire satisfaction, where there happen to be constraints on the kinds of desires that agents can have.

---

[18] Notable examples include Brandt (1972); Gauthier (1986); Goldman (2010); Mackie (1977); Nozick (1993).

[19] For a response to arguments involving cases of this kind see Street (2009).

Aside from my debt to Davidson's ideas on interpretation, action and rationality, I also make use of ideas about normativity which derive from the agent-constitutive approaches of Goldman (2010: 66-82 & 181-5), Korsgaard (1996: esp. §1.4.4 and §3.3.1; 2008, ch.3; 2009, esp. chs. 2 & 4-7) and Velleman (2000: chs. 1 and 8). However, as the chapter is largely an exercise in theory development, I do not attempt to directly engage with or resolve any existing debates in the literature on practical reasons, including those about normativity. The purpose of the chapter is to offer a reasonably determinate idea of how interpretivism about practical reasons might look.

An interpretivist account of practical reasons of the kind that I propose faces many issues. These include:

(i) Whether interpretability is a suitable notion on which to ground the concept of a practical reason.

(ii) Whether rational functioning is a suitable proxy for functioning mentally in a way which supports interpretation.

(iii) Whether the notion of rational functioning can be cashed out without invoking some interpretation-independent account of normativity.

(iv) Whether a plausible account of practical reasons' normativity can be given on an interpretivist approach.

(v) Whether a plausible account of standard reasons concepts can be given in terms of the concept of rational functioning.

These are significant issues. In outlining the account of practical reasons I attempt to go some way towards addressing them. Hopefully I go far enough to convince the reader that the proffered account of practical reasons is at least plausible.

Having outlined the proposed account of practical reasons in this chapter, I move on in chapter 3 to consider in detail what rational functioning involves. This involves discussing the aspects of rationality which determine the actions which rationally follow from an

agent's propositional attitudes. Whereas chapter 2 outlines the structure of the account of reasons—how the concept of a reason is to be cashed out in terms of rational functioning—chapter 3 fills in some of the content of what rational functioning involves. The aim of chapter 3 is to give enough content to the notion of rationality to show that it can support the account of reasons given in this chapter.

In the remainder of this chapter I explain the motivation for my interpretivist account of practical reasons, and set out the structure of the account, attempting to address some of the issues listed above as I do so.

The account that I offer is taken to be ontologically reductive: to be a practical reason is to be a set of attitudes from which some action follows (in a sense to be specified), given the constraints of interpretable functioning. However, I do not claim that the *concept* of a practical reason can be reduced in this way. One way of developing my position would be opt for both an ontological and a conceptual reduction of practical reasons. I refrain from pursuing this approach, as conceptual reduction is not a necessary feature of naturalism about practical reasons. The account is also mind-independently normative, in that practical reasons' normativity (discussed in section 5) is taken to be independent of agents' existing attitudes. Finally, the account is hypotheticalist, in the sense that the specific practical reasons that agents have are dependent on their existing motivations.

2. Why Interpretability?

The interpretability of agents features strongly in our practical lives. We use our ability to interpret others to make predictions about how they will behave, which (plausibly) helps us to coordinate our actions with theirs. The interpretation of others is also central to our evaluations of their character and their actions, which have an important influence on how we respond to and relate to them. We use our ability to interpret ourselves to help us to make decisions about what to do; understanding our motivations and relating these to different courses of action helps us in deciding how to act.

The importance of interpretability to practical life suggests that it may be a good place to start in developing a reductive account of practical reasons. However, justifying the development of an interpretivist approach to practical reasons requires more than the mere observation that interpretability is central to practical life, just as practical reasons are. It also needs to be shown that interpretability is a suitable candidate for *explaining* practical reasons; that the properties and relations bound up with practical reasons are apt to be explained in terms of the properties and relations which ground interpretability.

Interpretability is a result of our tendency to function in certain ways. Such functioning involves, among other things, acting in ways which are reliably correlated with the attitudes that we hold. If our behaviour were not reliably correlated with our attitudes, then we would be unable to attribute specific attitudes to each other (and to ourselves) in the light of action. The connections between attitudes and actions would be too unpredictable. This is borne out by the theories of radical translation proposed by Quine (1960: ch.2), and radical interpretation proposed by Davidson, (1973; 1974a; 1975). Both theories depend upon the existence of regular patterns between speakers' utterances and their intended referents, such that meaning can be attributed. As with speech acts and linguistic meaning, so with action and the meaningful attribution of propositional attitudes. Without regular patterns between action and attitude, no specific attitudes can be attributed to agents.

Given this requirement for attitudes to correlate reliably with actions, we can characterise interpretability as involving certain actions being paired-up with certain attitudes. In order to be interpretable, we must act in certain ways given the attitudes that we hold (or *vice versa*). For example, if I desire to avoid getting wet (and believe that it is raining) then, other attitudes being equal, I must act so as to avoid getting rained on if I am to be interpretable as having these attitudes. To the extent that I do not act to avoid getting rained on I cannot be interpreted as having a desire to avoid getting wet and/or a belief that it is raining (*ceteris paribus*). Thus my desire and my belief match up with certain actions that I am to perform/avoid performing, if I am to be interpretable as having those attitudes. More generally, if I am to be interpretable as having any attitudes at all, I must tend to perform

the actions which pair up with those attitudes, given the constraints of interpretable functioning.

One feature of practical reasons is that they single out certain actions for us to perform. Something can only be a reason for action if an action follows from it, in some sense.[20] This introduces a significant parallel between interpretable functioning and practical reasons. In the case of interpretability, certain actions are paired up with certain attitudes that we hold, given the constraints of interpretable functioning; in the case of practical reasons, certain actions are paired up with certain reasons that we have, given some relation which pairs reasons with actions.

With interpretability, it is a set of constraints on interpretable functioning which pairs actions with attitudes. In the case of reasons, there are two empty place-holders. Reasons (as yet unknown entities) are paired with actions by some unknown relation. My suggestion in this thesis is that these empty places can be filled in by mapping the concept of practical reasons on to an account of interpretable functioning. Thus, reasons are attitudes/sets of attitudes which are paired with certain actions by the constraints of interpretable functioning.

This view is motivated by the central role of interpretability in cementing our practical lives, combined with the suggestion that certain actions follow from our practical reasons in much the same way that certain actions follow from the attitudes that we hold, given what is necessary to be interpretable. The relation between reasons, agents and actions in the case of practical reasons seems sufficiently similar to that between attitudes, agents and actions in the case of interpretable functioning to make an account of reasons framed in terms of interpretability an appealing prospect, especially given the practical import of both reasons and interpretability.

[20] Part of the job for an account of practical reasons is to explain in just what sense an action follows from a practical reason.

3. Why Rationality?

Despite the rough statement of the view just given, I do not intend to cash out the concept of a reason for action directly in terms of the concept of interpretability. Rather, I intend to invoke the concept of rational functioning as a proxy for interpretable mental functioning (functioning psychologically in a way which supports interpretation). This raises several questions. Most immediately: (a) why choose rational functioning as a proxy for interpretable mental functioning? (b) Is it a good proxy?

One reason for identifying reasons in terms of rationality, rather than in terms of interpretability *per se*, is that interpretability involves both rational and causal factors, where the latter seem irrelevant to the attribution of reasons. For an agent to be interpretable her attitudes must connect with each other in the right ways *and* her attitudes must connect causally to her actions in the right ways—she must generally succeed in performing the actions that she intends to perform. So there is both a rational and a causal component to interpretable agency. When a person fails to execute the actions that she intends to perform, she cannot (easily) be interpreted as intending to perform those actions. This is not because she fails to act for a reason. Rather, it is because she fails to execute the actions that she has (or takes herself to have) a reason to perform. Her body lets her down, so to speak.

An example here might be a composer and pianist (Isabella) who sometimes fails to play the music that she has scored (she has bouts of nervousness on big occasions). Suppose that Isabella is a compositional genius but that, on a particularly big occasion, she is also completely ham-fisted. On any such occasion it is impossible to attribute musical genius to her. This is not because, on such occasions, Isabella's compositions are any worse than usual (they are just as inspired when she plays them badly as when she plays them well). Rather, it is because her compositional intentions are inaccessible, given her clumsy play. To interpret one of Isabella's performances as indicating her compositional genius requires both that she intends to play the music that she has scored and that she succeeds in enacting this

intention. Given her failure to play her pieces at all correctly on particularly big occasions, Isabella's compositions cannot be properly appreciated at these times, and a gift for compositional genius cannot be attributed to her.

Interpretability requires both that agents' mental processes are (sufficiently) rational and that their intentions are causally connected to their actions in the right ways. The second, causal element of interpretation does not seem to be relevant to an account practical reasons. We would not want to claim that Isabella's reasons to play A# are in any way diminished by her tendency to hit G# when she is nervous. Expressing the musical purpose behind her composition makes it rational for Isabella to play the notes that she has scored, even if she finds this difficult on occasion. Thus what it is rational for Isabella to play is independent of the causal connections which her ability to do so relies upon (at least, within the context of her being about to play at some particular performance).[21] Causal connections between intentions and actions can misfire without affecting the reasons that one has.

For this reason, I focus my attention on rationality as a mode of mental functioning which supports interpretation.[22] For agents to be interpretable, it is necessary that their attitudes

---

[21] Isabella's limitations do, perhaps, suggest that it might be better for her to avoid choosing her most difficult pieces to perform on big occasions. However, what is important is that her limitations do not suggest in any way that she should play the pieces that she *has* chosen for some occasion badly.

[22] A further reason for focussing on rationality rather than interpretability *per se* is that it allows a distinction (of sorts) to be drawn between acting for a good reason and not (i.e. between having a normative and a merely motivating reason). This distinction is commonly drawn in the literature on practical reasons (see, for example, Dancy, 2000: ch. 1; Smith, 1994: 94-8).

On my view, an agent acts for a good reason if she acts on the basis of a belief-desire pair from which an action rationally follows, where she does not have any further beliefs which (together) rationally undermine her belief that the action in question will satisfy the relevant desire. If an action is performed on the basis of a reason which involves a belief that is rationally undermined by some further belief/set of beliefs that the agent holds, then the agent's reason for acting is not a good one.

relate to each other in certain ways (on the whole). This includes forming certain intentions to act, given certain beliefs and desires. It is the rational relations which pair certain beliefs and desires with certain intentions to act (and, by extension, with certain actions—assuming that the right causal processes *are* in play) which seem apt to explain reasons for action. Thus reasons for action are, on my view, attitudes/sets of attitudes from which the intention to perform some action rationally follows, and from which the relevant action itself follows if 'normal' causal processes are in play.[23]

I hope that my focus on rationality is relatively uncontroversial. Many philosophers have accepted a strong explanatory connection between practical reasons and rationality. For instance, a popular contemporary trend has been to understand rationality as the capacity to respond appropriately to reasons (Parfit, 1997: 99; Scanlon, 1998: ch.1). In the other direction, Kantians (and other rationalists) have long insisted that agents' reasons for action are to be understood in terms of what it is rational to do. This project has both normative and metaethical sides, with philosophers such as Smith and Korsgaard claiming metaethically that reasons are to be explained in terms of rationality (Korsgaard, 2008: 2-5; Smith, 1994, ch. 5). My account of reasons also maintains that practical reasons are to be

---

For example, if I am aware that the odds of winning at roulette are terrible but neverthless continue to play it because I want to get rich (and, presumably, believe that I can, or even will, get rich by playing it) then I act for a bad reason.

[23] One might wonder whether I have changed the subject here, by turning reasons for action into reasons to intend to act. I do not think that I have. The beliefs/desires in question are rationally connected to intentions to act, for sure. But they are also rationally connected to eventual actions, in the sense that actions are the subject of agents' intentions. If, given certain beliefs and desires, it is rational for an agent to intend φ, at a time when she is able to execute this intention then, derivatively, it is also rational for her to φ at that time. An action derives its rational status from the rational status of the agent's intention to perform that action at the appropriate time. That is the proposed view, at least. This rational status withstands any causal deviations, as it is the intended action, rather than any rogue action which the agent ends up performing, which is the subject of a rational intention. To the extent that an agent has a rational intention that she fails to enact, her action is a failed one rather than an irrational one.

explained in terms of rationality. However, my conception of practical rationality departs strongly from the Kantian picture.

Even if there is nothing controversial about accounting for practical reasons in terms of rationality in general, one might wonder whether such an approach is appropriate, given the general interpretivist framework that I invoke. If reasons are attitudes from which certain intentions to act follow, given the constraints of interpretable mental functioning, we might ask whether rationality really is the correct adumbration of the kind of mental functioning required to support interpretability. In other words, is rational functioning a good proxy for functioning mentally in a way which supports interpretation?

Perhaps the biggest theoretical contribution to explaining what interpretability requires of us psychologically has been made by Davidson. One project of Davidson's was to explain what is psychologically necessary for it to be possible to attribute specific beliefs and desires to agents and specific meanings to their words, given the observable behavioural evidence (Davidson, 1980; 1995). In carrying out this project, Davidson made foundational use of many aspects of rationality, including those mapped by decision theory, logic and epistemology. These aspects are drawn together, among other places, in the paper 'A Unified Theory of Thought, Meaning, and Action', in which Davidson discusses what a theory which allowed us to interpret agents from scratch, given only behavioural observations and an ability to establish when they prefer that a sentence (an uninterpreted sentence, that is) is true, would be like (Davidson, 1980). According to Davidson, the assumptions that agents are logical, that they make choices in accordance with the axioms of decision theory, and that they form beliefs which they take to have the highest degree of evidential support, are required. In short, agents must be largely rational. This is because rationality provides a formal structure through which we can filter behavioural evidence to extract interpretations. Without them conforming to such a structure (and such a precise one at that) we would be unable to begin the process of attributing specific propositional attitudes to completely uninterpreted agents, given their utterances and behaviour.

Davidson's theory of interpretability in the abstract is taken to illuminate what is functionally necessary for agents to be interpretable in practice, when agents do not proceed in the manner supposed by the theory. Agents do not generally begin interpreting each other from scratch, but the extreme case shows the limits within which interpretable agency is situated.

There are, of course, many questions of detail which arise in relation to Davidson's theory. These include questions over: the degree to which an agent must be rational in order to be interpretable; whether the supposed requirements of radical interpretation in the abstract are a good guide to the requirements of interpretability in practice; to what extent rationality involves holding attitudes with certain contents as well as attitudes which are formally related to each other in certain ways. I address the third issue in chapter 4. The first two questions are not discussed in detail. This is not because they are unimportant, or because they are easy to address. Rather, they are hard questions which it is beyond the capacity of this thesis to address.

This does not mean that the focus on rationality as a proxy for interpretable mental functioning is unjustified. Rationality seems to be the only theory that we have to chart the kinds of patterns which must exist between agents' attitudes if such attitudes are to be attributable. Without an assumption of general rationality, we lack a starting point in the practice of interpretation (answering questions such as 'why did she do that?' almost always begins by searching for goals under which an action can be seen to be rational). Thus questions over the exact role of rationality in an abstract theory of interpretability, and of the match between such a theory and the requirements of interpretability in practice, appear to be questions of detail rather than questions of substance. Rationality, it can be assumed, is of central importance to interpretable mental functioning.

A further worry about my particular explanation of practical reasons in terms of rationality is that I propose to account for agents' practical reasons in terms of what it is rational for them to intend, given their existing beliefs and desires. This proposal conflicts with an established distinction which is often drawn between what is rational (from an agent's own subjective

standpoint) and what that agent has a reason to do. For example, Williams (1981) introduces a case in which an agent believes that a glass contains gin when it actually contains petrol. Williams uses this case to show that what is rational from the agent's subjective standpoint (to mix the contents of the glass with tonic and drink it) is not the same as what she has a reason to do (to avoid drinking the contents of the glass). Based on cases of this kind, a distinction between acting rationally and acting for a reason (Kolodny, 2005) or between choosing rationally and choosing correctly (Wedgwood, 2003) can be drawn.

As things stand, my proposed account of practical reasons does not respect this distinction. This appears to be a problem for the account. One response to this problem is to distinguish between two senses of the term 'reason' – a subjective sense and an objective sense, where subjective reasons are dependent on an agent's actual epistemic position and objective reasons are independent of an agent's actual epistemic position. Thus one might claim that, in a subjective sense, the agent in Williams' example does have a reason to drink the contents of the glass, while in an objective sense she does not. One can then treat this thesis as concerned with explaining agents' subjective reasons for action (on my account, reasons which comprise beliefs and desires that an agent has at the time for action, where these attitudes confer rationality on the performance of certain actions). These are the reasons which an agent is capable of being motivated by and are, therefore, of significant philosophical concern.

Controversially, I am dubious about the existence of a further category of what I have termed objective practical reasons. True, we often talk about an agent's having a reason to do something which it is not rational for her to do, given her existing epistemic standpoint (as Williams' example illustrates). However, I am prepared to regard this kind of talk as, strictly speaking, erroneous. Although it might be true that the relevant agent *would* have a reason not to drink the contents of the glass if she *was* aware that it contained petrol, I regard this counterfactual claim as the most that can be said of the situation. Given our own awareness of what is in the glass, we attribute a reason to the agent that (I think) she does not have, but which she would have if her beliefs were more akin to ours. This is because I

think that agents must, at least in principle, be capable of being motivated by the reasons that they have in any given situation; a reason which depends upon on a belief/piece of information that an agent does not have cannot motivate her in that situation.

However, many (if not most) would disagree with me here. Such opponents might wonder whether my account of practical reasons can also explain agents' objective reasons for action. One way of extending the account would be to follow Williams by invoking the notion of an ideally rational agent (an agent who has no false beliefs, all relevant true beliefs and who deliberates correctly—Williams, 1981). From this, my proposal would be that an objective reason is a set of attitudes that an agent would have in the obtaining circumstances if she was suitably idealised, where intending to perform some action is rational, given this set of attitudes. I do not consider the merits of this proposal further as the purpose of this thesis is not to explore such an account; its purpose is to consider how subjective reasons (the reasons which can actually motivate an agent to act, given her attitudes at the time for action) can be accounted for.

A more immediate concern is the suggestion that rationality itself is an independently normative notion, such that it cannot be invoked in providing an ontologically reductive account of what practical reasons are in terms of interpretability.

4. Is Rationality Independently Normative?

Davidson is explicit about the fact that his theory of thought, meaning and action is normative. For instance, he remarks that in the case of attributing beliefs, 'the guiding principles must derive..., as in the cases of decision theory or the theory of truth, from normative considerations' (Davidson, 1980: 156). To attribute beliefs, desires or meanings we must apply certain norms (or evaluative standards) to agents – norms by reference to which we can extract appropriate interpretations of their behaviour. These include norms of rationality, as well as norms of truth in the case of meaning.

One way of reading Davidson's use of rational norms, advanced by Timothy Schroeder, is to treat them as categorising the kinds of relations which must, largely, hold between agents' propositional attitudes in order for them to be interpretable (Schroeder, 2003). As mentioned in chapter one, Schroeder distinguishes between two features of any norm: the categorisation scheme of that norm and the norm's force-maker. The categorisation scheme of a norm is the way in which that norm '[divides] up domains into mutually exclusive and jointly exhaustive categories'(ibid: 2). For instance, norms of etiquette divide actions into three categories: polite, impolite, and neither polite nor impolite.

According to Schroeder, a categorisation scheme is not itself normative. There are many ways of categorising the world, not all of which involve norms. For a genuine norm to exist there must be an evaluative ordering which applies to the categories which come under some categorisation scheme. In Schroeder's terms, the categorisation scheme needs a 'force-maker' which 'puts the *normative force* into the categories' (ibid: 3). For instance, social convention acts as a force-maker by which polite behaviour is ranked above impolite behaviour. This gives the norms of etiquette their normative force.

The normative force-maker which turns a categorisation scheme into a normative ordering can, according to Schroeder, come from various sources including functions bestowed on objects by our intentions, social pressure, and natural selection. He also accepts that there may be categorisation schemes which are implicitly normative, such as the virtues. However, for Schroeder's purposes, the important thing is that the norms of rationality can be treated as having a categorisation scheme which can be applied independently of that scheme's normative force-maker.

Thus Schroeder claims that the rational categories which Davidson invokes can be applied independently of their having normative significance. To illustrate, he extracts two features of rationality which feature in Davidson's theory of mind: *consistency* of beliefs and *coherence* between beliefs, desires and actions. In both cases, it is argued that the relevant categorisation scheme can be applied independently of the presence of normative force. Schroeder gives the example that the first sentence of *Word and Object* may be found to be

consistent with the last sentence of *On the Plurality of Worlds*, even though there is no requirement for it to be so. In this case, consistency can be attributed between sentences without reference to the idea of normative force. Likewise, my actions may be found to be coherent with your beliefs and desires, even though 'there is no normative failure if I act in a manner you could not rationalise' (ibid: 5). Coherence, too, can be attributed without reference to normative force. This implies that both consistency of beliefs and coherence between beliefs, desires and actions are categories which are not implicitly normative.

Having argued that the categorisation schemes for the rational norms which Davidson invokes can be applied independently of their having normative force, Schroeder goes on to claim that Davidson's own application of these categories is indifferent to their normative force-makers. That is, he claims that Davidson's use of the rational norms of consistency and coherence are simply ways of identifying how agents must function in order to be interpretable, where the fact that these norms have normative force is insignificant so far as his theory is concerned. What is important is that agents' attitudes conform to certain patterns, not that these patterns have normative import. For instance, Schroeder cites Davidson's paper 'Mental Events', in which he writes:

> [W]e cannot intelligibly attribute any propositional attitude to an agent except within the framework of a viable theory of his beliefs, desires, intentions, and decisions...[W]e make sense of particular beliefs only as they cohere with other beliefs, with preferences, with intentions, hopes, fears, expectations, and the rest (cited in ibid: 7).

In this passage Davidson makes reference to coherence as a framework which must be applied in attributing mental states to agents. Schroeder's suggestion, based on this and other similar passages, is that Davidson does not make use of coherence *qua* rational norm, but simply as a structural constraint on the attribution of mental states. This leads Schroeder to conclude that Davidson's theory of mind is indifferent to the existence of a normative force-maker for the norms of rationality. What is important to Davidson's theory, on Schroeder's reading, is that the mind must function in accordance with certain categories

(including consistency and coherence), where these can be non-normatively ascribed. As such, Schroeder maintains that Davidson's theory of mind is, contra Davidson, non-normative.

This suggestion is controversial, especially given Davidson's insistence that his theory is normative. However, I am inclined to think that it is a plausible reading of Davidson's view. One reason for this is that one of Davidson's aims is to offer a unified theory of thought, meaning and action which has the rigour of any good scientific theory. He is aware that the normative dimension of his theory may be taken as a threat to this ambition. His response to this threat is as follows:

> The entire theory is built on norms of rationality; it is these norms that suggested the theory and give it the structure it has. But this much is built into the formal, axiomatizable, parts of decision theory and truth theory, and they are as precise and clear as any formal theory of physics (Davidson, 1995: 129-30).

Rationality, then, is something which Davidson thinks can be clearly and precisely specified. It seems to me that the role that rationality plays within Davidson's theory will be exhausted by the formally specified constraints on agents intentional states that it provides. This is because it seems implausible to think that a theory, such as Davidson's, which makes use of rationality as a set of systematic constraints on agents' attitudes could also depend upon the (supposed) fact that agents are rationally *obliged* to do certain things. That is, it is hard to see how the normativity of rationality could play any role in generating the interpretive outcomes of Davidson's proposed theory. Such outcomes, it seems, derive purely from the operation of the categorisation scheme for interpretation that formalised decision theory and truth theory provides.

This leads me to think that Davidson's use of normative terms does not commit him to the idea that the normative force of rationality plays a fundamental role in grounding interpretability. Rather, it seems that his use of notions like rationality and rational norm serve to pick out certain non-normatively specifiable categories which it is generally

necessary that we fall under, if we are to be interpretable. Such categories may be normative in the sense that, for any interpretable creature, these categories have a kind of unavoidable significance. But normativity of that kind can be seen as a explainable in terms of the necessity of falling under certain categories when it comes to functioning interpretably, rather than as an independent constraint upon it.

It seems entirely plausible to me that the categories which comprise rational functioning have normative significance in just the sense that we, as interpreters and interpreted, are constitutively constrained to apply them and to conform to them. That is, the normativity of the categories which underpin Davidson's interpretive scheme seems to me to be entirely explainable in terms of these categories' constitutive role with respect to interpretable functioning. Thus rather than seeing interpretable functioning as dependent on the normativity of certain categories, we might see the normativity of those categories as explainable in terms of their constitutive role with respect to interpretable functioning.[24]

I am not committed to the claim that Davidson's theory of mind, or of radical interpretation, is non-normative in the sense that it has no normative outcomes (such outcomes are exactly what I wish to insist upon). Rather, I am committed to the claim that nothing independently normative is involved in specification of his theory of interpretable functioning itself. I take Davidson's use of rationality to be consistent with this claim, given that the normative significance of rationality seems to be explainable in terms of its unavoidable significance to any interpretable creature, rather than as a prior feature which plays an explanatory role within Davidson's theory of interpretation.

I return to objections from irreducible normativity in chapter 4, where I discuss the objections that meaning and preference are subject to irreducibly normative constraints. For now, I hope to have shown that it is at least plausible to treat rationality as non-normative, insofar as it plays an explanatory role within Davidson's theory of interpretation. From now on I shall assume that the norms of decision theory, logic, and epistemology that Davidson

[24] I pursue this suggestion in section 5, below.

invokes supply the required attitudinal patterns in terms of which interpretation must be conducted, without reference to their normative status. That is, I assume that the charge that rationality imports unwanted, independent normativity into my account of practical reasons fails.

5. How is the Normativity of Practical Reasons to be Explained?

So far I have done the following: explained the general motivation for framing an account of practical reasons in terms of interpretability; introduced rationality as a proxy for interpretable mental functioning; attempted to show that using rationality in this way does not involve a commitment to independent and prior normativity. Before setting out the specific account of practical reasons that I wish to defend, there is one more issue to address. This is to indicate how, on my proposed account, the normativity of practical reasons is to be accounted for.

The approach that I adopt here is similar to the agent-constitutive approaches adopted by Korsgaard, Velleman and Goldman. In different ways, each of these attempt to ground normativity in what is constitutive of action. Each of them identifies a (different) constitutive aim of action, and then explain practical reasons' normativity in terms of this. I briefly gloss each account.

Korsgaard (2009: 25) claims that action's constitutive aim is the realisation of the self. In acting, we determine what kind of person we are. In order to determine the kind of person that we are, through the actions that we perform, we must be in a position to make choices about what to do (the process of choosing what to do is the process of deciding who to be). To be able to make choices involves adhering to certain constitutive constraints on choice (rational norms); one is committed to these by the very nature of what it is to make choices. These norms, for Korsgaard, are distinctly Kantian: choice involves a commitment to Kant's categorical imperative of willing only that which any agent could rationally will (Korsgaard, 2009: 153-8). Since we are condemned to make choices (we cannot choose not to choose, for to do that would be to make a further choice, governed by the constitutive norms of

choosing) we are committed to adhere to constitutive constraints on choice, viz. the norms of rationality. The constitutive role of rationality in the making of choices is what explains why rationality, and the practical reasons associated with it, is normative.

For Velleman, the constitutive aim of action is not self-realisation. Rather, it is the aim of behaving in ways that we can make sense of; the aim of 'knowing what we are doing' (Velleman, 2004: 236). For reasons of obvious circularity, Velleman cannot claim that to act in ways that we can make sense of is just to act in accordance with our practical reasons (practical reasons are what is being explained). Rather, he claims that to act in ways that we can make sense of is to act in ways that we can explain theoretically in some way—to act in ways for which we can give a 'comprehensive' and 'integrated' explanation of what we are doing (ibid: 231). The objects that we cite in our sense-making explanations of our behaviour are our reasons for action.

Velleman develops this idea in a narrative sense. Hence: 'reasons for acting are the elements of a possible storyline along which to make up what we are going to do' (Velleman, 2000: 28). To act for a reason (i.e. to act on the basis of some comprehensive and integrated explanation of what we are doing) is to allow ourselves to be guided to act by the motivations which feature in our chosen narrative explanation of what we are doing. This allows us make up our 'personal history', rather than being 'obliged to discover it' (ibid: 29). The normativity of practical reasons stems from the role that they play in our explanations of our own actions; they are features of a kind that we must pick out in constructing a rationale for an action, which is something that we must do if we are to act at all.

Adopting a more straightforward approach than Korsgaard or Velleman, Goldman claims that the constitutive aim of action is desire satisfaction. To act is to attempt to satisfy a desire. Successful actions are those which achieve this aim. Reasons are facts which would have a bearing on how any rational (i.e. coherent and relevantly informed) agent would go about satisfying some desire (Goldman, 2010: 34). Reasons are normative because attempting to satisfy our desires necessarily involves paying attention to facts which it is rational to treat as relevant to their satisfaction. To ignore our practical reasons, *qua* facts a

rational agent would treat as relevant to the satisfaction of our desires, is to fail to attempt to satisfy our desires at all.

Each of these three proposals is attractive in different ways. Moreover, the underlying thought that normativity is grounded in what is constitutive of action is appealing for the naturalist, who wishes to explain normativity in terms which are compatible with science. One supposed feature of each of the above accounts is that there are no *sui generis* normative entities or relations posited in their explanations of practical reasons' normativity.[25] Rather, normativity is explained in terms of there being some (supposed) aim at which action is constitutively directed, where agents must adhere to certain constraints on the pursuit of this aim if they are to count as acting at all.

However, one major criticism of the agent-constitutive approach to normativity involves asking why one ought to be an agent in the first place? Thus Enoch (2006) has suggested that, unless agency is a normatively significant category, there is no normative force behind the constraints that apply to being an agent. The agent-constitutive strategies canvassed above do not, in any way, show that agency is a normatively significant category. To show this would involve moving outside of a theory of what agency constitutively requires to give an explanation of why being an agent has normative significance, such that realising the constitutive aims of agency is somehow imperative. Thus Enoch concludes that the constitutive requirements of agency offer no ultimate explanation of normativity.

Enoch takes this point to hold, even if acting is something that we cannot avoid doing. The issue is not whether we are in a position to choose to be agents or not, but whether being an agent is normatively significant. Even if choosing commits us to choosing in accordance

[25] Certain apparently normative notions, particularly rationality, feature in the accounts as described. However, rationality is used as a summary notion here. For instance, in Korsgaard's account it is a placeholder for the categorical imperative as a formal constraint on choice; in Goldman's account rationality is a placeholder for the idea of following certain deliberative procedures and having a certain degree of information.

with certain constitutive requirements of choice, and even if we are condemned to make choices, the constitutive requirements of choice are not thereby normative. They are just constraints on something that we cannot help doing; that does not show why this thing matters, or that the constraints that apply to it matter either.

Although Enoch's criticism of the agent-constitutive approach to normativity has force against some interpretations of the view, I do not think it applies to all such interpretations. Specifically, if what is constitutive of agency is taken to ground normativity (i.e. to explain it without providing a reduction of it), I think that Enoch's attack is successful. For it is unclear why a constitutive requirement of an unavoidable practice would, by itself, generate any kind of normative significance around that practice. That, it seems, is normative bootstrapping.

However, if the agent-constitutive approach is taken to be a reductive explanation of normativity then things are different. In this case, there is no explanatory debt for the agent-constitutivist to pay. They do not need to explain why constitutive constraints on agency give rise to normativity because they have simply identified normativity in terms of the existence of these constraints.

The question for the normative reductivist here is not why the constitutive constraints of agency are normative as, for her, there is nothing more to normativity than the existence of these constraints. Rather, the question is whether the proposed reduction of normativity is a good one. I am hopeful that a reduction of normativity of this general sort is along the right lines. It seems plausible that our being functionally bound by certain constraints is what those constraints' normative force consists in. However, I do not think that it is constitutive constraints on agency which provide the right reduction base. Rather, I think it is constitutive constraints on interpretable functioning.

I shall now give a brief explanation of my preferred approach to normativity. I will not attempt to offer a fully-fledged account of it, as this would involve a detailed discussion and defence of certain claims which go beyond the scope of this thesis. Also, I will not try to give

any knock-down arguments against alternative approaches to normativity. What I will try to do is to indicate: what an account of normativity framed in terms of the constitutive requirements of interpretability would look like; why I am in favour of such an account; how it accommodates some of the strengths of the agent-constitutive approaches canvassed above.

The starting point for an interpretability based account of normativity is to assert the primacy of interpretation over intentionality. Metaphysically, this is the claim that intentional states and categories exist only with respect to the practice of interpretation: to have intentional states, or to fall under an intentional category, just is to be interpretable as such. Theoretically, this is the claim that intentional concepts can only be applied and understood with regard to their role in interpretive practices. Adopting one or both of these claims leads to a form of interpretivism about the intentional.

This kind of interpretivist position is strongly associated with the work of Davidson (e.g. 1975). So far as Davidson is concerned, we cannot get any theoretical purchase on intentional notions without reference to interpretation. Davidson takes this theoretical result to have metaphysical implications. Hence:

> We have the idea of belief only from the role of belief in the interpretation of language, for as a private attitude it is not intelligible except as an adjustment to the public norm provided by language. It follows that a creature must be a member of a speech community if it is to have the concept of belief. And given the dependence of other attitudes on belief, we can say more generally that only a creature that can interpret speech can have the concept of a thought.
>
> Can a creature have a belief if it does not have the concept of belief? It seems to me it cannot...Someone cannot have a belief unless he understands the possibility of being mistaken, and this requires grasping the contrast between truth and error—true belief and false belief. But this contrast, I have argued, can emerge only in the context of interpretation...(Davidson, 1975: 170).

In this passage we can see that Davidson takes belief (and by association, other propositional attitudes) to be dependent on interpretation, both theoretically and metaphysically. One cannot have the concept of belief without reference to the role it plays in interpretation; one cannot have beliefs without having the concept of belief. Thus, for Davidson, interpretation takes both theoretical and metaphysical primacy over the holding of beliefs, and other propositional attitudes.

Actions, like propositional attitudes, are intentional. Roughly, actions are behaviours which are (appropriately) connected to certain intentional, mental states. Thus on a Davidsonian picture, where intentionality depends on interpretation, the concept of action can be understood only by reference to its role in interpretation. Further, to perform an action requires that one could be appropriately interpreted as such. This is important because it suggests that it may be better to regard constitutive constraints on interpretation as fundamental to normativity than constitutive constraints on action/agency (where these are taken to be interpretation-independent).

It is this thought which leads to an account of normativity framed in terms of the constitutive requirements of interpretability, rather than the constitutive requirements of action. If to act just is to be interpretable as such, there cannot be a constitutive aim of action which can be delineated independently of interpretation. The constitutive aim of action, if it has one, is an aim which must be defined by reference to the (constitutive) constraints on interpreting some event as an action. As such, if a constitutive-constraint approach to normativity is to be adopted, and if the primacy of interpretation over intentionality is right, this approach must ultimately be grounded in the constitutive constraints on interpretation rather than the constitutive constraints on action.

Interpretability requires, among other things, that a creature can be treated as functioning in certain ways. For instance, an interpretable creature must be amenable to treatment in terms of having certain motivations, where they are disposed to change the world in accordance with these. More generally, they must be amenable to treatment in terms of the norms of rationality. These requirements apply at a global rather than a local level. A

creature must generally function in accordance with the requirements of interpretation if it is to have intentional states; local digressions from this mode of functioning do not denude a creature of intentionality. Thus Davidson remarks:

> The possibility of (objective) inconsistency depends on nothing more than this, that an agent, a creature with propositional attitudes, must show much consistency in his thought and action, and in this sense *have* the fundamental values of rationality; yet he may depart from these, his own, norms (ibid: 197).

The norms of rationality must be followed, in general, by anything which has intentional states. Local departures from these norms can occur, and can be attributed to agents 'against a background of rationality' (ibid: 196). However, where digressions from the norms of rationality occur, the determinacy of interpretation is liable to be reduced. There is always scope to reinterpret incoherence in terms of unusual beliefs and/or motivations (Hurley, 1989: chs. 4 and 5). This too, can only occur up to a point (see my discussion in chapter 4, section 3). However, in cases of potential irrationality, it can never be clear exactly which attitudes to attribute to an agent. This suggests that the greater the digression from the norms of rationality, the less interpretable a creature is.

If the constitutive requirements of interpretability admit of a certain degree of transgression, how is normativity to be understood in terms of these? Well, even if an interpretable creature can occasionally diverge from the kind of functioning which is constitutive of her being interpretable, her nature as an interpretable creature involves a general commitment towards functioning in these ways. Avoiding intentional states and actions which go against the constitutive constraints of interpretability is the essence of an interpretable creature's functioning; it is something that any interpretable creature generally does, and which she must generally be inclined to do, given that she cannot be what she is except by her tendency to adhere to these constraints. Thus interpretable creatures can be seen as generally committed to having intentional states, and to performing intentional actions, which comport with the requirements of interpretable functioning (in particular, rationality).

This idea of a commitment to function in accordance with the constitutive constraints of interpretable functioning decomposes into two elements: necessity and inescapability (cf. Korsgaard, 2009: 32). First, it is a necessary condition of being an interpretable creature that one generally functions in certain ways. Second, for anything which is an interpretable creature, functioning in these ways is inescapable. Choosing not to be an interpretable creature would, to appropriate Korsgaard's point (ibid: 1), require that one adhere to the constitutive constraints of choice, which on my account are the constitutive constraints of interpretable functioning. Thus although one can choose to put oneself in a position not to be able function interpretably (e.g. by choosing to commit suicide), one cannot choose to do something uninterpretable—doing so is a form of practical contradiction, or self-defeat.

Given that adherence to the constitutive constraints of interpretable functioning is both a necessary feature of our functioning as we do, and that functioning in this way is an inescapable feature of our mode of existence, it seems reasonable to treat us as being committed to function in accordance with these constraints. My proposal is that normativity reduces to our being so committed.

Thus the proposed account of normativity can be summarised as follows: (i) a requirement is normative if it is a requirement on the intentional states and/or actions of interpretable creatures; (ii) requirements of interpretability are normative (only) for those who are in the domain of being interpreted with respect to them; (iii) specific failures, on behalf of an interpretable creature, to comport with the general requirements which apply to the intentional states of interpretable creatures are normative failings in that such failures conflict with that creature's constitutive commitment to adhere to those requirements (i.e. her commitment to generally adhere to those requirements as part of her being an interpretable creature at all). Thus normativity is, on the proposed view, reducible to the existence of a constitutive commitment, on behalf of any interpretable creature, to adhere to certain constraints on her intentional states, where these constraints apply in virtue of her being interpretable as having such states at all.

This view is open to many lines of objection, and would need a great deal more support than I can offer it here to overthrow competing views of normativity. In particular, the assumption of the metaphysical and theoretical primacy of interpretation over intentionality would require significant support. The idea that normativity can be reduced to one's being, by one's very nature, constitutively committed to function in accordance with certain norms would also require a thorough-going defence. Finally, the claim that it is one's nature as *an interpretable being* which provides the relevant norms, to which one is constitutively committed and in terms of which normativity can be reduced, would need to be defended.

Although I cannot defend each of these claims here, I think that they are not implausible. The first two have been endorsed elsewhere in the philosophical literature. For instance, the theoretical and metaphysical primacy of interpretation over intentionality has been proposed by Dennett (1987), as well as by Davidson. A reductive account of normativity in terms of the constitutive features of one's natural mode of functioning has been proposed by philosophers following Aristotle. These include Foot (2001) and Korsgaard (2009: ch2) who, despite her Kantian take on what rationality involves, makes particular use of Aristotle's idea of natural function in explaining her view of normativity as deriving from constitutive constraints on choice. Of course, the relevant concept of one's nature on neo-Aristotelian views is that of human agency, not that of being an interpretable creature. However, the general strategy is the same and, if I am right about the primacy of interpretation over intentionality, one's nature as a human agent will turn out to be explained, in large part, by one's nature as an interpretable creature. In any case, it is the specific suggestion that it is one's nature as an interpretable being that provides the constitutive constraints in terms of which normativity is to be reduced that is peculiar to my particular account of normativity.

This claim derives from three considerations. First, that it is the constitutive constraints on some (relevant) mode of functioning which are apt to account for normativity (from agent-constitutivism). Second, that normativity applies to creatures only insofar as what they think and do is intentional (background assumption). Third, that interpretability takes

metaphysical and theoretical primacy over intentionality (from Davidson). Together these suggest an interpretation-constitutive approach to normativity, of the kind proposed above.

I refrain from attempting to provide further support for the suggested approach to normativity here. However, I will try to show how it accommodates some of the insights of the agent-constitutive approaches canvassed at the start of this section.

First, consider Velleman's approach. According to Velleman, normativity is ultimately to be explained in terms of constraints on acting in a way that one can make sense of. I think that Velleman's focus on sense-making is correct. However, in treating the constitutive requirements of action as primary, rather than the constitutive requirements of interpretability, I think his focus is wrong.

As well as the incompatibility of Velleman's focus on the constitutive requirements of action with my belief in the theoretical and metaphysical primacy of interpretation over intentionality, there is a further issue with Velleman's approach. This is that there are many different ways of making sense of an action by way of offering a theoretical explanation of it (examples include narrative explanations, genetic explanations, sociological explanations and so on). Velleman seems to have no immediate way of restricting his focus to one kind of explanatory relation—narrative explanation. As Millgram (2009) points out, the notion of sense-making as theoretical explanation which Velleman offers is, as it stands, broad enough to allow a very wide range of possible theoretical explanations of action to count as providing reasons for an action.

There is, of course, scope for Velleman to set limits on the kinds of theoretical explanation which are relevant to practical reasons, although a principled reason for doing this must be provided. However, this issue can be avoided entirely by focusing on the constitutive requirements of interpretability, rather than of action. This introduces a very clear notion of making sense: functioning in a way which allows for the meaningful attribution of intentional states to agents in light of their actions. Thus if interpretability is the focus of normative explanation then the notion of sense-making in play is already restricted, in a

very convenient way, while the idea of making sense retains a central role (as Velleman seems to rightly afford it) in the explanation of normativity.

If Velleman is right to focus on sense-making, my intuition is that Goldman is right about the constitutive aim of action being desire satisfaction. However, Goldman seems wrong to think that desire satisfaction's being the constitutive aim of action is what explains normativity. Supposing, as I do, that interpretation is primary to intentionality, then desire satisfaction is the constitutive aim of action only in the sense that (assuming a Davidsonian view of rationality) it is an aim that must be ascribed to agents in attributing the performance of intentional actions to them. Accepting my account of normativity allows for agreement with Goldman that the constitutive aim of action is desire satisfaction, while disagreeing that this is what explains practical reasons' normativity. What explains normativity is the existence of constitutive constraints on functioning in a way which can be interpreted (and, therefore, on functioning intentionally at all), where these constraints include attempting to satisfy one's desires through one's actions.

Finally, we can agree with Korsgaard that the norms of rationality are constitutive norms of choice, in that choice is an intentional notion. However, the constitutive norms of choosing are not, on my view, normatively significant because of their role with respect to the aim of self-realisation. Thus I disagree with Korsgaard's focus on self-constitution, although I can agree with some of her claims about the unavoidability of choice for agents, and the role of rationality (if not her view of its content) with regards to making choices.

Before setting out my account of what practical reasons are, I will make a few concluding remarks about my explanation of practical reasons' normativity. Specifically, on the proposed view, the normativity of practical reasons is mind-independent. This is because the requirements of interpretable functioning do not depend on the contents of any agent's existing propositional attitudes. These requirements provide a categorisation scheme according to which agents and their actions can be classified as either interpretable or uninterpretable. The normative force of these categories is taken to reduce to their being the categories which separate interpretability from uninterpretability, where functioning

interpretably is something that any creature which has intentional states is, by their nature, committed to do.

Perhaps the constitutive requirements of interpretability do not provide a good reduction base for normativity. I will not argue further for the claim that they do here. My purpose has been to make the suggestion, and to indicate why I think it is a promising one. The fundamental role that interpretability plays in grounding intentional action seems, I have suggested, to make it a better candidate for a reduction of normativity than the requirements of action alone. However, if action can be characterised independently of interpretation, perhaps the normativity of practical reasons can be explained in terms of its constitutive criteria instead. Nevertheless, I prefer to treat interpretation as primary.

6. What are Practical Reasons?

I now turn to the task of outlining my proposed account of what practical reasons are. My approach will be to run through a series of proposals concerning what reasons are, and to revise these in the light of counter-examples. The aim is to arrive at a firm statement of the proposed view, having shown why several alternative formulations are lacking.

The basic idea is that reasons are attitudes or sets of attitudes such that certain intentions to act follow, given the constraints of interpretable mental functioning (i.e. given rationality). According to the Davidsonian model of rationality on which I am operating, both a desire and a belief are required to rationalise an action. Therefore, I assume from the outset that the set of attitudes from which an intention rationally follows must contain a desire and at least one means-end belief about how to satisfy that desire.

A starting point for formalising the idea that reasons are attitudes from which certain intentions to act rationally follow is:

Reason 1: A practical reason is a set of attitudes, x, (containing a desire and at least one means-end belief) held by some agent, a, which entails that having

> an intention to perform some action, *p*, at a time when she is able to perform that action, is necessary *for a to be* rational, with respect to *x*.

On this formulation, an intention follows from a set of attitudes in the sense that having that intention is *necessary* to be rational, with respect to those attitudes.[26]

Reason 1 seems like a good first stab at characterising practical reasons. To adapt an example of Mark Schroeder's (2007: ch.6), one might desire to get to Mars and believe that in order to get to Mars one must board a Mars-bound spaceship. In this case it is necessary to be rational, given (only) this belief-desire pairing, to form an intention to board a spaceship headed for Mars (assuming that such things exist). Thus, according to Reason 1, a desire to go to Mars, together with a belief that to get to Mars one must board a Mars-bound spaceship, is a reason to board such a ship.

However, Reason 1 is open to an abundance of counter-examples. There are many actions that we have reasons to perform but which it is not necessary to form an intention to perform in order to be rational with respect to some relevant set of attitudes. For example, I might desire to have some cake and believe that two equally good ways of getting some cake are to go the local cafe or to go to the local supermarket. In this case, forming an intention to go to the cafe is not necessary to be rational, given my belief-desire pair, and

---

[26] I bring in the idea of an intention's being rational at a time when one is able to perform the relevant action to rule out cases, such as Kavka's toxin puzzle, in which there can be sets of attitudes which make it rational to have some intention but not at a time when one could perform the relevant action (Kavka, 1983). The toxin puzzle, which is discussed in detail in chapter 3, involves one's being offered a large reward (which will be received tomorrow morning), for intending, at midnight tonight, to drink a vial of sickness inducing toxin tomorrow afternoon. In this case it is (arguably) rational to intend, at midnight tonight, to drink the toxin tomorrow afternoon. But it is not rational to intend, tomorrow afternoon, to drink the toxin. Thus this is a case in which one's attitudes make it rational for one to intend to perform some action, only not at a time when one can actually perform it. As such, one does not have a reason to drink the toxin, even though there is *a* time at which intending to drink it is rational.

neither is forming an intention to go to the supermarket. Rather, forming either intention is sufficient to be rational, given my belief-desire pair.

Although, in this case, forming either an intention to go to the cafe or an intention to go to the supermarket is sufficient to be rational, and neither is necessary, I can nevertheless be accurately described as having reasons to go both to the cafe (this is a way for me to get some cake) and to the supermarket (this is an equally good way for me to get some cake). So this is a case in which I have reasons to perform two different actions, even though intending to perform each of these actions is only sufficient to be rational, given the relevant set of attitudes, and not necessary.

A second suggestion is:

> Reason 2: A practical reason is a set of attitudes, *x*, (containing a desire and at least one means-end belief) held by some agent, *a*, which entails that having an intention to perform some action, *p*, at a time when she is able to perform that action, is sufficient for *a* to be rational, with respect to *x*.

Reason 2 is better than Reason 1, as it accommodates the fact that there can be more than one intention which can be rationally formed, given some relevant set of attitudes. However, Reason 2 faces a different problem. This is that there can be reasons to perform certain actions, where intending to perform these actions is sufficient to be somewhat, but not entirely rational, given some set of attitudes. That is, rationality comes in degrees, as does the strength of reasons. For example, I might desire to have a nice pint of beer, while believing that the beer at the King's Head is quite nice and that the beer at the Ferret and Whistle is much nicer. In this case it seems that I have some reason to go to the King's Head but that I have more reason to go to the Ferret and Whistle. How is this to be characterised?

In terms of rationality, intending to go to the Ferret and Whistle is more rational than intending to go to the King's Head just because I believe that, by going to the Ferret and Whistle, I can better satisfy my desire for a nice pint of beer. Here, the most rational

intention for me to form, given the relevant set of attitudes, is the one corresponding to the action which I believe will best satisfy my desire. That action is also the one which I have the strongest reason to perform, given the relevant attitudes (other complications to do with the strength of my beliefs and desires aside). Nevertheless, intending to go to the Kings Head is still somewhat rational, and I do have some reason to go there. How is this to be captured within the account of reasons?

One suggestion would be to claim that I have a reason to perform any action which, I believe, will somewhat satisfy one of my desires. Hence:

> Reason 3: A practical reason is a set of attitudes, *x*, (containing a desire and at least one means-end belief) held by some agent, *a*, which entails that having an intention to perform some action, *p*, at a time when she is able to perform that action, is sufficient for *a* to be at least somewhat rational, with respect to *x*.

This characterisation might seem to introduce an unacceptable level of vagueness into the account. However, there are two important respects in which the picture is precise. First, there is a clear difference between believing that an action will do nothing to satisfy a desire and believing that it will do something to satisfy a desire. For example, perhaps I believe that the beer in the Rat and Sewer is simply awful. In that case, intending to go there is not at all rational as I believe it will do nothing to satisfy my desire for a nice pint of beer. Insofar as intending to go to the Rat and Sewer is irrational, given my desire for a nice pint of beer and my belief that the beer there is awful, I have no reason to go there.

The other sense in which the picture is precise is that I can have clear beliefs about which action will best satisfy a desire, and about which actions will better satisfy it than others (although perhaps not in every case).[27] Where a clear belief is held about the best way to

[27] Better satisfying here is not being more likely to satisfy (i.e. having a higher probability of satisfying some desire than an alternative), but being likely to satisfy more (i.e. having some degree of

satisfy a desire, the most rational thing to do, given that desire and the associated beliefs, is to intend to perform the action that I believe will best satisfy it. This allows my strongest reason, given some set of attitudes, to be clearly separable from my weaker reasons, even though no exact measure of each reason's strength is available. I can also distinguish between relatively stronger and relatively weaker reasons in this way, without needing to have totally precise beliefs about the degree to which an action will satisfy a desire. So, even though the picture may be somewhat vague, it seems that being able to rank actions in terms of their approximate degree of expected desire satisfaction is all that is needed for some intentions to count as more rational than others, and for some reasons to be stronger than others.

Reason 3 overcomes the problem that there can be actions which we have a reason to perform (such as going to the King's Head), even though these actions are not entirely rational, given the relevant set of attitudes.

However, Reason 3 is also problematic. This is because the term 'somewhat rational' is ambiguous. This can be seen if we consider another kind of example. Suppose that I desire to buy some decent tobacco to put in my pipe. One way to buy some decent tobacco is to go to the local high-quality tobacconist's, making sure that I take some money with me to pay for it. Here, picking up my wallet, leaving the house, walking to the shops, entering the tobacconists, and so on, all play a part in my buying some tobacco. The first of these actions (picking up my wallet) does not suffice to satisfy my desire to buy some decent tobacco to any degree whatsoever. So in one sense, having this intention is not sufficient for me to be at least somewhat rational.

---

probability of better satisfying a desire than some alternative would). Using Ramsey's model of decision-theory, on which we can use some value-neutral alternative as a reference point from which to derive agents' views about the probability of certain outcomes, it may be possible to be more precise about agents' beliefs regarding how much better some action is likely to be at satisfying a desire than some alternative (Ramsey, 1926). However, even if this is possible, this level of precision is unnecessary for present purposes (i.e. revising Reason 2 in light of the possibility that a desire can be more or less satisfied).

However, in another sense, intending to pick up my wallet is sufficient for me to be at least somewhat rational, given my relevant beliefs and desire (and, perhaps, the assumption that I believe that I will carry out the rest of my tobacco-buying plan). This is because picking up my wallet would help in the achievement of my tobacco-buying aim, under the right circumstances. So, intending to pick up my wallet is one of a number of intentions which jointly suffice for me to be (at least somewhat) rational, given my belief-desire pair, such that forming each of these intentions (perhaps on the assumption that I believe that I will enact the whole plan) seems also to at least somewhat rational.[28]

This ambiguity suggests that we need to distinguish between two senses in which forming an intention can suffice for an agent to be somewhat rational. The first sense (which I shall call the satisfying sense) is that of forming an intention to do something which one believes would be sufficient to at least somewhat satisfy a desire. On this sense, forming an intention is sufficient to be at least somewhat rational if that intention corresponds to an action which one believes would, by itself, suffice to somewhat satisfy a desire. The important thing here is that the intention is associated with an action which is itself sufficient to achieve at least the partial satisfaction of a desire.

The second sense in which forming an intention can suffice for an agent to be somewhat rational (which I shall call the contributory sense) is that of forming an intention to do something which one believes will *contribute* to a desire being (at least somewhat) satisfied. On this sense, forming an intention is sufficient to be somewhat rational if that intention applies to an action which one believes will contribute towards the satisfaction of some relevant desire (perhaps on the assumption that one believes that one will carry out the rest of a plan which is sufficient to at least somewhat satisfy that desire).

Now, forming an intention to pick up my wallet is not sufficient for me to be at least somewhat rational in a satisfying sense, but it is sufficient for me to be at least somewhat

[28] The caveat here is due to the issue of actualism versus possibilism, discussed below.

rational in a contributory sense (perhaps given the belief that I will carry out the rest of my tobacco-buying plan). That is, intending to pick up my wallet, if successfully enacted, is not sufficient to satisfy my desire to buy some decent tobacco to any degree whatsoever, but it is one of a set of intentions which, if successfully enacted, suffice for this desire to be at least somewhat satisfied.

Now that these two senses of the term 'somewhat rational' have been distinguished, how does Reason 3 fare? Well, the term 'somewhat rational' in Reason 3 is intended to be read in a satisfying sense. However, on this reading certain cases in which we do have reasons for action fail to qualify. For instance, I do seem have a reason to pick up my wallet in the above example, even though picking up my wallet does not suffice to satisfy my desire to buy some decent tobacco to any degree whatsoever. Because picking up my wallet contributes to the satisfaction of my desire (at least under the right conditions), it seems clear that I have a reason to pick it up (or, at least, that I would have a reason to pick it up if I believed that I was going to enact the rest of the relevant plan). So, we need an account of reasons which includes having reasons to perform actions which are somewhat rational in a contributory sense. This results in:

> **Reason:** A practical reason is a set of attitudes, *x*, (containing a desire and at least one means-end belief) held by some agent, *a*, which entails that having an intention to perform some action, *p*, at a time when she is able to perform that action, is sufficient for *a* to be at least somewhat rational (in a satisfying or contributory sense), with respect to *x*.

Here, 'somewhat rational (in a satisfying or contributory sense)' indicates that, to have a reason to perform some action, an agent must believe that performing that action will either: (i) at least somewhat satisfy a desire (satisfying sense) or (ii) contribute to a desire's being (at least somewhat) satisfied (contributory sense).

A further question, concerning actualism versus possibilism about practical reasons, arises at this point (Jackson and Pargetter, 1986; Woodard, 2009; Zimmerman, 1996: ch.6). Does my

having a reason to perform some action depend upon what I believe that I will actually do in the future, or only on what it is possible for me to do? For instance, if I know that I will not leave the house (I am snowed in and would have to dig myself out, suppose) but would nevertheless would like to buy some decent tobacco, do I still have a reason to pick up my wallet? Given that I know that the other parts of my tobacco-buying plan will not be enacted, does the contribution that picking up my wallet would, under the right circumstances, make to the fulfilment of my desire to buy some decent tobacco give me a reason to pick it up? Or, do I only have a reason to pick up my wallet if I believe that I will enact the other parts of my tobacco-buying plan?

This issue is not fundamental to my account of practical reasons. There is room to develop the account along either actualist or possibilist lines. I have actualist intuitions. Thus being 'at least somewhat rational in a contributory sense' in Reason, above, should be read in terms of being expected to make some actual contribution towards the satisfaction of a desire (rather than being merely being capable of making some contribution towards the satisfaction of a desire, if the right circumstances were to obtain). However, nothing of significance for my overall approach to practical reasons hangs on this decision.

Despite having a somewhat unfortunate degree of complexity, Reason is a categorical statement of the view of practical reasons that I wish to defend. Roughly, Reason is meant to capture, in terms of rationality, the idea that we have a reason to do anything, and each particular thing, that we believe will contribute, in some way, to the satisfaction of one of our desires. Thus the suggestion is essentially that a practical reason is any belief-desire pairing which entails that having an intention to perform some action, at time when one is able to perform it, has some degree of rationality. Forming an intention is rational, in some degree, if that intention corresponds to an action which we believe will help, in some way, to satisfy a desire.

In the next chapter I discuss some important features of rationality, including the rational relations between beliefs, desires and intentions which underpin this account. Prior to that,

there are some important issues regarding the strength of practical reasons, and the distinction between overall and *pro tanto* reasons, which need to be addressed.

7. How does the Account Accommodate Reasons of Different Strengths?

It should be apparent from Reason, or at least from my less formal characterisation of what Reason says, that we have a vast abundance of reasons for action. We have reasons to do anything and each thing which we believe will help to satisfy one of our desires to some degree. This might seem objectionable to some, in that there are many trivial actions which we have reasons to perform on this view (such as prodding the cat because I have a feint desire to see it squirm). In order to overcome the worry that we have too many reasons, to do too many things, some account of the weight of reasons must be given which shows why most of our reasons for action are too weak to be registered or responded to. This should allay the worry that an account of practical reasons like mine generates reasons which we intuitively do not have, by suggesting that they are not the kind of reasons that we would, or should, normally pay attention to.[29]

As discussed above, one way in which a reason can be relatively less weighty (other things being equal) is for the action associated with it to be perceived as a relatively poor way to satisfy a desire. For instance, eating pizza is a relatively poor way to look after my nutritional needs. So, my desire to eat a healthy diet provides me with only a weak reason for eating pizza. More generally, we only have (relatively) weak reasons to perform those actions which we believe will (relatively) poorly satisfy our desires (*ceteris paribus*).

[29] This kind of strategy is adopted by Mark Schroeder, in his book Slaves of the Passions (2007: ch. 5). Schroeder also raises some general problems with arguments premised on negative reasons existentials (i.e. arguments based on the claim that we obviously do not have a reason to φ) in his 'The Negative Reason Existential Fallacy' (Schroeder, unpublished).

A second way in which a reason can be relatively weak (other things being equal) is for one's degree of belief that the relevant action will (help to) satisfy a desire to be relatively low. The lower one's degree of belief in the desire-satisfying potential of an action, the weaker one's reason for performing that action (*ceteris paribus*).

The third way in which a reason can be relatively weak (other things being equal) is if the relevant desire is also relatively weak. The weaker one's desire for some outcome, the weaker one's reasons for performing the actions which one believes will (help) bring about that outcome (*ceteris paribus*).

A final, possible way in which reasons can be relatively weak (other things being equal) is for there to be a relatively large number of alternatives which we believe will (help to) satisfy a desire. Take the case of Buridan's ass. Suppose that the ass is hungry and has only one bale of hay to choose from. In this case, anthropomorphising somewhat, the ass has a relatively strong reason to eat from this bale. However, if, as in the standard example, there are two equally good bales of hay to choose from, one might think that the ass's reason for eating from either one is only half as strong. Likewise, if there were 1,000 bales, one might think that the ass's reasons for eating from each of them would be a thousand times weaker than his reasons for *eating from one of the available bales*. Intending to eat from any particular bale is one of a thousand intentions which would suffice for rationality, given the ass's desire for food. As such, perhaps the degree of rationality associated with forming each such intention is only one-thousandth of the degree of rationality associated with forming an intention to *eat from one of the available bales*. Plausibly, the more options one has for satisfying a desire, the weaker one's reasons for taking any particular option (*ceteris paribus*).[30]

[30] Others' intuitions may diverge from mine on this claim. I have no specific arguments in favour of it. For this reason, I introduced it as a 'possible' way in which reasons can be relatively weak, and discussed it in correspondingly tentative terms.

Given these four different ways in which a reason can be relatively weak, it should not be too much of a problem that the account of practical reasons allows that we have many reasons to do lots of petty things. Such reasons are, for one reason or another, relatively weak.

8. What about Overall Reasons?

So far I have been discussing practical reasons which consist of some subset of an agent's beliefs and desires (a desire together with at least one means-end belief). In line with contemporary terminology, I shall refer to these as *pro tanto* reasons—reasons so far as they go. However, for the account of practical reasons to be successful, it must be able to account for overall reasons for action. How do our *pro tanto* reasons combine, on the proposed account, to give us overall reasons to perform certain actions?

Well, given the decision-theoretic model of practical reasoning that I invoke, accounting for overall reasons turns out to be a relatively straightforward task. On such a model, practical rationality consists in intending to perform those actions which will bring about maximal desire satisfaction, given one's beliefs. Thus a (maximally) rational agent forms a set of intentions to perform those actions which, given her beliefs, will maximise her degree of desire satisfaction overall. Given this model of practical rationality, I propose that an agent has an overall reason to perform any action which she believes will contribute towards the maximal satisfaction of her desires overall. This brings me to the following account of overall reasons:

Overall Reason: An overall reason is a set of attitudes, x, (containing all of the desires, and all of the (relevant) means-end beliefs) held by some agent, a, where these entail that having each and every particular member of some set of intentions, $i_1, i_2, i_3,...i_n$, to perform the corresponding actions, $p_1, p_2, p_3,...p_n$, at times when she is able to perform those actions, is sufficient for a to be most rational, with respect to x.

Here, the set of attitudes, *x*, is an overall reason for *a* to perform each and every action in the set of actions, $p_1, p_2, p_3,...p_n$,. This is because intending to perform each of these actions (at a time when she is able to perform them) contributes in part to *a*'s having each of the members of the set of intentions, $i_1, i_2, i_3,...i_n$, which it is most rational for her to have, at the times when she is able to perform the relevant actions (assuming, for actualist reasons, that she has every other member of this set of intentions at an appropriate time).

Having each and every member of that set of intentions is sufficient for *a* to be most rational, but not necessary, as there may be other sets of intentions which, if enacted, would also bring about maximum desire satisfaction (there can be many equally good ways of satisfying any particular desire).

So, the proposal is that the total set of an agent's desires and her (relevant) means end beliefs are an overall reason for her to perform certain actions.[31] The actions that these attitudes are an overall reason for her to perform depend upon the content and strength of the attitudes in question, as it is the content and strength of an agent's beliefs and desires that determines which courses of action are sufficient to maximally satisfy her desires, according to her means-end beliefs, and thus which intentions it is rational for her to have at the relevant times.

9. Conclusion

In this chapter I have set out the interpretivist account of practical reasons that I wish to defend. The view is essentially that a practical reason is a belief-desire pairing which entails that an agent's intending to perform some action (at a time when she can perform that

[31] It is worth noting that I stipulate that the means-end beliefs which feature in an overall reason must be relevant in the sense that they relate to the satisfaction of one of the desires that the agent in question has. Any contradictory beliefs (from which anything follows) are ruled-out in that the belief with less credence is taken not to be relevant.

action) is sufficient for her to be at least somewhat rational. Belief-desire pairs qualify as practical reasons in this sense just because forming the intentions which rationally follow from them is necessary for agents to be interpretable. Thus it is the requirements of interpretability which, ultimately, select rational mental functioning as an appropriate categorisation scheme for practical reasons, and which provide its normative force. If interpretable mental functioning involved something other than rational functioning, the account of practical reasons would be vastly different.

The proposed account of practical reasons is ontologically reductive, identifying practical reasons as sets of propositional attitudes which entail that performing actions of certain kinds is sufficient to be interpretable. No *sui generis* normative notions are invoked in this account (not, at least, if rationality can be invoked as a categorisation scheme for interpretable mental functioning, without reference to an independent notion of normative force).[32] The account of practical reasons' normativity is also mind-independent, in that their normative force does not depend on any agent's existing attitudes. Rather, I have suggested that normative force depends on interpretable creatures' constitutive commitment to function in accordance with the requirements of interpretability. Finally, the account is hypothetical, in that the specific practical reasons that agents have depend on the contents of their desires.

It is hoped that the account of practical reasons that I have proposed is at least plausible (or, at the very least, not entirely implausible). This account is motivated by a naturalistic bias, together with a concern to maintain the mind-independent normative status of practical reasons, which seems to be their hallmark. As explained at the beginning of the chapter, I take the requirements of interpretable functioning to be a good candidate in explaining practical reasons. This is because of interpretability's centrality to practical life. It is also because the requirements of interpretability can be seen to pair certain attitudes with certain actions, much as practical reasons are paired with certain actions by practical reasons' normative scheme. Given that, for the naturalist, finding an appropriate normative

[32] And not if the objection from irreducible normativity, discussed in chapter 4, can be avoided.

scheme by which to categorise practical reasons is problematic, it seems sensible for them to consider interpretivism as a candidate for providing a naturalistic explanation of what practical reasons are, and of their normativity.

In the next chapter, I take a more detailed look at the kinds of rational relations which underpin interpretable mental functioning. Thereafter I attempt to respond to a couple of key objections to my account of practical reasons: (i) that it relies on irreducibly normative properties or relations; (ii) that it cannot accommodate the possibility of moral reasons or action. It is hoped that I manage to set these objections aside, such that my proposed account of practical reasons can be treated as a viable naturalistic option.

Chapter 3: Rational Agency

1. Introduction

In chapter 2 I set out an interpretivist account of reasons for action. According to this account, practical reasons are belief-desire subsets from which certain intentions to act (and, derivatively, certain actions) rationally follow, where rationality is a mode of interpretable mental functioning. In this chapter I discuss the aspects of rationality that determine which intentions rationally follow from a belief-desire subset. I approach this task by considering how rationality determines the intentions which follow from an agent's overall set of beliefs and desires, and subsequently localise this to particular subsets of beliefs and desires.

First, I give an overview of the elements of practical psychology which feature in the account of rational agency. These include beliefs, desires, all things considered judgements, intentions, and some principles of rationality which connect these attitudes (notably, Davidson's Principle of Continence—Davidson, 1969: 41). I then discuss the Principle of Continence in detail. I suggest that this principle applies to intentions and consider (but reject) a possible revision to it to accommodate cases where we expect our all things considered judgement about an action to change in the future. Third, I discuss alternative principles which might rationally order our intentions, given our total set of beliefs and desires (in particular, a Principle of Maximisation). I suggest that it is unclear whether the Principle of Maximisation takes rational priority over the Principle of Continence but that, in any case, it is the latter which takes psychological priority. Fourth, I discuss the puzzle of how incontinence can be psychologically possible, given that we hold a principle which directs us towards continence. Given the Principle of Continence, it is puzzling how any agent can deliberately fail to do what she judges that she has most reason to do. Psychological explanations of incontinence are important, not least because they support the claim in section 3 that continence takes psychological priority over maximisation. Fifth, I

suggest a Principle of Weak Continence, which confers some degree of rationality on intentions to act in ways which we believe will promote desire satisfaction but not to the highest available degree. This principle is central to the explanation of *pro tanto* reasons given in chapter 2. It also marks the limits of interpretable action; it is a minimal criterion which applies even to incontinent actions. Finally, I conclude by drawing the different threads together and relating the discussion of rational agency in this chapter to the account of practical reasons given in chapter 2.

2. Elements of Rational Agency

The view of rational agency that I adopt is a maximising, decision theoretic view.[33] Roughly stated, this is the view that agents are rational to the extent that they perform actions which maximise preference satisfaction, given their means-end beliefs (i.e. maximise expected utility).[34] This is the view of agent rationality which underpins Davidson's view of interpretability (Davidson, 1980; 1995). However, in the first instance rationality concerns the relations between agents' mental states. Rationality only concerns the relations between agents' mental states and their bodily movements derivatively (i.e. insofar as these movements are associated with some mental state). As such, some mental state needs to be specified as the attitudinal counterpart to actions, such that actions can be treated as open to rational criticism.

One suggestion would be that Davidsonian all things considered (ATC) judgements about what it is best to do are the attitudinal counterpart to actions (Davidson, 1969). These are second order beliefs about what it is best to do overall, given one's various beliefs and desires. However, given that an agent can intentionally act against her better judgement

[33] For purposes of simplicity I set aside more complex views of rational agency than those framed directly in terms of acting in ways which maximise expected levels of preference satisfaction. These include Gauthier's model of constrained maximisation and McClennen's view of dynamic choice (Gauthier, 1986; McClennen, 1990). The plausibility of such views is not questioned.

[34] The term 'maximising expected utility' here is given its subjective, decision-theoretic meaning rather than its objective, utilitarian meaning.

(that is, act incontinently), ATC judgements cannot be the required attitudinal counterpart to actions.

A second suggestion would be that it is individual belief-desire pairs which are the attitudinal counterpart to actions. This seems to be the view taken by Davidson in 'Actions, Reasons and Causes', where he identifies anything which an agent does intentionally as an action, where to do something intentionally is to be caused to do it by an appropriate belief-desire pair (i.e. a reason) (Davidson, 1963). On such a view, the rational criticism of actions could be seen as the criticism of actions in terms of the particular beliefs and desires that an agent acts upon, given her total balance of preferences. For instance, suppose that Stella's total balance of preferences (given her means-end beliefs) favours donating £100 to charity, but that she fails to do so because she wants to buy a new dress. In this case, failing to donate the money to charity is irrational because Stella's omission to make the donation is caused by a reason which goes against her greater overall balance of reasons. Here it is not the desire for a new dress which is irrational, but Stella's acting upon it, given her overall balance of reasons. Thus on this view the rational criticism of actions does not concern the attitudes one holds, but rather what they cause one to do.

This may be a plausible view, but I will not discuss it further because there are good reasons for introducing a further kind of attitude into the account, namely intentions. One reason for introducing these is that the bare belief-desire view of intentional actions cannot sufficiently accommodate the possibility of incontinent actions, given certain other intuitive premises. This is argued by Davidson in 'How is Weakness of the Will Possible?' (Davidson, 1969). Davidson defines incontinence as follows:

> D. In doing *x* an agent acts incontinently if and only if: (a) the agent does *x* intentionally; (b) the agent believes there is an alternative action *y* open to him; and (c) the agent judges that, all things considered, it would be better to do *y* than to do *x* (ibid: 22).

Incontinent actions are intentional actions performed against one's better judgement. There are two premises that Davidson finds strongly intuitive which, together, appear to contradict the claim that there can be incontinent actions. These are:

> P1. If an agent wants to do *x* more than he wants to do *y* and he believes himself free to do either *x* or *y*, then he will intentionally do *x* if he does either *x* or *y* intentionally (ibid: 23).
>
> And
>
> P2. If an agent judges that it would be better to do *x* than to do *y*, then he wants to do *x* more than he wants to do *y* (ibid: 23).

P1 ties wanting more to acting; if an agent wants to perform one action more than another, she will perform that action if she performs either action intentionally (assuming that she believes that she is free to perform either action). This premise is strongly intuitive. Davidson takes it to illustrate the basic nature of wanting something; to want something is to be disposed to try to get it (ibid: 22). Davidson strengthens this observation by invoking Hampshire's suggestion that if a person wants to do something then, other things being equal, she will do it if she can. Here he takes 'other things being equal' to mean 'provided there is not something he wants more' (ibid: 22-23). The result is P1.

P2 ties judging better to wanting more; if an agent judges one action better than another then she wants to perform that action more. This premise may seem less intuitive, particularly as the notion of 'judging better' is somewhat opaque. However, it does seem intuitive that there is at least *some sense* in which judging an action better entails wanting it more. Thus Davidson writes that:

> It is easy to interpret P2 in a way that makes it false, but it is harder to believe there is not a natural reading that makes it true. For against our tendency to agree that we often believe we ought to do something and yet don't want to, there is also the

> opposite tendency to say that if someone really (sincerely) believes he ought, then his belief must show itself in his behaviour (and hence, of course, in his inclination to act, or his desire) (ibid: 27).

Taken together, P1 and P2 seem to rule out:

> P3. There are incontinent acts (ibid: 23).

According to Davidson's definition of incontinence, acting incontinently involves acting intentionally against one's better judgement about how to act, all things considered. However, according to P1 and P2 combined, if an agent judges that it is better to do *x* than *y*, then she will want to do *x* more than *y*, such that, if she believes that she is free to do either *x* or *y*, she will intentionally do *x* (if she does either *x* or *y* intentionally). That is, P1 and P2 entail that an agent will only intentionally act in ways that she judges better, given what she takes herself to be free to do. This appears to rule out incontinence, as defined above.

Davidson's solution to this problem is to distinguish between two senses of judging better, an ATC sense and an all-out sense.[35] I have already explained the notion of judging an action better ATC. According to Davidson, the attitude of judging an action better all-out is a *sui generis* propositional attitude; it is the attitude of favouring an action outright or, in more familiar terms, intending that action (Davidson, 1978). Intentions do not necessarily cause actions; they may come into existence along with an action or perhaps, as Davidson

[35] He also refers to these as conditional and unconditional senses of judging better, but it has been pointed out to me that this is problematic. ATC judgements do not have the content '*if* these are all of my reasons, then I should *x*', rather, they have the content 'I should *x*, given all of my reasons'. Thus ATC judgements are better regarded as judgements relative to one's total belief-desire set than 'conditional' judgements and all-out judgements are better regarded as outright judgements than 'unconditional' ones.

suggests, be identical to actions in some instances (ibid: 98-99). They can also exist independently of deliberation and successful execution (ibid: 89).

By introducing all-out judgements (that is, intentions) into the account of agency, Davidson can overcome his problem concerning the logical possibility of incontinent actions. This is because, if P2 is read in terms of all-out judgements, the conjunction of P1 and P2 does not contradict P3, given Davidson's definition of incontinence. Even if judging an action better all-out entails wanting it more, and thus performing it if one does anything intentionally, this does not entail judging that it is better ATC. What is important in the case of incontinent actions is that an agent judges against an action, ATC, but still performs it intentionally. Reading P2 in terms of all-out judgements permits this. Thus, by introducing two senses in which an agent can judge an action better—an all-out sense and an ATC sense—Davidson can square P1 and P2 with P3.

I think that this is a good reason to accept Davidsonian all-out judgements. I find each of his three premises strongly appealing. P2 is the weakest, but I am inclined think that there is at least some sense of judging better on which judging an action better entails wanting it more. On Davidson's proposal, intending (judging an action better all-out) is just this sense.

However, Davidson's line of reasoning may be unconvincing to some. It may appear question-begging, for instance, for him to appeal to a somewhat opaque sense of judging an action better in P2, and then use the ensuing conflict between P1, P2 and P3 to argue that an all-out sense of judging better needs to be posited, such that P2 can be accommodated, given P1 and P3. If it is so intuitive that judging an action better is tied to wanting it more, one might expect the exact sense of judging better in question to be clear from the start. If this were so then the existence of all-out judgements would have been assumed prior to the argument which is taken to support introducing them.

Nevertheless, Davidson can perhaps be more charitably interpreted as flagging up the fact that there is some sense in which judging an action better entails wanting it more and that,

given P1 and P3, this cannot be an ATC sense. So it must be something else. His suggestion: that there is an all-out sense in which we can judge one action better than another.

Even so, it may be suggested that there are alternative ways of overcoming the perceived conflict between P1, P2 and P3. For instance, perhaps there are two different senses in which an agent can want something more, such that P1 and P2 can be squared with P3, even if P2 is read in terms of ATC judgements. One proposal here is to distinguish between *motivational* and *evaluative* senses of wanting an action more. Suppose that an agent wants to perform an action more in a *motivational* sense if she is more strongly disposed to do it than some alternative. By contrast, suppose that an agent wants to perform an action more in an *evaluative* sense if that action will better satisfy her overall balance of desires than some alternative action, given her means-end beliefs.[36]

Now, if 'wanting more' in P1 is read in a motivational sense, whilst 'wanting more' in P2 is read in an evaluative sense, then these two principles do not necessarily contradict P3, even if 'judging better' in P2 is read as judging better ATC. That is, if wanting more in a motivational sense entails acting, and judging better ATC entails wanting more in an evaluative sense, then this does not necessarily rule out the possibility of judging that an action is better ATC and yet intentionally performing some alternative. An agent could judge an action better ATC and thus want it more in an evaluative sense, while simultaneously wanting some alternative more in a motivational sense, thus performing that action.

However, this suggestion is problematic because it relies on the claim that judging an action better ATC entails wanting it more in an evaluative sense. This claim is implausible because it rules out the possibility that we can be mistaken in our ATC judgements (the possibility that we can sometimes judge a non-expected utility maximising action best, ATC). Sometimes, I assume, we can judge an action better ATC even if it is not the action which

[36] Heil makes a similar distinction in his discussion of the psychological conditions required for incontinent action. He suggests that incontinence occurs where an agent wants something with more motivational force than she does evaluative weight (Heil, 1989: 581).

will maximise preference satisfaction, given our means-end beliefs. This is because we are prone to errors of calculation and to errors of omission; sometimes we fail to take certain courses of action into proper account when weighing up what it is best to do. Therefore, even though the suggestion that 'wanting more' is equivocal would accommodate the possibility of incontinence, given P1 and P2, without the need to introduce Davidsonian all-out judgements, it would nevertheless rule out the possibility of a different kind of practical mistake; judging the wrong (i.e. non-expected utility maximising) action better, ATC.

So, in order to square P1 and P2 with P3, perhaps an all-out sense of judging better does need to be introduced. Even so, it may not be agreed that there is any sense in which judging better does entail wanting more. If P2 is simply rejected then Davidson's claim that there must be both an all-out and an ATC sense of judging better lacks support.

However, there is another reason for introducing intentions into the account of agency. This is that it is possible to have a pure intention: an intention which is formed in the absence of deliberation and which is never enacted. For instance,

> Someone may intend to build a squirrel house without having decided to do it, deliberated about it, formed an intention to do it, or reasoned about it. And despite his intention, he may never build a squirrel house, try to build a squirrel house, or do anything whatever with the intention of getting a squirrel house built' (Davidson, 1978: 83).

If we accept that we can intend to do things without deliberation or any attempt at execution, then there must be more to intending than acting intentionally, as the class of intendings is broader than the class of intentional actions.

Davidson argues that pure intending cannot be a kind of believing or a kind of desiring. Reasons for believing are different in kind to reasons for intending. Desires are *prima facie* while intentions are not (ibid: 91-99). Rather, Davidson claims that pure intending must be a *sui generis* propositional attitude—the kind of all-out judgement introduced above. Given

that it would be artificial to distinguish between cases of having such an attitude and not having it simply on the basis of whether we actually perform an action or not, Davidson holds that this kind of all-out judgement in favour of an action is present in all cases of intentional action (ibid: 88-89). Thus on a Davidson's picture, intentions are the attitudinal counterpart to actions.[37]

I propose to accept this picture. As well as being inclined to accept the argument for all-out judgements which emerges from Davidson's account of incontinent actions, I think that his suggestion that such judgements can be seen to accompany intentional actions is right. If cases of pure intention involve having a distinctive propositional attitude then it seems entirely arbitrary (if not absurd) to hold that this attitude only obtains where we *fail* to act for a reason. Thus, from now on, I will assume that intentions play the role of attitudinal counterpart to actions.[38]

One important caveat here is that not all intentions are attitudinal counterparts to actions. It is intentions which are held at the time of action which correspond to what an agent does, and in terms of which an agent can be rationally criticised for doing it. Intentions held at times other than the time of action do not, quite clearly, correspond to any particular action. A related point here concerns the rationality of performing an action. This is not to be

---

[37] Davidson thinks that sometimes performing an action is identical to making an all-out judgement in favour of it (e.g. cases where the conclusion of our practical reasoning is to act 'straightaway') (Davidson, 1978: 98-99).

[38] A further argument to support introducing intentions into the account of agency comes from Bratman (1999:ch.1), who points out that agents need the capacity to fix what they are going to do in the future, such that they can manage their present actions accordingly. For Bratman, intentions play the role of defeasible resolutions. Intentions can change but where an intention is formed this will be enacted at the time of action (assuming that this is registered by the agent) unless there is cause to reconsider it (ibid: ch.7). The default is that intentions are retained and enacted. Differences between Bratman's and Davidson's accounts of intention aside, considerations of Bratman's kind do not establish that intentions accompany actions in cases where deliberation is immediately prior to action, or where there is no deliberation about some action at all.

judged simply in terms of the rationality of holding an intention to perform that action but, rather, in terms of the rationality of holding an intention to perform that action *at a time when one is able to perform it*. For instance, it might be rational for me now (on Wednesday) to intend to visit my newborn nephew at the weekend. However, suppose that I come down with a heavy cold on Friday such that, on Saturday morning, it is no longer rational for me to intend to travel down to visit him. In this case, travelling down to visit my nephew is not a rational action, despite the fact that intending to travel down to visit him was rational earlier in the week. At a time when I am able to *visit my nephew at the weekend* (i.e. Saturday or Sunday), intending to do so is not rational and, therefore, actually doing so is not rational either.

This is important for the account of reasons given in chapter 2. An attitudinal subset is a reason for action only if it entails that intending to perform some action is (at least somewhat) rational *at a time when one is able to perform it*. Attitudes which rationalise intentions at times other than a time when one is able to act upon them are not reasons to perform some action.

Given the above discussion, the account of rational agency that I am employing makes reference to four types of attitudes: beliefs, desires, ATC judgements (second order beliefs about what it is best to do) and intentions (all-out judgements in favour of an action). Of these, it is intentions held at the time of action which serve as the attitudinal counterpart to actions – as the attitudes in terms of which actions can be rationally criticised. Meanwhile, it is an agent's beliefs and desires which entail that forming certain intentions is rational. For this to be the case, there must be certain rational principles (or rules of practical inference) which connect intentions to beliefs and desires. In this section I introduce two such principles: the Principle of Continence and the Principle of Rational ATC Judgement. I discuss the Principle of Continence in detail in section 3. In section 4 I introduce another rational principle: the Principle of Maximisation, which is a general maximising principle of the sort assumed by standard models of decision theory.

The Principle of Continence is posited by Davidson in his explanation of why incontinent actions are irrational (Davidson, 1969). Davidson's principle of continence states: 'perform the action judged best on the basis of all available relevant reasons' (ibid: 41).[39] On Davidson's formulation, the principle of continence connects ATC judgements to intentional actions (actions caused by a reason). In the next section, I suggest that the principle of continence should be seen as applying to intentions, rather than to intentional actions. The Principle of Continence, thus revised, connects ATC judgements with intentions. This is important because I am treating intentions as the attitudinal counterpart to actions, in terms of which actions can be rationally criticised and in terms of which the account of practical reasons is framed. On this account, the Principle of Continence is a primary principle for ordering agents' intentions, determining (in conjunction with the Principle of Rational ATC Judgement) what their various beliefs and desires are reasons to do.

The Principle of Continence does not provide a complete account of which intentions rationally follow from an agent's beliefs and desires; it accounts only for the intentions which rationally follow from agents' ATC judgements. Therefore, a second principle of practical rationality is required to connect an agent's beliefs and desires to her ATC judgements. This principle is: 'judge best whichever action will maximise desire satisfaction, according to one's means-end beliefs'.[40] I call this the Principle of Rational ATC Judgement. I will not argue for this principle. This is because I take it as intuitive that this principle is in keeping with the general decision-theoretic approach towards rationality that I adopt. According to decision theory, rational actions are actions which maximise expected utility. According to the Principle of Rational ATC Judgement, a rational ATC judgement is one on which the expected utility maximising action is judged best. The content of rationality is the same in both cases. The only difference is that the maximising principle in decision theory

[39] Note that for Davidson, as for me, reasons are belief-desire pairs.

[40] I do not mean to claim that agents must consciously apply this principle, or even consciously entertain an ATC judgement in deciding how to act (although I assume that sometimes they do consciously entertain such judgements). To make an ATC judgement is to be interpretable as having made one, but not necessarily to consciously entertain it.

concerns actions whereas the Principle of Rational ATC Judgement concerns second-order beliefs about which actions are best.

One might worry that the Principle of Rational ATC Judgement that I propose is at odds with a decision-theoretic approach to practical rationality, given Lewis' desire-as-belief result (Lewis, 1988; 1996). Lewis shows that the value (i.e. expected utility) of some proposition, *A*'s, obtaining cannot be identical to the degree of credence that an agent places in some proposition, *Å*, which concerns *A* (for instance the proposition that *A*'s obtaining would be good). According to my proposal, a rational agent will believe the proposition (*Å*) that acting so as to bring about that *A* is best, if and only if *A*'s obtaining maximises expected utility. One might translate this into formal decision-theoretic terms as follows: the degree of credence that a rational agent places in *Å* (the proposition that acting so as to bring about that *A* is best)= the expected utility of *A*'s obtaining. This appears to directly conflict with Lewis' result.

Lewis directs his proof against certain anti-Humean theories of practical rationality (the sort which claim that an agent desires some outcome to the extent that she believes that outcome would be good). There have been various anti-Humean responses to Lewis. These include: (i) suggesting that the anti-Humean's position is more moderate than Lewis supposes, claiming merely that we can only be motivated by desires which result from beliefs, but not that these desires must involve valuing an outcome to the extent that one believes that it is good (Broome, 1991); (ii) distinguishing evaluative and non-evaluative propositions about an outcome, where *Å* involves the former and *A* involves the latter (Bradley and List, 2009); (iii) claiming that Lewis' particular decision-theoretic formulation of the desire-as-belief thesis has implausible implications for practical judgements and so must be wrong (Daskal, 2010; Weintraub, 2007).

These suggestions, whether succesful anti-Humean strategies or not, are not appropriate for my purposes, given my proposed formulation of the Principle of Rational ATC Judgement (above). However, one suggestion, from Hájek and Pettit (2004), is appropriate. Hájek and Pettit point out that Lewis' result only holds if the halo proposition has the same semantic

value in all contexts (e.g. if '*A*'ing is good' means the same thing regardless of any changes in the attitudes that an agent holds). However, if the content of the halo proposition is treated as indexical, Lewis' result does not hold. The desire-as-belief result depends on there being a halo proposition with a univocal content, where a (rational) agent's credence in this proposition cannot, as he shows, vary equally with changes in the expected utility of an action. But if the content of the halo proposition also changes along with an agent's credence in that proposition then Lewis' result no longer applies as there is no single halo proposition towards which the agent's degree of credence varies (ibid: §II).

Applying Hájek and Pettit's proposal to ATC judgements would involve claiming, for example, that '*A*'ing is best' means that *A*'ing will maximise *current* desire satisfaction (or, at least, that it is the most favourable action, given one's *current* attitudes) . This proposal seems entirely appropriate, given that ATC judgements are judgements about what is best from an agent's own standpoint (i.e. given her current beliefs and desires). Adopting this suggestion would therefore involve claiming that the the semantic value of ATC judgements changes when an agent's desires change, such that Lewis' result is avoided.

Hájek and Pettit suggest that Lewis anticipated this kind of strategy and saw it as a trivial way of avoiding his result (ibid: 84). However, they are unconvinced by this claim, while also maintaining (as Lewis does in his original paper) that 'a trivial truth is still a truth' (ibid: 84). In any case, even if the connection between ATC judgements and expected utility turns out to be trivial, given decision theory and the proposed account ATC judgements, this does not necessarily pose a problem for my account of practical rationality. If the nature of ATC judgements trivially entails that a rational agent's degree of credence in an ATC judgement is identical to the expected utility of the relevant action, this seems to be a perfectly acceptable outcome for my (Humean) project. As such, I maintain that Lewis' desire-as-belief result does not undermine my proposed Principle of Rational ATC Judgement.

Given the principles of Rational ATC Judgement and Continence, the intentions which rationally follow from an agent's total set of beliefs and desires can be specified. Moreover, the actions associated with these intentions are the same as those judged rational according

to standard, maximising models of decision theory. Acting on an intention which corresponds to a rational ATC judgement (i.e. one on which the expected utility maximising action is judged best) just is performing the expected utility maximising action. Thus the Principle of Rational ATC Judgement and the Principle of Continence serve, together, as analogues for standard decision theoretic maximising principles, selecting the expected utility maximising action as the action which it is rational to perform. They are decision-theoretic principles of rationality applied to an agent psychology which includes ATC judgements and intentions (as human psychology plausibly does).

The two nominated principles provide a way rationally deriving intentions, and their associated actions, from an agent's beliefs and desires. Or, in reverse, they provide a way of deriving an agent's beliefs and desires from her actions (given a large enough sample of actions and certain other assumptions about rationality, such as that agents are generally logical, follow principles of epistemic rationality and so on). That is, they are principles of interpretation.

In the next section I discuss the Principle of Continence. I suggest that it should be regarded as an ordering principle for intentions (rather than just intentional actions), given ATC judgements. I also discuss a possible refinement to the principle to accommodate cases in which our expectations about what we will judge best in the future differ from our judgements about what is best now. In section 4 I discuss the Principle of Maximisation as an alternative rational ordering principle by which intentions can be derived from beliefs and desires, such that agents can be interpreted. The conclusion will be that the two principles suggested (the Principle of Rational ATC Judgement and the Principle of Continence) are one pairing from within a class of possible rational ordering principles. Agents are interpretable, so long as they function in accordance with (at least) one appropriate set of ordering principles. However, the various principles are all extensionally equivalent at the point of action, such that beliefs and desires are reasons to perform the same actions, whatever rational ordering principles are applied.

3. The Principle of Continence

Discussions of incontinence often characterise it as action against one's better judgement (Charlton, 1988; Davidson, 1969; Hurley, 1989: ch.8; Watson, 1977).[41] This is true, but misleading. That is, it is possible for an agent to act against her better judgement, where such actions can correctly be described as incontinent. However, what is incontinent about the action is not that what is done goes against what the agent judges it better to do, but that, in so acting, the agent favours acting in a way that she has already judged against. That is, the agent *intends* to act against her better judgement.[42]

Although Davidson formulates the principle of continence in terms of actions, he also includes intentions in his description of incontinence. Hence: 'weakness of the will is a matter of acting intentionally (or forming an intention to act) on the basis of less than all the

---

[41] Incontinent actions are characterised as actions against one's 'better judgement', as opposed to actions against one's 'best judgement' because all that is required for incontinent action is that one acts in a way that goes against a judgement that is more comprehensive than the one on which one acts. 'Comparative judgements suffice for incontinence' (Davidson, 1969: p22).

[42] Mele suggests that there can be cases of incontinence in which an agent goes against her intention (Mele, 1987: 34-5). In Mele's example, John has a biology assignment which involves finding out his blood type. To do so, he must poke a small needle into the tip of his finger. He forms an intention to do so but, as the needle nears his fingertip, he is suddenly unable to do it. According to Mele, John acts incontinently against his intention to prick his finger.

I do not count this as a case of incontinent action. Either it is not a case of intentional action at all, or it is a case in which John's intention changes. If, despite his intention to prick his finger, John finds that he cannot, I would describe his behaviour as involuntary. Failing to prick his finger in this case is not an action, but rather a failure to act (perhaps brought about by fear). On the other hand, if John *decides* not to prick his finger, then I assume that his intention has changed. Either way, I do not take the example to show that agents can act incontinently against an intention which they have at the time of action.

reasons one recognizes as relevant' (Davidson, 1986: 200).[43] Presumably the reason for including intentions in this description of incontinence is that what is important in cases of incontinent action is the agent's state of mind (i.e. her intention) rather than what her body does. As such, it would be artificial to maintain that intentions cannot be incontinent as well as actions.

Audi also proposes that intentions can be weak-willed, along with what he terms 'predominant wants' (Audi, 1979; 1990). I shall not discuss the latter here. Audi gives the following example of weak-willed intention:

> *Suppose S judges that he should not take a drink and quite consciously tries to resist* doing so. He may still form the intention to take one. May he not have thereby exhibited weakness of the will? Even if he is not able to take one because the bottle is empty, he has already failed in the kind of inner struggle that often precedes incontinent action (Audi, 1979: p181).

The point is that in cases where physical circumstances conspire so that we cannot act incontinently, we may nevertheless be described as incontinent in that our state of mind is such that, had circumstances been different, we would have intentionally acted against our better judgement.

Given that, strictly speaking, it is intentions that are incontinent, I wish to maintain that the principle of continence should be seen as applying to intentions rather than to actions. Hence:

> PC. Intend to perform the action judged best on the basis of all available relevant reasons.

[43] Note that Davidson treats incontinence and weakness of the will as synonymous.

This seems to be an appropriate ordering principle for our intentions, given our ATC judgements. However, PC is open to potential counter-examples. For example, suppose that I am deciding whether to visit my parents for Christmas or whether to have a quiet Christmas at home. Right now I am not very keen on the idea of having to make a long journey on icy roads at a very busy time of year. I also find the idea of spending some quiet time in my own home at Christmas very appealing. So my ATC judgement now is that it would be better not to go to my parents' for Christmas. However, suppose I know that come December I will be missing my parents and that I will feel a certain weight of obligation to go to visit them. Suppose that I also expect the idea of being alone during the festive season to become less appealing as I start to hear about other people's festive plans. These considerations lead me to expect that when Christmas time arrives, I will judge it better to go to visit my parents.

Now, suppose that my parents have just invited me for Christmas and are expecting a reply. Although I can change my mind either way in the future, I need to make some sort of decision now. The question is, should I decide to visit my parents at Christmas or not? If I decide to visit them then I will be going against my present ATC judgement about what it is best to do for Christmas but keeping in line with what I expect I will judge best at the time. If I decide not to visit them, I will be going against what I expect to judge best at the time but keeping in line with my current ATC judgement about what it is best to do. What does rationality require: intending what we judge best now, or what we (most) expect ourselves to judge best at the time of action?

On the one hand, it seems like it is our expected judgement at the time of action which is important, so far as forming rational intentions is concerned. If I expect that, come Christmas time, I would rather be at my parents house than at my own, then I should now form an intention to go to my parents' house for Christmas. On the other hand, we can imagine cases in which it seems clearly rational to intend in line with our present ATC judgement rather than our expected future judgement. For example, suppose that I am aware that in the future I will, against my wishes, be brainwashed by the leader of a religious cult into subscribing to what I currently regard as a debased value system. Once

brainwashed, I will judge that it is best ATC to engage in certain debased rituals. Although I may not be able to help judging that it is best to engage in these rituals at the time, it nevertheless seems irrational for me to now intend to engage in these rituals which, on the basis of my present values, I judge against performing ATC. How can it be rational for me to intend to do something which I currently regard as debased and which I only expect to regard as sacred on the assumption that I am going to be unwillingly brainwashed?

These two examples suggest that it is sometimes rational to intend in line with our expected future ATC judgements and sometimes rational to intend in line with our present ATC judgements about how it will be best to act in the future. If that is true then we are no further forward in establishing whether the Principle of Continence relates intentions to ATC judgements that we expect to make at the time of action, or to ATC judgements that we make now.

One suggestion is that, in the Christmas case, my present ATC judgement does not go against visiting my parents' for Christmas but in favour of it. Perhaps I should be seen as having a general desire to satisfy the desires that I expect myself to have when Christmas time arrives. I expect myself to desire to see my parents during the festive season, to desire to satisfy an obligation to visit them for Christmas, and to desire to spend the festive season with my family rather than alone. Perhaps I endorse these desires in the sense that they are desires that I have a current preference for satisfying, on the assumption that I develop them in the future; they are desires I am now in favour of my future self satisfying, should he have them. If so, perhaps my ATC judgement now includes the judgement that to best satisfy my current preferences, including for the satisfaction of the desires that I expect to have at Christmas, I should go to my parents' house for Christmas. If this preference is strong enough then my current ATC judgement should be in favour going to my parents' house for Christmas, despite my current lack of a desire to go there.

This suggestion is compatible with the intuition that in the brainwashing case it is irrational to intend to perform the rituals which I currently think are debased. Here I do not have a preference for the satisfaction of the desires that I expect myself to have in the future. In

fact, I prefer that these desires are not satisfied. So my current ATC judgement is against engaging in the debased rituals.

Assuming that the above suggestion is right, the Principle of Continence remains as stated in PC, where judged best means 'judged best now'.

4. Alternative Ordering Principles for Intentions

Is PC (in conjunction with the Principle of Rational ATC Judgement) the only rational ordering principle for intentions? Perhaps not. There are certain cases where it seems rational to form intentions which go against PC. For instance, Kavka's toxin puzzle is a case in which it appears rational to intend against one's ATC judgement about what it is best to do (Kavka, 1983).

In this case, an eccentric billionaire offers to pay you £1 million tomorrow morning if, at midnight tonight, you intend to drink a vial of toxic liquid tomorrow afternoon. If consumed, the liquid will cause a day's painful illness. The billionaire advises you that you do not need to drink the toxin tomorrow afternoon to secure the money; you simply need have an intention at midnight tonight to drink the toxin tomorrow afternoon. This is sufficient for him to pay you the money in the morning.

One issue raised by this puzzle concerns the possibility of forming an intention to do something which you have no reason to do, and strong reason not to do. On this issue Kavka claims that 'you cannot intend to act as you have no reason to act, at least when you have substantial reasons not to act' (ibid: 35).

Another problem is that it is unclear what it is rational to intend in this case. On the one hand it seems that intending to drink the toxin is irrational, given that you have no reason to drink it and significant reason not to. On the other hand, it seems that intending to drink the toxin (or at least trying to bring it about that you intend to drink the toxin) is rational, given that by intending to drink the toxin you stand to gain £1 million. If this second intuition is

even partially correct then the Principle of Continence cannot be the only rational ordering principle for intentions. Assuming that you judge, ATC, that it would be better not to drink the toxin, the Principle of Continence stipulates not intending to do so. This conflicts directly with the intuition that intending to drink the toxin is rational insofar as it will earn you a large sum of money.

This problem leads Kavka to remark that 'when reasons for intending and reasons for acting diverge... confusion often reigns. For we are inclined to evaluate the rationality of the intention both in terms of its consequences and in terms of the rationality of the intended action' (ibid: 35-6). For present purposes I will assume that the rationality of your intention in the toxin case can be correctly evaluated both in terms of its consequences and in terms of the rationality of the intended action (drinking the toxin). Not intending to drink the toxin is rational due to the Principle of Continence and your ATC judgement against drinking it. Intending to drink the toxin is rational due to the benefits of holding such an intention. The question is, why do the benefits of intending to drink the toxin make holding such an intention rational? What principle of rationality supports forming this intention, given these benefits?

My suggestion here is very simple. This is that, in addition to the principles of Rational ATC Judgement and Continence, we also subscribe to the following Principle of Maximisation:

PM. Maximise expected levels of preference satisfaction.

This Principle of Maximisation is more general than the Principle of Continence. It is an ordering principle for any- and everything that we do. This raises issues concerning the rational evaluation of beliefs, in that it can generate conflicts with more specific epistemic principles, such as Carnap and Hempel's Requirement of Total Evidence for Inductive Reasoning (Davidson, 1969: 41). This counsels believing that which you take the total balance of evidence to best support. The problem here is that forming beliefs which go against one's evaluation of the balance of evidence can sometimes maximise expected desire satisfaction. For instance, forming religious beliefs when one is about to die can bring

solace at a time of need. According to the Principle of Maximisation, forming such beliefs will be rational in cases where they maximise expected levels of desire satisfaction (for instance, by significantly reducing one's fear of death when one's strongest desire is to avoid such fear). Thus, accepting the Principle of Maximisation as a general rational principle poses a problem if one wishes to claim that the rationality of beliefs should be evaluated *purely* in terms of their conduciveness to truth.

I do not intend to discuss this problem here, other than to suggest that one might reasonably suppose that following *a tendency* to form inductive beliefs only if these comply with the Principle of Total Evidence for Inductive Reason is apt to maximise expected levels of desire satisfaction. Believing in God when one is about to die might maximise preference satisfaction, given one's other beliefs, but having a tendency which prevents one from doing this might mean that one's total belief set is more conducive to overall preference satisfaction than it would be if one lacked this tendency. A tendency for one's beliefs to track the available evidence is plausibly conducive to maximal preference satisfaction (it will certainly be practically useful in the pursuit of many of one's ends). If so, having such a tendency is counselled by the above Principle of Maximisation.

However, this proposal leads directly back to questions over whether rationality should be understood in terms of individual states and actions or in terms of agents' tendencies. These questions have been set aside at present for purposes of simplicity. It also leads to questions concerning the independence of principles of epistemic rationality from principles of practical rationality. Such questions are not directly relevant to this dissertation and will not be addressed. Finally, there is the empirical question of whether a tendency to follow The Requirement of Total Evidence for Inductive Reasoning actually does maximise expected desire satisfaction.

Let us suppose that these questions have satisfactory answers, such that the Principle of Maximisation can be retained as a general principle of rationality (a principle of rationality which applies to everything and anything that an agent does). This is not strictly necessary for present purposes, as a restricted version of the Principle of Maximisation might be

posited, such as one which counsels forming practical attitudes which will maximise expected levels of preference satisfaction. However, the Principle of Maximisation as stated is more simple and will suffice for the present discussion, the above issues having been noted.[44]

As far as intentions are concerned, the Principle of Maximisation counsels forming whichever intentions will best satisfy one's preferences, given one's means-end beliefs. Thus, in the toxin puzzle, it counsels intending to drink the toxin (if one can), as this will best satisfy one's strong desire to receive a large sum of money. Thus the Principle of Maximisation serves as an ordering principle in terms of which the rationality of intending to drink the toxin can be explained.

This leaves us in a position of conflict. According to one rational principle—the Principle of Continence—intending to drink the toxin is irrational. According to a second rational principle—the Principle of Maximisation—intending to drink the toxin is rational. Thus intending to drink the toxin is irrational in one respect and rational in another. This seems to be the most that can be claimed about the rationality of intending to drink the toxin.

This may seem unpalatable: perhaps a clear answer as to whether intending to drink the toxin is rational or not is a requirement of a theory of rationality. Perhaps it seems undermining for any such theory to permit the application of different rational principles, where these can generate conflicting results. Thus one might ask what rationality really consists in: continence or maximisation?

---

[44] It should also be noted that although the Principle of Maximisation applies to everything an agent does, this does not entail that agents can be rationally criticised for, say, sneezing at an inopportune moment. Rational criticism is the criticism of attitudes in relation to the principles of rationality, or the criticism of behaviour associated with such attitudes (i.e. the criticism of actions). One's involuntary movements are not the subject of rational appraisal as they are purely physical; such movements should be treated as something that happens to a person rather than something that she 'does'.

I am not in a position to answer this question. In fact, I am not sure that it *can* be given a principled answer (that would, after all, explain why the toxin puzzle is so puzzling). Most importantly, I am not sure that the question needs a determinate answer. One way of arguing for this claim is to think of rational principles primarily as principles which (assuming that they are followed) support interpretability. Interpretation is a matter of assigning attitudes to agents in light of their behaviour. So long as the behavioural outcomes which follow from two different sets of rational principles are the same, these principles can be seen as equally valid when it comes to interpretation.

Now, an important point to note here is that the principles of Rational ATC Judgement and Continence are extensionally equivalent to the Principle of Maximisation at the point of action. A continent intention formed on the basis of a rational ATC judgement is an intention to act in a way which accords with the Principle of Maximisation (i.e. to act in a way which will *maximise preference satisfaction, given one's means-end beliefs*). Assuming that what we intend to do at the time of action corresponds to what we actually do (issues of incompetent execution aside), following the principles of Rational ATC Judgement and Continence ensures that, at the time of action, we intend to act (and thus do act) in accordance with the Principle of Maximisation.

This is important so far as the question of competing rational principles is concerned because it shows that the principles of Rational ATC Judgement and Continence do not conflict with the Principle of Maximisation when it comes to the eventual actions which they support. Both sets of principles support intending to do (and therefore doing) the same things when the time for action comes (i.e in the toxin case, not drinking the toxin). This means that the principles of Rational ATC Judgement and Continence have the same interpretive upshot as the Principle of Maximisation. Interpretation is a matter of attributing attitudes on the basis of actions (including speech-acts). Given that the eventual actions which follow from these different principles are always the same, they function equivalently when it comes to attributing beliefs and desires to agents on the basis of how they act. They are not competing principles of interpretation, even though they are principles which can support holding different intentions at points prior to that at which interpretive data (i.e.

action) is generated. This suggests the conclusion that, to be interpretable, agents must function in accordance with some principle/set of principles which generates the same actions, given their beliefs and desires, as those which follow from the principles of Rational ATC Judgement and Continence *and* the Principle of Maximisation.

This might be disputed. For example, it might be suggested that the interpretive data *must* be affected by the intentions that an agent holds, otherwise we would have no reason for attributing future intentions at all. Further, given that PC and PM support having different future intentions in the toxin case, surely the interpretive data which follows from these intentions will be different.

This is an extremely complex point which I cannot fully address here. All I can do is to indicate that I think the way in which the interpretive data will be affected by an agent's future intentions in the toxin case will be primarily a matter of her, for example, saying certain things (e.g. 'I intend to drink the toxin'). My suggestion is that the interpretation of such utterances will be the same according to both PM and PC. That is, on both principles we will be lead to attribute to the relevant agent a belief that she intends to drink the toxin and a desire to sincerely report what she intends. Whether we attribute the *intention* to drink the toxin or not will not depend on which principles of interpretation we use, but rather on our views about the psychological constraints which apply to intending (i.e. on what sort of state we take intending to be). If we take it to be possible to intend without having reasons to act then both PM and PC will permit the attribution of an (incontinent) intention to drink the toxin in cases where an agent sincerely claims to have one (other things being equal). If we take intentions to require reasons to act, then neither principle will allow such an interpretation.[45]

[45] One way in which the principles of Continence and Rational ATC Judgement might generate different interpretations to the Principle of Maximisation is that the former might licence the attribution of ATC judgements while the latter might not (or not, perhaps, as often). However, even if this did apply, it would not be a matter of the different sets of interpretive principles generating inconsistent or conflicting results. It would be a case of one set of interpretive principles supplying a

There may be other principles which have the same interpretive results as the Principles of Rational ATC Judgement and Continence and the Principle of Maximisation. These will also count as rational ordering principles for interpretable agents. What is important is that, so long as an agent's propositional attitudes are ordered by a set of principles which are extensionally equivalent to the principles of Rational ATC Judgement and Continence/the Principle of Maximisation at the point of action, that agent will be interpretable. Any such ordering principles will count as rational principles. There seems to be no reason to suppose that practical rationality must reduce to a single rational ordering principle or combination of ordering principles to which all agents must subscribe.

Nevertheless, one might wonder whether the Principle of Maximisation is the dominant rational principle, to which the principles of Rational ATC Judgement and Continence are subordinate, or whether the principles of Rational ATC Judgement and Continence, and the Principle of Maximisation, are independent rational principles of equal status. I have no clear answer to this question. Although the Principle of Maximisation is perhaps more simple, as is a view according to which there is only one primary principle of practical rationality, this does not suffice to show that the principles of Rational ATC Judgement and Continence are mere heuristics which apply to humans only insofar as they *seem* to depart from maximisation in cases such as the toxin puzzle. I refrain from further speculation about the rational priority of the principles discussed.

As well as rational priority, there is a question of psychological priority. Do we function primarily in terms of the principles of Rational ATC Judgement and Continence, or in terms of the Principle of Maximisation? Or is it a mixture of both?

One reason to think that continence takes psychological priority is Kavka's suggestion that, in the toxin case, it is impossible to form the (incontinent) intention to drink the toxin,

---

richer view of agent psychology that the other, where the specific attributions of attitudes that both sets of principles licence attributing remain consistent.

despite its maximising credentials. However, this is a limiting case because it is a case in which there are no reasons for drinking the toxin and a strong reason not to. Taking a causal view of action (as Davidson does), the absence of any attitudes which rationalise drinking the toxin means that there is nothing to cause us to drink it intentionally, or to form an intention to drink it. Thus in cases where there are no reasons to perform some action but there are reasons not to, we seem bound to intend continently, even if intending incontinently would maximise expected levels of desire satisfaction.

What about in cases where there is some reason to act, albeit a weak one? For example, suppose that you are offered £1 million tomorrow morning for intending, at midnight tonight, to drink the toxin tomorrow afternoon, and a further £1 tomorrow afternoon for actually drinking it. Presumably, in this case, you would still judge ATC against drinking the toxin. It can hardly be worth suffering a day's painful illness for just £1. Nevertheless, perhaps in this case you *are* able to intend to drink the toxin, despite this intention being incontinent. Perhaps you are able to incontinently form the intention to drink the toxin because of your weak reason for drinking it. Cases of incontinence are possible, after all, and in this case you have a strong maximising reason for intending, incontinently, in line with your weak reason for drinking the toxin.

However, as will be discussed in the next section, psychological explanations of incontinence generally rely on the idea that an agent has a strong desire to perform the incontinent action. That is, to act incontinently, some strong motivation to perform the incontinent action must exist which overrides the motivations associated with an agent's ATC judgement against performing some action. However, in the case just described, there is no strong motivation to act incontinently. One only stands to gain a further £1 by doing so. So, if incontinence requires the presence of a strong motivation to perform the incontinent action, it looks as if continence will take priority over maximisation in the above-modified toxin case. In fact, continence will take priority over maximisation in all but those rare cases in which: (a) one has a strong enough motivation to perform the incontinent action; (b) stands to maximise expected levels of desire satisfaction by intending incontinently. Normally in cases where (a) is satisfied, (b) is not (incontinence is not generally a maximising

strategy). Thus, on the whole, continence seems to take psychological priority over maximisation. Most cases in which continence fails are not maximising cases, while even in cases where a maximising but incontinent intention could be formed, this would not be formed *because* of any maximising considerations but simply because one was strongly motivated to do the incontinent thing. We are, it seems, creatures of continence except in cases where certain of our desires just happen to get the better of us.

Nevertheless, we do seem to function at least partly in terms of the Principle of Maximisation in that, where considerations of continence are indeterminate, the intentions we form are governed by considerations of maximisation. For instance, Pink suggests an example in which what it is rational to intend is indeterminate given what one judges it best to do, ATC (Pink, 1991: §6). In this example Dan, a stuntman, must decide now whether to perform a stunt in six months' time, as he needs to decide whether or not to organise publicity. If Dan attempts the stunt without having organised publicity he will undertake a serious risk for very little gain; if Dan organises publicity and then fails to undertake the stunt he will look very foolish. These are both 'mismatches'. Dan's strongest desire is to avoid a mismatch.

Given that Dan's strongest desire is to avoid a mismatch, it is indeterminate at the present moment whether undertaking the stunt would be better than not undertaking it. Which action is best depends on what publicity has been organised, and that is yet to be decided. So Dan is without a determinate ATC judgement about what it is best to; his ATC judgement concerning the performance of the stunt, insofar as he has one, is simply to do whatever will best avoid a mismatch.

Nevertheless, it turns out that Dan has recently had an accident and is feeling risk averse. If the time for performing the stunt was now, he would not want to perform it. He also expects that in 6 months' time he will be considerably less risk averse and expects that he may even decide to perform the stunt even if no publicity has been organised (a mismatch). None of these facts change the fact that Dan most wants to avoid a mismatch and that it is

indeterminate at this point whether performing the stunt in 6 months' time or not will best avoid a mismatch.

So, in the stuntman case, there are considerations which make it rational to intend to perform the stunt which do not influence the rationality of actually performing it. By intending to perform the stunt (and, thereby, to organise publicity) it is more likely that Dan will avoid a mismatch as, in 6 months' time when he is likely to perform the stunt whether or not he has publicised it, he will have already organised publicity. So for maximising reasons, it is rational for Dan to now intend to perform the stunt. However, because no publicity has yet been organised, it is indeterminate at this time whether performing the stunt or not will best avoid a mismatch. So, given that Dan's strongest desire is to avoid a mismatch, it is indeterminate at this time whether, ATC, Dan should perform the stunt. This means that considerations of continence are indeterminate between intending to perform the stunt and intending not to.

In the case as described, Dan (we can suppose) decides to perform the stunt and subsequently organises publicity. In making this decision it seems clear that Dan would be guided by the Principle of Maximisation, rather than by the Principle of Continence. His current ATC judgement is indeterminate between performing the stunt and not performing it such that there is no determinate intention for him to continently form. This shows that we can sometimes be lead to (rationally) form intentions because doing so is the best way to promote desire satisfaction, despite there being no rational support for these intentions from the Principle of Continence.

My conclusion is that where considerations of continence apply (such as in the toxin case), these will psychologically trump considerations of maximisation (at least, in all but those extremely rare cases in which there is some incontinent act which: (a) we have overwhelming motivation to perform and (b) intending to perform just so happens to be maximising). Where considerations of continence are indeterminate, considerations of maximisation will determine which intentions we form.

5. The Psychology of Incontinence

Even if maximising considerations *can* lead us to intend incontinently, at least in certain extremely rare cases, we are nevertheless strongly inclined to intend in accordance with the Principle of Continence. In fact, when we fail to intend to act as we judge best, this often provokes a certain confusion within us (as Davidson puts it, the incontinent agent '*recognizes*, in his own intentional behaviour, something essentially surd'—Davidson, 1969: 42). Failing to intend to act as we judge best generally results in actions which are both incontinent and inferior from the point of view of maximisation (given that continent intentions formed on the basis of rational ATC judgements are, when held at the time of action, maximising intentions). Continence is generally conducive to maximisation and, even if maximisation is not the rationally dominant principle, the Principle of Continence is strongly and independently rooted in our psychology (as evidenced by our reaction in the toxin case).

This makes it puzzling that we are able to stray from continence at all. How is it that we can form intentions to do things that we judge inferior, ATC, when we strongly subscribe to a principle which directs us not to do so? I discuss three suggestions.

The first suggestion, advanced by Davidson (1982; 1986), is that incontinence can occur only if the mind is divided into 'semi-independent' substructures. A mental substructure is a collection of mental states, where consistency between the various states in a substructure is greater than the consistency of the mind as a whole (ibid: 181). Mental substructures of this kind can be seen as analogous to groups of individuals operating within a society. Such groups have a narrower, more consistent set of interests than society as a whole. They also have the capacity to influence what happens at the level of the whole.[46] Mental

[46] The analogy between mental substructures and groups of individuals operating within a society is drawn by Hurley (1989: ch.8: esp. §1).

substructures are not entirely separate in that mental states can be members of more than one substructure (just as individuals can be members of more than one social group). As such, mental substructures should be seen as 'strongly overlapping territories' (Davidson, 1986: p211).

Having posited that the mind can be divided in this way, it is possible to allow that the causal impact of any substructure can go beyond its reason-giving power. For example, my 'pleasure module' (crudely put) might contain a desire to enjoy certain tastes, together with the belief that eating cake will lead to the enjoyment of such tastes. This is a reason for eating cake. My 'health module', on the other hand, might contain a desire to lose weight together with a belief that eating cake will frustrate this desire. Likewise, my 'self-image module' might contain a desire to look thin, together with a belief that eating cake will frustrate this desire. These are reasons against my eating cake. Having weighed-up my competing reasons, my ATC judgement is against eating cake. Nevertheless, my pleasure module might cause me to make an all-out judgement in favour of eating cake. Ignoring my better judgment (which the principle of continence requires that I follow) I eat some cake.

In the situation just described, the act of eating cake is intentional. I have a reason to eat cake (because it is tasty), and I form an intention to eat cake on the basis of this reason, such that I eat some cake. However, the reason for eating cake is not a reason for ignoring my better judgement, given the Principle of Continence. So, although eating cake is rational in terms of the substructure, it is irrational in terms of the mind as a whole. Given these facts, it must be the reasons associated with the substructure which cause me to form an all-out judgement in favour of eating cake, as there is nothing at the level of the mind as a whole which supports this decision. If reasons cause and the mind is unified, there cannot be incontinent acts, for incontinent acts go against reason at the level of the mind as a whole. As such, it is only if the mind is divided into substructures that we can be caused to act both irrationally, and for a reason (i.e. intentionally). This, at least, is Davidson's suggestion.

This suggestion is criticised as needlessly complicated by Heil, who suggests that for incontinence to occur all that is required is that the motivational force of a desire can be

disproportionate to its evaluative weight (Heil, 1989: 581). If a desire has a disproportionately strong motivational force then it might cause an agent to act against the balance of reasons (which is weighed in evaluative terms). This suggestion might be true, although a plausible explanation of how the motivational and evaluative weight of desires can be separated would then need to be offered. However, even if it is true, there is still the fact (of primary concern for Davidson's psychological explanation of incontinence) that, in acting incontinently, an agent does something that she does not have a reason to do *viz.* ignoring the overall balance of reasons, which she ought to heed given the principle of continence.

Davidson's contention is that an agent cannot be caused to do this by her mind as a whole, as her mind as a whole (i.e. her overall balance of reasons), by definition, supports acting continently. Heil's suggestion, drawing on his distinction between evaluative weight and motivational force, is that while an agent's overall evaluative balance of reasons might support acting in one way, her overall motivational balance of reasons might favour acting in another. If it is motivational force which determines how we act, then the agent's motivational balance of reasons can cause her to act against her better evaluative judgement about what to do. That is, she can be caused to act incontinently by something at the level of the mind as a whole: her overall motivational balance.

I am sceptical of Heil's proposal. This is because I am inclined to understand the evaluative weight of desires in terms of their motivational force. That is, I endorse the claim that we see an object as desirable to the extent that we desire it. This leads me to reject an explanation of incontinence premised on a distinction between the evaluative weight and motivational force of individual desires.

However, I am prepared to countenance the idea that desires do not sum, or combine, motivationally in the same way that they do evaluatively (and, particularly, the idea that they do not sum motivationally at all). This means that an agent might regard her best option as the one which the aggregate of her various motivations support (i.e. the evaluatively preferred option), while nevertheless being motivated to do something else

because this is what she has the strongest single desire to do. However, I am then inclined to see incontinence as arising when a single desire is motivationally efficacious, despite being only part of the evaluative picture. This is exactly what Davidson's divided mind hypothesis proposes.

Alternatively, it is logically possible that desires can sum motivationally, and that the way that they sum motivationally is different to how they sum evaluatively. However, I am not in favour of an account of incontinence which invokes this idea unless some explanation of the discrepancy between motivational and evaluative summation can be provided. That is, if desires can sum motivationally at all, why do they sum differently to how they sum evaluatively? In the absence of an explanation, I am inclined to think that either desires do not sum motivationally at all (which is conducive to the divided mind hypothesis), or that they sum in the same way motivationally as they do evaluatively. If the latter is the case, this contravenes Heil's proposal (assuming, as I have, that the evaluative weight of any single desire is the same as its motivational force).

A third, 'picoeconomic' explanation of the psychology of incontinence is offered by Zheng (2001). According to Zheng's approach, the mind is not divided into semi-independent substructures. Rather, it is suggested that our motivations change over time and, in particular, that we are prone to motivational spikes when a reward is within close temporal proximity. For instance, when I am in the coffee shop, my desire to enjoy a tasty slice of cake might dramatically increase relative to its normal background level. According to Zheng, this leads to a temporary change in the balance of reasons such that, in the immediate term, my preferences will be best satisfied by my (say) eating cake. That is, short term motivational spikes change the balance of reasons for a short period of time immediately prior to the availability of a reward such that, at the level of the whole agent, the most rational thing to do *at that time* is to pursue the immediate reward.

One might wonder how this is an account of incontinence, given that the suggestion is that pursuing an immediate reward becomes, near the time of its availability, the rational thing for an agent to do. The answer is that, for Zheng, ATC judgements are not judgements about

what will best satisfy an agent's immediate preferences. Rather, ATC judgements are cross-temporal judgements about what will best satisfy one's balance of preferences over time. So, when making an ATC judgement, one considers how important various outcomes are on average and then comes to a judgement about how it would be best to act in order to best satisfy the average of one's preferences across time. Incontinent actions occur when one intentionally acts against one's ATC judgement because the balance of reasons in the immediate term goes against what one judges best from an ATC (i.e. cross-temporal) point of view.

In effect, Zheng's proposal replaces mental division with temporal division. Incontinent actions are the result of motivational proximity effects; how we act depends on our motivational state at the time of action whereas our ATC judgement depends on our (perceived) balance of motivational states over time.

I will not try to arbitrate between Davidson's and Zheng's accounts of the psychology of incontinence here. It may even be that both views have some truth to them. Perhaps we are sometimes hijacked by a part of our mind, such that we act against our better judgement about what it is best to do (even with respect to our immediate preferences). Perhaps we are sometimes hijacked by our whole mind at a particular time, such that we act against what we judge best from a cross-temporal point of view (i.e. believe ourselves to prefer on average over time). In either case, the conclusion seems to be that in order to act incontinently, some kind of division between the scope of ATC judgements and the scope of our action-causing attitudes needs to be made. Either these attitudes are only a subset of our overall belief-desire set, or they are a subset of the beliefs and desires that we (take ourselves to) have across time, or both.

If our action-causing attitudes always extended as far as our ATC judgements about what it is best to do, then we would be unable to intentionally act against an ATC judgement, given the Principle of Continence. If we form intentions on the basis of everything we take to be relevant then we are necessarily continent, as we are guided by a principle which directs us to intend to do what we judge best, given everything that we take to be relevant.

Incontinence is a kind of tyranny in which what we judge best ATC is over-ruled by some set of considerations which is less inclusive than the considerations which inform our ATC judgement.

6. The Principle of Weak Continence

Sometimes we are hijacked by our desires (or certain subsets of them) into intending, and subsequently acting, in ways that we judge against, ATC. However even in these cases we act on the basis of some belief-desire pair; we act for a reason even if we go against the (perceived) overall balance of reasons. In cases where we fail to act on the basis of any of our beliefs and desires whatsoever, we fail to act at all. Such behaviour cannot be interpreted; no attitudes can be attributed to an agent in the light of it.

This leads me to posit a Principle of Weak Continence, which directs us to: intend to perform only those actions which it is judged will contribute to (at least) some degree of preference satisfaction. This principle marks the limits of interpretability. One cannot hold an intention to perform an action which one believes will not contribute towards preference satisfaction in some way or other.[47] Thus although, in order to be interpretable, one must generally make rational ATC judgements and exhibit (strong) continence, one can nevertheless be interpreted on occasion as incontinent so long as one's incontinence is not so severe as to break the Principle of Weak Continence.

For instance, if I were to succumb to my desire to kick a queue-jumper in the shin, despite my judgement that refraining from doing so would be better ATC, I would be interpretable

[47] It might be objected that sometimes we choose to do the moral thing, say, even though it will not contribute towards the satisfaction of any of our preferences. Given that my account of intentional action, following Davidson, essentially involves both beliefs and desires, I set this kind of suggestion aside. To the extent that Davidson is right that interpretability involves acting in ways which can be rationalised by some relevant belief-desire pair, I take the Principle of Weak continence to mark the limits of interpretable action. If Davidson's account of interpretable action is wrong about the essential role of desires, then this claim will also be wrong.

as having a vengeful desire to see him suffer that had got the better of me. In this case my action contributes to the satisfaction of my vengeful desire even though it goes against my better judgement. Given a general background of continent actions (acting in ways that I judge best), incontinence can be assigned in this case, along with the attribution of vengeful desires. That is, I can be interpreted as a person who succumbed to a vengeful desire without being interpreted as someone who thinks that kicking other people in the shin over minor grievances is acceptable or good.

The Principle of Weak Continence (or analogues to it) underpins the notion of a *pro tanto* reason, as outlined in the previous chapter. *Pro tanto* reasons are reasons to perform actions which can be defeated by the overall balance of reasons but which nevertheless render an action intelligible. As characterised in chapter 2, they are belief-desire subsets which entail that intending to perform some action (at a time when one is able to perform it) is at least somewhat rational. The sense in which an intention can be at least somewhat rational, given some belief-desire subset, is that of being weakly continent; of being an intention to do something which will contribute to at least some degree of preference satisfaction, given one's means-end beliefs. Thus the Principle of Continence, together with the Principle of Rational ATC judgement (or analogues of these) determines what agents have overall reason to do, given their beliefs and desires. The Principle of Weak Continence determines what agents have *pro tanto* reason to do, given their beliefs and desires.

7. Conclusion

In this chapter I have discussed the aspects of rational agency which underpin the account of practical reasons proposed in chapter 2. I have focused on the Principle of Continence as a primary rational ordering principle for intentions. This principle, when combined with the proposed Principle of Rational ATC Judgement, delivers intentions which are extensionally equivalent to those generated by standard decision-theoretic principles, such as the Principle of Maximisation, at the point of action. For this reason, it was suggested that to be rational (and therefore interpretable) agents must generally have intentions, at the time of action, identical to those which follow from the Principles of Rational ATC Judgement and

Continence *and* the Principle of Maximisation. So long as agents generally function in terms of at least one of these sets of principles (or any other extensionally equivalent set of principles), they will be interpretable. Practical rationality does not necessarily reduce to a single rational ordering principle; it may be that there are various independent rational principles, where functioning in terms of at least one set of these is sufficient for rationality.

Despite the arguably simpler picture given by the Principle of Maximisation, I refrain from treating this as the dominant rational ordering principle. Continence might simply be a proxy for maximisation, or it might be of independent rational status. In any case, I focused on the principles of Rational ATC Judgement and Continence in my discussion, as these seem to take psychological priority in the case of human agents. Although we could run the model of rational agency purely in terms of the Principle of Maximisation, there is much to be added by considering the specific agential psychology that we inhabit as humans. At the very least this affords the above conclusions that: (i) the Principle of Maximisation need not be the only, or even dominant, principle of practical rationality; and (ii) to be interpretable, agents need only to function in terms of (at least) one set from within a number of different sets of possible rational ordering principles which are extensionally equivalent at the point of action.

So far as practical reasons are concerned, what is important is that there are specific rational constraints which order agents' intentions at the time for action, such that they can be interpreted (assuming that these intentions are, in general, successfully executed). Agents have overall reason to act in ways which it is rational for them to intend, at the time of acting, given the principles of Rational ATC Judgement and Continence. However, sometimes agents act irrationally, failing to heed the overall balance of reasons. In such cases they must still act on the basis of some practical reason or other; they must perform an action which, at the time of acting, it is at least somewhat rational for them to intend, given the Principle of Weak Continence proposed in section 6. Agents have *pro tanto* reasons to perform each and every action that follows from their desires and means-end beliefs, given the Principle of Weak Continence. Agents' overall reasons support only those

actions which follow from their beliefs and desires, given the Principle of Rational ATC Judgement and the Principle of Continence.

Chapter 4: Objections from Irreducible Normativity

1. Introduction

In chapters 1-3 I introduced the problem of accounting for practical reasons in a naturalistic way and proposed an interpretivist approach to meeting this challenge. My suggestion was that practical reasons can be treated as attitudinal subsets from which certain intentions to act follow, given the constraints of interpretable mental functioning. These constraints were explained in terms of rationality. It was further suggested that the normativity of practical reasons can be reduced to interpretable agents' commitment to function in accordance with rational principles, given that functioning in accordance with these is constitutive of being interpretable.

In this chapter I discuss one kind of objection to this approach. This is the objection that the constraints of interpretable functioning cannot be fully specified without invoking some normative feature of other. As such, the suggestion is that interpretability relies on normativity of a kind which cannot be reduced to interpretable agents' constitutive commitment to function in certain ways. If this commitment can only be specified in terms of something normative, normativity cannot be reductively explained in terms of it.
One species of this kind of objection was set aside in chapter 2. This was the objection that rationality, as invoked by a Davidsonian account of radical interpretation, is normative. Following Timothy Schroeder (2003) I suggested that Davidson's use of rationality in his theory of radical interpretation can be regarded as non-normative, in that it makes use of rationality's categorisation scheme but not its normative force-maker. In this chapter I consider two further objections from irreducible normativity. These are the objections that meaning and preference are (respectively) irreducibly normative.

The first of these objections is owed to Kripke (1982: 37), who argues that for us to be able to attribute determinate meanings to speakers' words and sentences, there must be

semantic norms which prescribe how they are to be used. The second objection is owed to Hurley (1989: chs. 2-6) who argues that for us to be able to attribute determinate preferences to agents in the light of their actions, there must be objective values which normatively govern such preferences.

As may be apparent from my characterisation of these two objections, they have much in common. Hurley takes her problem over the indeterminacy of preference attribution to be analogous to the problem of indeterminacy in the attribution of meaning (ibid: 53). In the next section I discuss the meaning problem. I explain the nature of this problem and discuss a Davidsonian solution to it, in terms of the principle of charity as a substantive constraint on interpretation. I claim that charitable interpretation can be construed non-normatively, such that meaning can be determined, on a Davidsonian approach, without reference to normativity. Meaning attribution requires adhering to certain constraints on interpretation, but there need not be anything normative about these constraints.

My discussion of the meaning problem is relatively brief. Its purpose is twofold. First, I wish two show how a Davidsonian can defend the idea that words and sentences have determinate meanings, without succumbing to the claim that meaning is normative. The aim here is to illustrate what a non-normative species of the Davidsonian approach to meaning would be like. I do not wish to argue that this is the correct understanding of Davidson's position or, indeed, that it provides a successful account of meaning in non-normative terms. Rather, the aim is to show that there is a non-normative option available for the Davidsonian to pursue when accounting for meaning. As such, the Kripkean suggestion that meaning is normative does not necessarily apply, or undermine my account of practical reasons.

The second purpose of my discussion of the meaning problem is to provide a context for the related worry that preference attribution is subject to normative constraints. I discuss this worry in section 3, where I consider Hurley's suggestion that objective values provide the substantive constraints on preference attribution which are required for the attribution of determinate preferences to be possible. I dispute this claim, suggesting that substantive

constraints on preference attribution can be accounted for in non-normative terms. As such, it is hoped that both the objection from meaning and the objection from preference attribution can be met.

2. The Objection from Meaning

Wittgenstein famously argued that there can be no such thing as a private language—a language whose meanings are accessible only to the author of that language (Wittgenstein, 2001: §§243-71). This is because such a language would not admit of the possibility of error; its rules would depend upon the author of the language in such a way that 'whatever is going to seem right to me is right' (ibid: 78). If any use of a word can be deemed right, then no determinate account of what that word means can be given; the word has no meaning.

Kripke (1982) treats this indeterminacy problem as general, suggesting that it applies to any language, public or private. Kripke emphasises the role of Wittgenstein's rule-following considerations in the private language argument and proposes that these considerations apply more generally to all language, public or private (ibid: ch2; Wittgenstein, 2001: §§185-202). For any word to have a meaning there must be some rule which determines whether that word has been correctly used, where this rule must have determinate application across an infinite range of possible cases. Such a rule cannot be provided by anything to do with an agent's uses of a word. No agent's inner states can determine the application of a rule of meaning across an infinite range of possible cases. Nor can an agent's *dispositions* to use a word in a particular way establish what would count as a correct or incorrect use of that word: any use of a word would reflect an agent's dispositions to use that word in a particular way, such that there is no possibility of her using it in error (Kripke, 1982: 22-37). For Kripke, nothing about an agent or her uses of a word can determine how that word ought to be used. Without the possibility of correctness and error, meaning is radically indeterminate. Thus, according to Kripke, there is no fact of the matter concerning what some agent means by any given use of a word (ibid: 21).

The supposed alternative to explaining meaning in terms of the intentions, dispositions, or inner states of some agent is to claim that meaning is fixed by the community in which a word is used (Wittgenstein's own suggestion—Wittgenstein, 2001: §§138-202). However, according to Kripke, this is a sceptical solution to the problem in that it involves accepting that there is no fact of the matter as to what anyone means by the use of some word (Kripke, 1982: 69; 71). There are simply the linguistic conventions of a community in which that word is used, where these determine the word's 'meaning' but not its meaning. Uses of a word are accepted or rejected within a community, but nothing fixes what is acceptable or unacceptable (there are no determinate rules according to which the community operates).

Regardless of this sceptical conclusion, one supposed upshot of Kripke's argument is that meaning is normative (Wikforss, 2001: 203). For a word to have a determinate meaning, it must be possible to distinguish correct uses of that word from erroneous ones, acceptable uses from unacceptable ones. The meaning of a word must establish normative constraints on its proper use. Therefore meaning (if such a thing exists) is normative.

This presents a worry for my account of practical reasons. One feature of interpretation is the attribution of meanings to agents' words in the light of their speech-acts. It is only if we can reliably interpret what agents mean by their words that we can attribute beliefs, desires and the like to them in the light of what they say (Davidson, 1974a). However, the worry for my view of practical reasons is that if the interpretation of speech-acts already involves a commitment to certain normative constraints on meaning, the constitutive constraints of interpretable functioning are already normative in kind. This precludes giving a reductive account of normativity framed in terms of such constraints.

Put another way, the suggestion is that in assigning a meaning to some word when interpreting an utterance, we implicitly acknowledge the existence of certain normative constraints on the use of that word. For instance, suppose that someone describes a letter box as 'red'. Here, we can only interpret them as meaning that the letter box is <u>red</u> if this implicitly entails that the word 'red' *should not* be used to describe things which are green, blue or yellow too. Thus in interpreting 'red' as meaning <u>red</u>, we are committed to the

existence of certain normative constraints on the use of that word. Otherwise, our interpretations would be radically indeterminate. Interpreting 'red' as meaning red would be null without this involving normative constraints on use, in that without such constraints this interpretation would not rule out the possibility of 'red' being (non-erroneously) used to describe non-red things. Thus linguistic interpretation implicitly involves the existence of normative constraints on language; such constraints are part and parcel of a word's having some particular meaning. If interpretation depends on the existence of normative constraints, it cannot be invoked in the explanation of what normativity is. This, at least, is the worry for my view of practical reasons.

One way of avoiding this worry is to propose a non-normative account of correctness and error in language. For example, one might propose that correctness is not about how one *should* use words, but about truth. Thus:

> The notion of semantic correctness is non-normative in just this sense: [t]hat an application of *e* [some expression] is correct, does not entail that it ought to be made, and, conversely, incorrect applications do not immediately imply that *S* [a speaker] has violated any semantic prescription. If 'green' means *green* then *S* applying it to a red object implies that her statement is false, but it does not thereby follow that she has failed to do what she ought, semantically, to do (Glüer and Wikforss, 2009: §2.1.1).

The suggestion here is that for some word (or expression) to have a meaning, it must be determinate when that word can be used within a true sentence (or when that expression can be truly uttered). This is not a matter of normativity at all, it is a matter of there being determinate conditions under which sentences are true. Thus the issue of semantic correctness is, on this suggestion, to do with the existence of determinate truth conditions for a sentence, and not with what a speaker should do with her words.

This suggestion might be resisted, by claiming that the truth conditions for an utterance cannot be spelled out without reference to that utterance's meaning. Thus accounting for

semantic correctness in terms of truth is a non-starter. Without knowing what an utterance means, we cannot know when it can be truly uttered or not. However, this objection does not apply to a Davidsonian account of meaning, which is framed in terms of truth (Davidson, 1967). If the meaning of a sentence is given by its truth conditions, these must be independent of the meaning of that sentence. Thus it appears to be open for the Davidsonian to claim that the interpretation of language only requires some method of distinguishing between truth and falsehood; it does not require the acceptance of any normative constraints on how words are to be used. To interpret someone as meaning red by 'red' we are only committed to the attribution of falsehood to certain sentences (say, ones in which non-red objects are described as 'red'), and not to their being any norms which prescribe that 'red' is only to be used to describe red things.[48]

For the purposes of this thesis I shall assume that this suggestion is correct. I do not take it to have been adequately established. However, I do take it to be at least open for the Davidsonian to claim that semantic correctness and error can be accounted for in terms of truth. Plausibly, adopting this kind of approach allows the worry that meaning is normative to be avoided (worries over the normativity of a sentence's truth-conditions notwithstanding).[49] However, this does not entirely resolve the issue of meaning scepticism. Even if it is proposed that semantic correctness can be accounted for in terms of truth, it still needs to be shown that words can have a determinate meaning, where whatever determines the meaning of a word must also determine the truth conditions for an infinite variety of sentences involving that word (given the above characterisation of semantic correctness in terms of truth).

[48] There is a further question of what determines the truth conditions which it is appropriate to apply. On a Davidsonian view this is determined by charity in interpretation, as discussed below.

[49] For further discussion of the allegation that meaning is normative see Boghossian (1989); Glüer and Wikforss (2009); Wikforss (2001). For an argument that Davidson's philosophy of language is non-normative see Glüer (2001).

On a Davidsonian approach, it is a speaker's uses of a word which determine its meaning, as constrained by charitable interpretation (Davidson, 1973; 1974a). The principle of charity is an essential feature of Davidson's account of radical interpretation, which sets out to establish what is necessary for a speaker to be interpreted from scratch and, therefore, what is necessary for her to be interpretable at all (Davidson, 1973). For the radical interpreter, neither the meanings of a speaker's utterances, nor the beliefs (or other attitudes) that they express, can be known in advance of interpretation. Rather, the purpose of radical interpretation is to establish both what a speaker means, and to assign beliefs, desires and the like to her which are consistent with the meanings of her sentences.

So far as meaning is concerned, radical interpretation (according to Davidson) involves providing Tarskian T-sentences: sentences in a meta-language which define the truth conditions for sentences in an object language. Thus, on a Davidsonian account, the meaning of a sentence is given by the truth conditions for that sentence, as encapsulated by a relevant T-sentence. For instance, the meaning of 'Snow is white' is provided by the T-sentence ' "Snow is white" is true if and only if snow is white' (Davidson, 1974b: 194). The meanings of words are determined by their relation to true sentences in which they appear.

There are many issues which arise in relation to this kind of theory, not least of which is finding a way of restricting the kinds of T-sentences which are relevant to the meaning of a word (' "Snow is white" is true if and only if grass is green' may be a true T-sentence, but it seems bizarre to suggest that it can provide the meaning of 'Snow is white' (Davidson, 1967: 25-27)). However, I will not discuss the plausibility of defining the meaning of a word in truth-theoretic terms here.[50] I will simply assume that radical interpretation is able to provide the meanings of sentences by establishing the conditions under which they can be held true, and that the meanings of particular words derives from their role in such sentences. These are controversial assumptions, but they are part and parcel of the general Davidsonian approach that I adopt in this thesis. The important question, for present purposes, concerns the interpretive process by which a sentence's truth conditions are

[50] For discussion of this approach see LePore and Ludwig (2007).

established (the process by which a radical interpreter arrives at an appropriate T-sentence, given some utterance/set of utterances).

The problem here, as Davidson points out, is that belief and meaning are interdependent. Thus:

> A speaker who holds a sentence to be true on an occasion does so in part because of what he means, or would mean, by an utterance of that sentence, and in part because of what he believes. If all we have to go on is the fact of honest utterance, we cannot infer the belief without knowing the meaning, and we have no chance of inferring the meaning without the belief (Davidson, 1974a: 142).

To determine belief, we need meaning; to determine meaning, we need belief. How are we to determine either in the absence of both? The answer is that we need a method of interpretation which imposes constraints on both structure and content. For Davidson, these constraints are provided by the principle of charity, which requires that we find the majority of a speaker's beliefs to be true. Doing so requires that we maximise agreement between ourselves and those who we interpret; that we treat them as sharing our basic logical presumptions and as largely agreeing with us on matters of truth. Thus: 'if we cannot find a way to interpret the utterances and other behaviour of a creature as revealing a set of beliefs largely consistent and true by our own standards, we have no reason to count that creature as rational, as having beliefs, or as saying anything' (Davidson, 1973: 137).

So, the beliefs and logic of the interpreter constrain the interpretations that she is able to attribute to those who she interprets. This rules out the possibility of radical indeterminacy: the kinds of things that a speaker means by her words, and the kinds of beliefs that she has, must be largely similar to what we mean and what we believe. If there were too much disagreement, the possibility of attributing meaning and belief would disappear altogether (Davidson, 1973: 137; 1975: 168-9). Thus for Davidson, the problem of radical indeterminacy over meaning is solved by there being constraints on the meanings which can be determined by a speaker's uses of a word. These constraints are provided by the requirement that we

find a speaker's beliefs largely consistent and coherent with the truth (as defined by our own beliefs and notion of consistency).

One potential worry for my account of practical reasons here is that charity is a norm of interpretation. If so, (radical) interpretation is subject to normative constraints, in the sense that there are certain ways in which it ought to be carried out (i.e. ways which maximise agreement). If interpretation is normative, normativity cannot be explained in terms of interpretation. However, this worry can be avoided if the principle of charity is regarded as defining a constitutive constraint on interpretation. It is not that we *ought* to interpret speakers as largely true and consistent. Rather it is that to be able to provide any kind of interpretation of an agent at all, we must find them to be largely true and consistent, given our own beliefs and notion of consistency. As suggested above, without following a procedure of maximising agreement between ourselves and those who we interpret, we have no basis for determining what they believe and mean; treating others as believing and meaning (roughly) the same things as us is necessary for us to be able to assign determinate content to their attitudes and utterances, in light of their (speech) behaviour.

Insofar as the notions of truth and consistency can be non-normatively specified (as I assume is at least a possibility, on a Davidsonian approach), the requirement for charity can be regarded as a non-normative constraint on interpretation.[51] So, it appears to be at least possible to avoid the objection from the normativity of meaning, given a Davidsonian approach.

However, a second worry is that charity does not place a sufficient constraint on the contents which can be attributed to speakers' beliefs, or on the meanings which can be ascribed to their utterances by an interpreter. This worry arises because it seems possible, in

[51] I have discussed the availability of a non-normative notion of consistency in chapter 2, section 4. I assume that, for the Davidsonian, truth can be characterised along Tarskian lines, in terms of a logical operator. The availability of this kind of formal approach to truth for Davidson is noted by Engel (2001: 39).

cases where we may be inclined to ascribed a false belief to a speaker, to simply alter the meaning attributed to some utterance of hers (by altering the T-sentence which we attach to it), such that the belief which it expresses comes out to be true. For example, suppose that Jane (sincerely) says that 68 plus 57 equals 5. In this case, we *could* interpret her as meaning plus by 'plus' and attribute her a false belief. But we could also interpret her as meaning quus by 'plus' and find her belief to be true.[52] The problem is that charity seems to require something along the lines of the latter interpretation, rather than the former one. This is because charitable interpretation involves finding a speaker to be maximally true and consistent, by our own lights. Thus if, on some occasion, Jane insists that '68 plus 57 equals 5', we seem bound by charity to assign some non-standard meaning to her utterance (at least in that instance) such as treating 'plus' as meaning quus. Doing so allows us to attribute Jane a true belief (e.g. that 68 quus 57 equals 5).

This raises two problems for the Davidsonian view of belief and meaning. The first is that it seems to show that charity, as the sole substantive constraint on interpretation, has absurd results, such as that we must assign non-standard meanings to speakers' utterances in cases where doing so will maximise their degree of true belief. The second is that it seems to show that, without some further substantive constraint on the kinds of meanings which we can attribute to speakers' utterances (that is, on the kinds of truth conditions which we can apply when interpreting them), what they mean and believe is radically indeterminate. For there will be any number of non-standard interpretations according to which we can find a sentence to be true.

The response to this worry is to point out that it shows a misunderstanding of charity as a method of interpretation. Charitable interpretation involves more than just finding an interpreted speaker's beliefs to be maximally true and consistent. It also involves applying our own concepts, and interpreting others by reference to the beliefs that we hold given

[52] Kripke defines quaddition (in the first instance) as a function which is identical to addition for all calculations involving numbers below 57; for calculations involving numbers above 57, quaddition generates the answer 5 (Kripke, 1982: 9).

those concepts. It is only by ignoring the central role of an interpreter's concepts in interpretation that we might be tempted to think that attributing beliefs involving strange contents, such as that of quaddition, is a viable method of charitable interpretation.

Not that we *cannot* attribute such non-standard beliefs in certain particular instances. We can define what quaddition is, for example, using more familiar concepts, and we can attribute a belief that 68 quus 57 equals 5, say, to anyone who says so (at least under certain conditions). The point is not that we can never attribute non-standard meanings to speakers' utterances. Rather, it is that we cannot find a person to inhabit an entirely alien conceptual scheme (i.e. to hold beliefs which generally involve alien concepts, such as quaddition). Attempting to do so would involve losing any grip that we have on what the contents of such a person's beliefs are. As Hurley puts it: 'before reaching anything that would count as a radically different conceptual scheme, we lose grip on the very idea of a conceptual scheme, on the very notion of belief' (Hurley, 1989: 52).

So, to interpret someone, we must generally apply familiar concepts and find them to have largely true beliefs involving these concepts. Any interpretation involving a non-standard concept is at odds with this requirement, and therefore requires appropriate interpretive support (e.g. a determinate, somewhat limited context which supports the application of some particular non-standard concept, where this concept can be defined in terms of other, more familiar ones—perhaps a group of philosophers playing a quaddition game, for example). More generally, charitable interpretation involves finding others to hold largely true beliefs, where the beliefs that they hold must generally have the same kinds of contents as the beliefs that we hold (Davidson, 1974b: 197-8).

Once the role of a shared conceptual scheme is recognised in the process of charitable interpretation, the worries that charity requires absurd interpretations and that it allows for radical indeterminacy seem to be avoidable. There is then a further question about what fixes the shared conceptual scheme that speakers inhabit. On a Davidsonian approach, this is a built in feature of interpretable creatures' existence within a world, given that for Davidson there is no distinction to be drawn between conceptual scheme and empirical

content (Davidson, 1974b: 187-93; 197-8). In fact, for Davidson, there is no such thing as a conceptual scheme in the (traditional) sense of a scheme of concepts which organises the world around us in some way (ibid: 187-93). Thus Davidson claims that that there is no 'theory-neutral reality' that we impose a conceptual order upon (ibid: 195).[53] The contents of our beliefs are built into our existence within the world, as it were. If conceptual sheme and empirical content cannot be distinguished then it seems impossible for there to be creatures who have radically different conceptual schemes, given the shared reality that they inhabit (ibid: 194-5).[54]

Returning to the central question of meaning, a Davidsonian approach seems able to avoid the worry that meaning is normative. Insofar as a non-normative understanding of semantic correctness is available (say, in terms of truth), Kripke's contention that meaning must establish normative constraints on use can be avoided. Moreover, insofar as charitable interpretation can be seen to constrain what speakers can mean by any given use of a word, the worry that meaning is radically indeterminate, given use, can also be avoided (at least on a Davidsonian account of meaning). Assuming that the principle of charity is not a normative principle but, rather, a non-normative constraint on interpretation, this solution to the worry that meaning is radically indeterminate does not involve the introduction of an unwanted normative element into my theory of practical reasons. It simply involves the constraint that we apply our own beliefs and concepts when attributing meanings and beliefs to others in light of their utterances. Thus we might conclude that, on a Davidsonian account, meaning and its attribution need not be normative. I take this conclusion to be sufficient for the purposes of my project, which aims to develop a plausible naturalistic approach to practical reasons but not to establish that this approach is correct.

[53] For discussions of this issue see Child (1994); McDowell (2001). See also Davidson's response to McDowell (Davidson, 2001).

[54] This is disputed by Hacker (1996).

3. The Objection from Preference

Hurley (1989: chs. 3-5) takes the worry that meaning is radically indeterminate to apply more generally to intentional content. Just as there is the worry that meaning may be radically indeterminate, given linguistic behaviour, so there is the more general worry that intentional content (both the meanings of words and the contents of agents' attitudes) may be radically indeterminate, given behaviour in general. In the cases of meaning ascription and belief attribution, the solution comes (on a Davidsonian approach) from the principle of charity, which provides a substantive constraint on interpretation. What agents mean by their words is constrained by the requirement that the beliefs such words are used to express must come out to be largely true. However, the principle of charity as it applies to belief and meaning does not provide a substantive constraint on the attribution of preferences (preferences are not truth-apt).

This leads to a problem of radical indeterminacy in the attribution of preferences. This can be illustrated by the following type of example. Suppose that I am playing cricket and that my team mate (Ann) edges a ball to first slip for an easy catch. Two possible interpretations of her behaviour are: (i) that she wanted to keep her wicket, and mistakenly believed that she should play a shot; (ii) that she wanted to lose her wicket (so that she could return to the pavilion for an early lunch) and believed that she could get out by edging the ball to slip. The question is, what determines which interpretation is correct and, therefore, which (if either) preference Ann has?

Suppose that we take the formal constraints of decision theory to be exhaustive of the constraints on preference attribution. The preferences that we attribute to agents must be such that their behaviour can be seen to maximise expected levels of preference satisfaction, within certain other formal constraints (such as transitivity). The problem is that such constraints admit of radical indeterminacy. Hurley (1989: 59) gives the following example. Suppose that I prefer apples to oranges and oranges to pears; transitivity requires that I also prefer apples to pears. What, then, if on some occasion I choose to eat a pear

over an apple? Have I violated transitivity? Not necessarily, for perhaps my preference is actually for green apples over oranges, for oranges over pears, but for pears over red apples. So long as my behaviour is consistent with the ascription of this preference (e.g. that I chose a green apple over an orange, an orange over a pear, but a pear over a red apple), I can be found to comply with the transitivity requirement.

However, what if my choices are not consistent with this preference ordering? Have I violated transitivity in that case? Again, not necessarily. Perhaps 'I chose an apple with several leaves attached over an orange, but a pear over an apple with no leaves' (ibid: 59). Or else, 'perhaps I made one choice in a shop which also had bananas, and another in a shop which didn't; I chose an apple-in-the-presence-of-bananas over an orange, but a pear over an apple-not-in-the-presence-of-bananas' (ibid: 59). There are potentially infinite distinctions which might be drawn between my different choices, such that my preferences can always be found to be transitive. Thus transitivity does not sufficiently constrain the preferences that I might have, given my choice of fruit, for any determinate preference ordering to be attributed to me. The upshot of this, and other similar examples involving other decision-theoretic principles is that, given only the formal constraints of decision theory, it is radically indeterminate which preferences one has.

Thus Hurley's proposal is that, in addition to the formal constraints provided by decision theory, there must also be certain substantive constraints on the attribution of preferences to agents (ibid: ch.4; ch.5 §1). That is, there must be constraints on preference attribution which go beyond the requirement that any preferences attributed must fit with the behavioural evidence, given the formal constraints of decision theory. There must also be substantive constraints on the kinds of preferences which are eligible to be attributed to agents, in light of their actions.

The notion of eligibility here derives from the metasemantic debate over the inscrutability of reference (see, for example, Lewis, 1984; Williams, 2007). To avoid the worry that, on an interpretationist metasemantics, any word can be treated as referring to anything, Lewis has suggested that constraints other than fitting with the linguistic data must apply to the

eligibility of interpretations (for instance, that interpretations must be simple) (Lewis, 1984: esp. 224-9). Thus the notion of eligibility is meant to capture the idea that an account of some term's reference needs to do more than just fit with the linguistic data. An eligible interpretation of a term's reference must also conform to certain other constraints which rule out its having some crazy-referent (Williams, 2007).

Eligibility might be thought to be a normative notion—a notion which involves there being certain interpretations which we ought to ascribe to agents and others that we ought not. However, the idea of something's being the best, or most eligible, interpretation of some term's reference (or of some agent's preferences, or whatever) might also be understood in terms of its being the interpretation with the greatest likelihood of being true, or else the interpretation which most closely accords with the constitutive constraints which apply to interpretation (recalling that such constraints operate at a holistic level). I do not wish to discuss the normative status of eligibility here. I will simply suggest that it is at least *prima facie* plausible to regard the eligibility of interpretations as a matter of establishing criteria which bear on an interpretation's likelihood of being true, or else on its qualifying as an interpretation to begin with.

Hurley's suggestion is that the eligibility of preference attributions must be more than just a matter of fitting with the data, while falling within the constraints of formal decision theory. As illustrated above, there will be an infinite number of possible preference attributions which meet these criteria. For preference attributions to avoid the problem of radical indeterminacy, substantive constraints on the eligibility of preferences are required.

Hurley frames this suggestion specifically in terms of the eligibility of distinctions which can feature in agents' preferences. Some distinctions are more eligible to feature in agents' preferences than others. For example, a preference for love over money, we might suppose, is more eligible than a preference for love over money except on alternate Wednesdays.

The question is, what is it that explains which distinctions are (more) eligible to feature in preference attributions than others? The answer cannot derive from agents' actual

preferences themselves. The preferences that agents hold are downstream of interpretation such that they cannot ground substantive constraints on preference attribution.

Hurley also argues that the answer cannot derive from extended preferences ('preference[s] between being in one person's objective situation with all his subjective features, including his preferences, and being in another person's objective situation with all his subjective features, including preferences' (Hurley, 1989: 109)). Extended preferences remove all points of difference between agents, treating these as the objects of preference. Once all such differences have been abstracted away from, there are no differences left which might lead people to have different extended preference orderings. What we are left with might be termed 'featureless bare egos' which, insofar as they are identical in kind, will hold the same ordering of extended preferences between the different extended options available (ibid: 112). Given this shared extended preference ordering, it might be suggested that we can explain substantive constraints on the eligibility of preferences in terms of a requirement that any preferences which agents hold are broadly consistent with the single extended preference ordering that all agents would hold if we abstracted away from all points of individuality.

Ignoring the problem of whether the concept of a featureless bare ego is even intelligible, this proposal does not help. This is because treating all possible differences between agents as objects of extended preference does not suitably constrain the kinds of distinctions which can be made between the extended alternatives available. Hence:

> We cannot incorporate the sources of concern with all possible distinctions among alternatives into the extended alternatives, leaving nothing to constrain the eligibility of the contents of the extended preference ordering, for then the extended preference ordering will be completely indeterminate (ibid: 112).

I take the idea here to be that there must be some constraint on the kinds of distinctions which can be made between the extended options available. If any distinction at all can be made between the extended options available, then any of an infinity of extended

preference orderings will be compatible with the ranking of lives which emerges when all differences between individuals are abstracted away from. Life A might be ranked above life B because of a preference for happiness over misery, or because of a preference for schmappiness (happiness on all but the 2011th day of a life) over misery, or because of a preference for any other of a potential infinity of categories which might distinguish these lives. To avoid radical indeterminacy at the level of extended preferences, the kinds of distinctions which are eligible to be drawn between the extended options available must themselves be substantively constrained.

However, when all possible differences between agents are treated as objects of preference, there is nothing left about agents themselves which could substantively constrain the kinds of distinctions between the available extended options that they might make. The 'sources of concern with all possible distinctions among alternatives' have been removed. As such, there is nothing about the featureless bare egos which remain after the process of abstraction to determine the kinds of distinctions which they might make between the extended alternatives available. Thus the supposed single extended preference ordering which is obtained by abstracting away from agents' specific situations remains radically indeterminate, in the absence of something to substantively constrain the kinds of distinctions which are eligible to be drawn with respect to this ordering.

Hurley's proposal is that substantive constraints on (extended) preference are provided by objective values (ibid: 118). That is, the kinds of distinctions which are eligible to feature in (extended) preference attributions are constrained by their relation to value. So, for instance, a preference for a long life unless one is in pain is more eligible than a preference for a-long-life-unless-one-was-born-on-a-Monday-and-in-that-case-life-for-an-even-number-of-weeks (to use one of Hurley's examples—ibid: 122). This is because avoiding pain has value, whereas living-for-an-even-number-of-weeks-if-one-was-born-on-a-Monday does not.

On Hurley's view, value attaches to one's form of life: it is with respect to one's (human) nature that certain things have value. Thus, in abstracting away from individuals' specific

circumstances, we are not left with featureless bare egos, but rather with a theory of human nature, where this is partly a theory of value. This leads Hurley to suggest that 'what emerges as part of our theory of human nature is something like a theory of primary goods... Distinctions drawn by reference to these goods provide eligible distinctions' (Hurley, 1989: 115).

The role of (human) nature in an account of value means that, on Hurley's proposal, the constraint of preference by value has both an *a priori* component and an *a posteriori* one. According to Hurley, it is an *a priori* truth that eligible distinctions must derive from the ordering of values (in the absence of anything about agents' preferences themselves which can provide the required substantive constraints on eligibility). Simultaneously, what is of value is partly an *a posteriori* matter of our (human) nature. Further, Hurley's claim that value is partly determined by (human) nature leads her to propose that value and preference are interdependent. What is valuable depends (in part) on the kind of creature that one is. The kind of creature that one is is partly a matter of one's being disposed to make certain kinds of choices, where these can be seen as a reflection of preference. Thus Hurley writes: 'I do not put forward the Platonistic claim that values are prior to and independent of preferences, but merely deny that preferences are prior to and independent of values' (ibid: 57).

So, for Hurley, one's having determinate preferences, given choice, requires value as a substantive constraint on the eligibility of distinctions. But what is valuable for a given creature is partly determined by the kinds of choices that creatures of that kind make—by creatures of a certain kind having preferences for certain kinds of outcomes over others. Thus: 'constraints on interpretation are necessary, but at no point is it suggested that decision theory could possibly do without the brute input of activity to be interpreted, taken at least *prima facie* to be expressive of preference' (ibid: 94). Preference is constrained by value; value, in part, depends on how it is in one's (human) nature to respond to one's environment.

Even though Hurley is not adopting a Platonistic account of value, she still holds that value is a distinct normative kind (that it does not reduce to preference) *and* that it acts as a substantive constraint on preference. This presents a distinctive worry for my account of practical reasons, which attempts to explain the normativity of practical reasons in terms of the constitutive constraints of interpretable functioning. If preferences cannot be attributed independently of value, and value is a distinct normative kind, then value places a normative constraint on interpretation. As such, normativity (including the normativity of practical reasons) cannot be reductively explained in terms of the constitutive requirements of interpretable functioning.

How might this worry be avoided? One suggestion, inspired by Lewis' conception of eligibility, is that the eligibility of preference attributions is determined by their degree of simplicity. The simpler the distinctions which feature in potential preference attributions are, the more eligible these preferences are to be attributed. Thus, for example, a preference for apples over oranges is more eligible to be attributed than a preference for apples-in-the-presence-of-bananas over oranges because it is simpler.

On this proposal, relatively more eligible interpretations are those which involve attributing relatively simpler preferences, given the behavioural evidence and subject to the formal requirements of decision theory, and the formal (logical) and substantive (truth-preserving) requirements of belief attribution. Any preference involving a relatively more complex distinction can be attributed only if it reduces error with respect to some other interpretive constraint. However, to the extent that additional complexity in preference attributions is to be avoided just as much as breaches of other interpretive constraints, sometimes the attribution of a simpler preference will be more appropriate, despite this involving the attribution of formal inconsistency or erroneous belief. Interpretation is a matter of ascribing particular attitudes and meanings which balance the various interpretive constraints that apply to agents on the whole.

An immediate question which arises in relation to this suggestion is: what does simplicity mean in this context? Lewis' proposal, regarding the inscrutability of reference, is that

simplicity is (partly) a matter of syntactic structure (Lewis, 1984: 227-9; Williams, 2007: 15). The fewer syntactic connectives within a theory, the simpler it is. However, lack of syntactic complexity alone is not enough to account for simplicity. There is also a question concerning the level at which some component of a theory is syntactically simple. For instance, a theory which refers to 'imaginary trees' may be more syntactically simple on the surface than one which refers to certain complex physical structures which constitute real trees. However, if a theory involving imaginary trees were to be stated in its most primitive (or physically fundamental) terms—terms which refer to the physics underpinning both the brain states involved in tree imaginings and to the physical structure of trees—that theory would be more syntactically complex than a theory which only referred to the physical structure of trees. Lewis' proposal is that the eligibility of a theory is given by its syntactic structure when stated in its most primitive (i.e. fundamental physical) terms (1984: 228).

This is a very brief gloss of what is a very complicated issue. However, it is not clear that Lewis' notion of simplicity is appropriate when it comes to the kinds of distinctions which are eligible to feature in preference attributions. The syntactic complexity of any given distinction, when its objects are spelled out in fundamental physical terms, might not match up in any way with our intuitive grasp of whether some distinction is eligible or not. For instance, a preference for sleep-when-one-is-not-tired-but-has-to-fly-to-Australia-at-5am-the-next-morning-to-give-a-talk-on-the-nature-of-physical-exhaustion-in-*homo-sapiens* seems to be more eligible for attribution than a preference for sleep-when-one-is-not-tired over sleep-when-one-is-tired. However, the latter preference seems likely to be more syntactically simple when its contents are spelled out in fundamental physical terms than the former, which involves far more physical entities and categories which require elucidation, at the very least. Thus some other notion of simplicity seems to be required for an interpretation of eligibility in terms of simplicity to get off the ground. In the absence of any obvious candidates, I set this approach aside.

A second suggestion is to claim that substantive constraints on the eligibility of preferences are simply brute. That is, perhaps certain distinctions *just are* more eligible to feature in preference attributions than others. Thus it might be a brute fact about interpretation that a

preference for a long life is more eligible for attribution than a preference for a-long-life-unless-one-was-born-on-Monday-and-in-that-case-life-for-an-even-number-of-weeks. To have preferences one must, on the whole, be subject to distinctions which have a relatively high degree of brute eligibility.

One problem with this kind of brute eligibility response is that it seems to require a capacity to detect which distinctions are brutely more eligible than others. Preferring chocolate to cheese, for example, seems to require an awareness that a distinction between chocolate and cheese is more eligible than a distinction, say, between chocolate and cheese at high tide. This kind of awareness, one might think, is entirely mysterious given the suggestion that eligibility is brute. Faced with competing interpretations, how could we know that attributing either one of them involved drawing a distinction which is, *quite simply*, less eligible to be drawn? What does our capacity to detect degrees of eligibility consist in, if degrees of eligibility are just a brute feature of distinctions which might feature in preference attributions?

One response to this objection is to suggest that Hurley's view of eligibility is a partner in guilt. After all on her view, some method of detecting value is required if agents are to be interpretable as having preferences. If this method of value detection is entirely mysterious, Hurley's view faces the same epistemic problem as the brute eligibility view. If some positive proposal for detecting value is available, perhaps this proposal can be adapted to apply to the detection of brute eligibility.

The method for 'detecting' value, on Hurley's view, derives from the mind's being constitutively constrained by the world. Thus, for Hurley, it is simply in one's nature to be constrained by certain values (i.e. to form preferences and make choices which are informed by these values). This is not a matter of accessing value *qua* platonic form, and then adapting one's preferences accordingly. It is a matter being sensitive to the evaluative features of one's environment (i.e. its reason-giving features—Hurley, 1989: 99). This idea might seem a little opaque. However, I take the general idea here to be that sensitivity to value is to be explained by the value-ladenness of preference, which is part of the more

general world-ladenness of mind (ibid: 92-4).[55] Thus 'the contents of the mind are [not] given in some way independently of the subject's environment and his relations to it' (ibid: 94). Our mental life is infused by the world in which we live. On Hurley's view, this world is in part evaluative. Thus to know what is valuable just is part of what it is to have a mental life. Such knowledge falls out of there being 'constitutive constraints on the relations in which the mind stands to the world, where the direction of individuation may be from world to mind, but not wholly *vice versa*' (ibid: 94). In short, the contents of the mind are constitutively constrained by the world in which it exists; value is a feature of that world, and constrains preference in virtue of that fact.

Can this suggestion be adapted to apply to the suggestion that eligibility is brute? Not entirely, for the brute eligibility of distinctions is not a feature of the world as such. Rather, it must be a constitutive feature of interpretable functioning; interpretability, so the suggestion might go, constitutively involves drawing certain kinds of distinctions over others.

Thus, perhaps it is part of any interpretable creature's nature to invoke certain kinds of distinctions rather than others. This is not because such distinctions line up with certain features of her natural environment (i.e. value), where it is constitutive of her nature to be influenced by these features (Hurley's view). Rather, it is that applying such distinctions is a brute feature of an interpretable creature's mode of functioning. Just as a capacity to speak a language might constitutively involve having certain grammatical tendencies, so a capacity to function in an interpretable way (as well as a capacity to interpret others) might constitutively involve having a tendency to draw certain kinds of distinctions over others. Such distinctions, the suggestion goes, are brutely more eligible to be drawn with respect to preference attribution. Thus one might suppose that substantive constraints on the eligibility of preferences do not necessarily derive from objective values; they might derive from brute facts about the nature of interpretable functioning (i.e. that interpretable

[55] Hurley follows Davidson here in collapsing the scheme/content distinction.

functioning necessarily involves being subject to certain kinds of distinctions when it comes to preference).

A second objection to the brute eligibility suggestion, *qua* non-normative hypothesis, is that some of the distinctions which are eligible to feature in preference attributions seem to be essentially evaluative in kind. Thus a preference for brave over cowardly actions seems to be eligible, even though bravery and cowardice are essentially evaluative categories. This point is made by Hurley, as a part of her argument for objective values as constraints on the eligibility of preference attributions (ibid: 102-5; 110). Given that some of the distinctions which are (intuitively) eligible with respect to preference attributions involve essentially evaluative categories, an account of the eligibility of these distinctions couched in non-evaluative terms seems unavailable.

This objection is a significant threat to the suggestion that eligibility is brute, at least insofar as it is taken to support a non-normative account of substantive constraints on preference. Avoiding normative constraints on preference by retreating to the suggestion that eligibility is brute cannot succeed if certain eligible distinctions are essentially evaluative in kind.

One possible response here is to suggest that not all distinctions that people make, and which appear to be (or have appeared to be) eligible, are good ones. Thus some people distinguish between miracles and everyday occurrences, others have distinguished between special substances like phlogiston and more familiar ones like wood. Neither of these distinctions are good ones, in the sense that neither of them involves a true separation between one kind of thing and another. Yet they appear to be distinctions which we can attribute to people in the context of belief. We can attribute a belief in the miracle of Jesus' resurrection, or in the existence of phlogiston. Just as such beliefs are (I assume) false, so we might hold that preferences for bravery over cowardice, or for fairness over unfairness are, strictly speaking, empty, in that nothing can instantiate the kinds of properties which they invoke.

If there is no genuine distinction between bravery and cowardice, or between fair outcomes and unfair ones, then any preferences which are attributed with respect to such things will be null, just as any beliefs based on spurious distinctions will necessarily be false. Yet, just as we can attribute false beliefs based on spurious distinctions, so we can attribute empty preferences based on such distinctions too. In both cases all of the hallmarks of belief and preference appear to be met, but the beliefs in question are necessarily false and the preferences in question are necessarily empty (i.e. the preferences which an agent actually holds in this domain are either confused, related to something different, or else entirely lacking).

This kind of response may be tenable, but I am not in favour of it. This is because it involves the attribution of both a high degree of erroneous belief, and of empty preference. I am inclined to think that notions like bravery, cowardice, fairness and unfairness are not entirely spurious, unlike those of miracles and phlogiston. People can have true beliefs about brave and cowardly actions, I think, and they can equally have preferences for, say, exhibitions bravery over displays of cowardice.

An alternative response is to suggest that distinctions between, say, bravery and cowardice are eligible to feature in preference attributions, but that substantive constraints on the eligibility of preferences are not, thereby, evaluative in kind. One way of pursuing this kind of response would be to suggest that essentially evaluative categorisations, such as those of bravery and cowardice, supervene on non-evaluative ones. Thus, for example, one might invoke the traditionally non-cognitivist idea that thick evaluative concepts have both a descriptive component and an evaluative (affective, on non-cognitivist approaches) component (see, for example, Burton, 1992). If the descriptive component of, say, the concept of a brave action can be distinguished from its evaluative component, then it is open for the proponent of brute eligibility to claim that it is simply a brute fact about interpretation that distinctions between actions drawn in terms of these descriptive features are relatively more eligible to feature in preference attributions, regardless of their

(supervening) evaluative component.[56] In this way, a distinction between brave and cowardly acts can be treated as relatively more eligible than certain other distinctions, without reference to bravery or cowardice appearing in the explanation of why this is so.

In figurative terms, one might suggest that the descriptive component of evaluative distinctions carries the interpretive baggage, while the evaluative component comes along for the ride. More literally, the suggestion is that having an evaluative preference involves distinguishing between acts or outcomes with certain descriptive features (where such distinctions have brute eligibility), while simultaneously attaching some kind of evaluative significance to this distinction.

One way in which this evaluative significance might be attached is through affect. This kind of non-cognitive approach would involve explaining value in terms of our having preferences/affective states which relate to certain kinds of situations (actions, or whatever), as subsumed under certain descriptions. An alternative suggestion would be that the evaluative component of evaluative distinctions is attached by its simply being a brute, substantive constraint on interpretation that we generally prefer situations (or actions) to possess the descriptive features in question. In other words, the suggestion is not that value is to be explained non-cognitively, in terms of preference/affect. Rather, it is that value *just is* the fact of our being substantively constrained to prefer certain situations, as subsumed under certain descriptions. Thus far from value placing a substantive constraint on the eligibility of preferences, it might be suggested that certain brute, substantive constraints on the eligibility of preferences provide a reduction base for value.

This is a controversial suggestion. Like several other controversial suggestions which I have either made, or relied upon, in this thesis, I do not intend to try to present a positive argument for it here. The argument of this thesis is that it is *possible* to account for practical

[56] Some non-cognitivists maintain that there is no consistent evaluative dimension to so-called 'thick concepts' (e.g. Blackburn, 1992). If this is right then perhaps evaluative preferences are relatively ineligible, or else must be given a situation-specific treatment.

reasons, including their normative force, in terms of the constitutive constraints of interpretable functioning. In order for this argument to succeed, I must show that an account of practical reasons, including their normativity, *can* be given in terms of the constitutive constraints of interpretable functioning. This can only succeed if these constraints are *not necessarily* normative in kind. Thus to establish this possibility, I am required to show how a non-normative account of the constraints of interpretable functioning can be given, but not that this account is successful. Although I hope that my suggested account has a certain degree of plausibility, it is beyond the scope of this thesis (whose purpose is primarily exploratory) to attempt to prove the various claims on which the theory of practical reasons that I propose depends.

Nevertheless, I am required to defend any proposals on which my theory relies from certain significant objections. One objection to the suggestion that it is the descriptive, rather than the evaluative, features of certain distinctions which render them eligible with respect to preference attribution is that some eligible distinctions are purely evaluative in kind. For instance, suppose that I prefer good outcomes to bad ones. On the surface, such a preference appears to be relatively eligible for ascription. However, in that it is purely evaluative, it might be argued that it cannot piggyback on a distinction between different descriptive categories.

My response to this objection is to draw on Hurley's non-centralist hypothesis that thin evaluative features are to be explained by thick ones, and not the other way around (Hurley, 1989: 27-9). If 'goodness' is a function of thick properties, such as fairness, kindness, justice and the like, then the proponent of brute eligibility can argue that distinctions between good and bad outcomes are eligible in virtue of their connection to the descriptive features of thick evaluative distinctions (supposing, as I have above, that these can be distinguished from their evaluative features). Thus it is possible to claim that the distinction between good and bad outcomes is not, despite appearances, a purely evaluative distinction even though the terms which it involves do not have any specific descriptive meaning.

So, it seems that the worry that preference attribution requires normative constraints on the eligibility of preferences may be avoidable. Perhaps it will be suggested that this worry could have been avoided more easily, by adopting a fuller conception of charitable interpretation. According to Davidson, charitable interpretation is not merely a matter finding agents to have largely true and consistent beliefs, but also of finding their preferences to be largely in agreement with ours. Thus in interpreting someone 'we will try for a theory that finds him consistent, a believer of true beliefs, and a lover of the good (all by our own lights, it goes without saying)' (Davidson, 1970: 222).

Just as agreement in belief relies on a shared conceptual scheme, so we might claim that agreement in preferences relies on a (largely) shared preferential scheme, as we might call it.[57] Attributions of strange preferences may only be possible against a background of broad similarity between the contents liable to feature in an interpreter and an interpreted agent's preferences. Supposing that this suggestion is true, it might be suggested that the solution to the problem of indeterminacy in preference attribution might simply be that attributing non-standard preferences goes against the requirement for maximal agreement between an interpreter and an interpreted agent's preferences.

This proposal seems right, so far as it goes. But one might then ask what it is that determines the specific preferential contents that interpretable agents are (generally) bound to share. Even if a requirement for agreement ensures that agents cannot generally have preferences for different kinds of things, it does not establish the specific contents which agents' preferences must generally share (just as the requirement for general agreement between an interpreter and interpreted speaker's beliefs cannot, by itself, determine the contents of the beliefs which are eligible to be attributed).
It is here that Hurley's objective values seem to play a role. Just as, on a Davidsonian view, the nature of the world fixes the contents of our beliefs about it, so one might claim (as Hurley does) that the evaluative nature of the world fixes the contents of our preferences. Thus charitable interpretation involves treating people as susceptible to certain values, such

[57] Here a preferential scheme is a set of contents over which an agent's preferences range.

that they can be seen to have certain specific preferences (Hurley, 1989: 26). In the absence of an alternative proposal, the fact that agents share *the particular* preferential scheme that they do remains unexplained. A constraint of broad agreement between agents' preferences does little to determine the actual preferences that agents have in the absence of something to fix what it is that agents must generally prefer.

Thus, even though a requirement for broad agreement in preferences applies, there is still a need to explain the particular kinds of contents that agents' preferences must have. My proposed, non-normative solution to this worry is that having preferences which range over certain kinds of contents is a brute feature of being an interpretable agent. That is, it is a brute fact that certain kinds of contents are liable to feature in an interpretable being's preferences. The kinds of contents which are liable to feature in an interpretable being's preferences just are the ones which any interpretable being will tend to attribute when interpreting others (or, for that matter, herself).

Before concluding this chapter, it is worth noting an interesting consequence of the brute eligibility proposal. Hurley suggests that, given the need for substantive agreement between agents' preferences, rationality has both procedural and substantive dimensions. Hence: 'the interpretation of action is dependent on a non-optional principle of charity that reaches to the substantive rationality of desires as well as to their consistency' (ibid: 26). Thus it might seem that substantive constraints on agents' preferences entails that rationality has both substantive and procedural dimensions.

However, the brute eligibility proposal allows for the existence of substantive constraints on preference without entailing that rationality has a substantive dimension. On this view, it is not that there are certain values which we must treat agents as responsive to in attributing preferences to them. There is no requirement of practical rationality for agents to discover what they ought to desire, and to adjust their preferences such that these more closely *correspond* to what is valuable. Rather, the suggestion is that practical rationality merely involves deliberating about how to best satisfy the desires that one has, where the kinds of desires that one can have are subject to substantive constraints (i.e. must generally fall

within a certain preferential set). Thus although, on the brute eligibility proposal, the kinds of preferences that one has is constrained, there is no sense in which practical rationality involves modifying the preferences that one has to make them more compatible with the preferential set that interpretable beings must generally conform to. If I happen to have a preference which is relatively ineligible (or even just to lack one which is relatively eligible) that is not a rational error, but merely a fact about my preferential set which is consistent with the general constraint that my preferences are more rather than less eligible on the whole. Thus if brute eligibility is an option, the debate over substantive versus procedural conceptions of rationality does not seem to be resolvable simply in terms of the existence of substantive constraints on the eligibility of preferences.

4. Conclusion

In this chapter I have discussed two related objections to my proposed account of practical reasons. These are the objections that meaning and preference are, respectively, irreducibly normative. Both of these worries arise from the threat of radical indeterminacy. Thus it has been alleged that without normative constraints on meaning, it is radically indeterminate what a speaker means by her utterances. Insofar as such constraints introduce a normative dimension to interpretation, they are incompatible with a reductive account of normativity which is framed in terms of the constitutive constraints of interpretable functioning.

My response to this objection was twofold. First, I suggested that an account of semantic correctness in terms of truth, rather than in terms of the operation of semantic norms, can be adopted. Second, I suggested that, on a Davidsonian approach, meaning (as understood in truth-theoretic terms) can be sufficiently constrained by use, given the interpretive principle of charity. That is, the truth conditions for some sentence are sufficiently determinate, given the context in which that sentence is uttered and the constraint of charity on (radical) interpretation. The objection that charity is normative was forestalled by the claim that charity can be regarded as a constitutive constraint on interpretation, rather than as a normative principle. Thus it is not that one ought to interpret others charitably, but that one must, if one is to interpret them at all.

I then moved on to the second worry, that preference is radically indeterminate given choice. In the ensuing discussion I considered three possible accounts of substantive constraints on the eligibility of preferences. The first suggestion, from Hurley, was that preferences are substantively constrained by objective values. This suggestion threatens my view of practical reasons, by positing what are essentially normative constraints on interpretation. If interpretation is normatively constrained, normativity cannot be reductively explained in terms of the constitutive constraints of interpretable functioning (*contra* my account of practical reasons).

This threat led me to consider two alternative suggestions. The first was that simplicity acts as a substantive constraint on the eligibility of preferences. This suggestion was set aside because of worries over providing an account of simplicity which tracks our intuitive judgements about the kinds of distinctions which are eligible to feature in preference attributions. Some distinctions which appear relatively more eligible than others also seem to be relatively more complex, syntactically speaking, when spelled out in fundamental physical terms. Promising alternatives to a syntactic account of simplicity are not obvious, given that it is the content of distinctions which feature in preference attributions which are at issue. Thus the suggestion that eligibility is a matter of simplicity was set aside.

The second suggestion was that substantive constraints on the eligibility of preferences are a brute feature of interpretability. That is, interpretability simply involves making certain kinds of distinctions over others, with respect to preference. It is a brute fact that interpretation involves attributing preferences for certain kinds of things rather than others.

One worry with this suggestion was that it seems to require that interpretable agents have a capacity for detecting which kinds of distinctions are brutely more eligible than others, where such a capacity is arguably mysterious. However, I suggested that this worry could be avoided by claiming that it is simply in the nature of interpretable beings to draw certain kinds of distinctions over others. Thus no capacity for 'detecting' which distinctions have

brute eligibility is required. All that is required is a tendency to draw certain distinctions, where possessing this tendency is part of what makes a creature interpretable.

A second worry with the brute eligibility response was that certain distinctions (such as between bravery and cowardice) are essentially evaluative in kind. As such, the worry was that an account of the eligibility of such distinctions cannot avoid reference to their normative nature. My response to this worry was to suggest that perhaps the descriptive component of such distinctions can be identified independently of their evaluative component. This would allow their eligibility to be construed in descriptive terms, while their evaluative component is treated as a kind of *post hoc* attachment (either to be explained purely in terms of preference/affect, or in terms of the fact that it is a substantive constraint on interpretability that one must tend to have preferences involving the descriptive component in question).

A follow up objection was that some distinctions are purely evaluative. However, on a non-centralist view of value, thin evaluative distinctions are a function of thick ones. Invoking non-centralism, the proponent of brute eligibility can claim that the application of thin evaluative distinctions is eligible only insofar as these rely on the descriptive features of thick evaluative distinctions.

In the absence of further objections to brute eligibility, I take this to be an available view of the substantive constraints required for preference attribution. This proposal may seem unappealing, given that it involves positing that substantive constraints on preference are a brute feature of interpretable functioning. However, the alternative seems to involve treating the normativity of value itself as a brute feature of reality. I take this suggestion to be even less attractive. In any case, I conclude that the objection from the normativity of preference attribution can be avoided on an interpretivist account of practical reasons.

Chapter 5: The Objection from Moral Reasons

1. Introduction

A satisfactory account of practical reasons should allow for the possibility of moral reasons for action. There may be no such reasons, but they should not be ruled out in advance by an account of what practical reasons are. This chapter concerns the objection that the possibility of moral reasons is excluded by my account of practical reasons.

The initial force of this objection comes from the claim that moral reasons are/must be categorical (i.e. independent of the contents of agents' desires). This claim has been widely endorsed, particularly by Kantians, who hold that moral reasons are grounded in a categorical imperative of rationality. Others also hold that moral reasons are, or must be, categorical. These include many (though not all) utilitarians, who maintain that the moral requirement to maximise the good is independent of the contents of agents' desires, as well as many moral sceptics who, following Mackie, doubt the existence of moral reasons because of their supposedly categorical nature (Mackie, 1977: ch1).[58]

The account of practical reasons offered in this thesis seems unable accommodate the existence of categorical reasons for action. On this account, practical reasons are explained in terms of the specific desires that agents have. As such, it seems that they cannot apply regardless of the contents of these desires. Thus, if we accept that moral reasons must be categorical and that an account of practical reasons must allow for the possibility of moral reasons for action, my account of practical reasons appears to fail. However, as will be discussed in section 2, Mark Schroeder (2007) offers an account of moral reasons which

[58] As discussed in chapter 1, the normativity of practical reasons is problematic, regardless of whether they are taken to be categorical or hypothetical. Thus Mackie's specific concerns over the queerness of categorical, moral properties seem to be slightly misplaced.

attaches them to agents' desires while (potentially) retaining their categorical status. As will be seen, I am not in favour of Schroeder's approach. However, if such an account were to succeed, then the categoricity of moral reasons could potentially be accommodated within a desire-based theory of practical reasons (although not without some major adjustments, in the case of my particular theory).

Setting aside Schroeder's account of moral reasons, I assume that my approach cannot accommodate the existence of categorical moral reasons for action. This means that, to avoid the charge that my account of practical reasons cannot accommodate the possibility of moral reasons at all, I must deny that such reasons are necessarily categorical. In pursuing this response, the challenge is to show that a credible hypothetical account of moral reasons can be given, such that moral reasons can plausibly exist even if practical reasons are hypothetical in kind. This involves finding a way to account for moral reasons in terms of the contents of agents' particular desires, while respecting our everyday intuitions about the nature and status of moral reasons. In particular, it must be shown that we have reasons to perform the kinds of actions that are typically considered moral.

In what follows I refer to four prototypical moral act-types: (i) helping others; (ii) avoiding harming others; (iii) keeping one's promises/commitments; and (iv) promoting just social outcomes. I take these to be core examples of the kinds of actions that an account of moral reasons must be able to explain the existence of reasons to perform. In addition, certain *other criteria* must be met: the moral reasons that an account furnishes must be relatively strong, reasonably entrenched and admit of some kind of fairly unified explanation.

I consider three existing varieties of hypotheticalism about moral reasons: Foot's account of morality as a system of hypothetical imperatives (Foot, 1972); Hobbesian contractarianism (Gauthier, 1986; 2003; Hobbes, 1996: esp. chs. 14 & 15); Railton's account of moral reasons as reasons to do what is 'socially rational' (Railton, 1986).[59] I claim that each of these

[59] The term 'hypotheticalism' is adopted by Schroeder to label his own particular account of reasons for action. However, as explained in chapter 1, this is somewhat misleading in that some of the

accounts is promising, in certain respects. However, I suggest that none of them offers a complete explanation of moral reasons as it stands. Foot's approach side-steps some crucial issues regarding the content, generality and strength of our other-regarding desires. Gauthier's approach leaves us without reasons to perform certain typically moral actions, when a tendency towards performing these is not advantageous to us. Railton's approach relies on a concern for aggregate welfare which we may not have, while failing to capture reasons to promote justice or keep commitments where these conflict with aggregate welfare.

I propose a fourth species of hypotheticalism, similar in some ways to Foot's approach. According to the proposed account, morality is indexed to certain other-regarding desires typical of human agents, and (potentially) core to the desire-profile of any interpretable agent. This account will not be fully developed. Rather, my aim will be to show that a hypotheticalist account of moral reasons of this kind is sufficiently viable to release my account of practical reasons from the charge that it cannot accommodate the possibility of moral reasons for action.

Although the proposed account is taken to be the most promising explanation of moral reasons of the options considered, I also contend that a pluralist approach (which also invokes Gauthier's, and perhaps Railton's, views) may be best. This allows the strengths of each position to make up for the weaknesses of the other(s), explains certain conflicts in our moral judgements (such as between justice and utility) and is consistent with a reasonable over-determination hypothesis about the explanation of moral reasons.

It should be noted that some of the accounts of moral reasons discussed in this chapter are (as things stand) inconsistent with my general view of practical reasons, as stated in

---

reasons which are generated on Schroeder's view are (plausibly) categorical (i.e. independent of the specific contents of agents desires). I use the terms 'hypothetical' and 'hypotheticalism' generally, to refer to accounts on which reasons are dependent upon the specific contents of agents' desires. I refer to Schroeder's view simply as 'Schroeder's view'.

chapters 2 and 3 of this thesis. I try to flag up any inconsistencies where relevant. However, I do not take these inconsistencies to be crucial; I think that they can be avoided by suitably modifying either my account of practical reasons, the suggested account of moral reasons, or both.[60]

2. Schroeder's View

Mark Schroeder takes seriously the objection that Humean accounts of practical reasons cannot sufficiently account for moral reasons. He states that 'by any reasonable standard, all have failed' to account for 'the content of moral requirements' (Schroeder, 2007: 115). According to Schroeder, there are good reasons to think that moral reasons apply to everyone, regardless of what they desire, and that they do so with equal weight. A central concern of Schroeder's book *Slaves of the Passions* is to offer a Humean account of practical reasons (i.e. an account of practical reasons which makes essential reference to agents' existing psychological states—ibid: 2) which is consistent with there being moral reasons that apply to everyone, regardless of what they desire, and with equal, relatively high weight.

I have already explained Schroeder's general view of practical reasons in chapter 1. I now discuss his particular attempt to explain the existence of moral reasons of the kind introduced above. Here Schroeder takes his problem to involve showing how there can be

---

[60] For example, on Gauthier's view it is the consequences of agents' tendencies/dispositions to perform certain types of actions which determines what is rational, rather than the consequences of any particular actions in themselves. By contrast, my account of practical reasons (as set out in chapters 2 and 3) relies on an act-based account of practical rationality. Thus the two accounts are incompatible as they stand. I take this incompatibility to be avoidable, in that I think the rational principles which underpin interpretable functioning can, potentially, be cast in terms of agents' tendencies to intend certain types of actions. As mentioned in chapter 3, it is unclear whether the best account of rationality is act-based or disposition-based. The act-based approach was adopted for reasons of simplicity, given the purposes of this thesis, but it can be avoided if there are good reasons to opt for a disposition-based approach.

equally weighty reasons for everyone to perform some (moral) action, regardless of what they desire. Thus he does not want to show merely that there can be categorical moral reasons (i.e. reasons to perform some action, regardless of what one desires) but that these reasons are also agent-neutral (i.e. that they apply equally to *all* agents).

Showing that agent-neutral reasons for action exist appears to be impossible, given that on Schroeder's view a fact/state of affairs only counts as a reason to act if an agent has some desire, the object of which will be promoted by performing some action (given that fact/state of affairs). If reasons are indexed to desires, how can they be universal to all agents, and with equal weight? Here Schroeder has to overcome two major hurdles. The first is to show that there can be genuinely agent-neutral reasons for action. The second is to show that these agent-neutral reasons can be equally weighty for everyone (and significantly weighty too).

Schroeder's strategy in overcoming the first hurdle is to offer a very weak understanding of the promotion relation, such that reasons are *very* easy to come by. If there is an abundance of reasons to do all sorts of things then it is reasonable to suppose that there can be genuinely agent-neutral reasons (i.e. reasons for everyone to perform some action, regardless of what they desire). This is because agent-neutral reasons are 'massively overdetermined' in the sense that they are facts which explain why acting in a certain way will promote the satisfaction of (almost) any desire (ibid: 109).

Schroeder's chosen view of the promotion relation is that '*X*'s doing *A* promotes *p* just in case it increases the likelihood of *p* relative to some baseline' where this baseline 'is fixed by the likelihood of *p* conditional on *X*'s doing nothing—conditional upon the status quo' (ibid: 113). On this view, it is at least theoretically possible for there to be actions which will raise the probability of almost any desire being satisfied. Reasons to perform such actions would exist for everyone, regardless of what they desired. Schroeder's hope is that moral actions, such as helping a stranger in need will, in his weak sense, promote the satisfaction of almost any desire such that the fact that somebody is in need will be a reason for everyone to help her, regardless of what they desire.

Setting aside concerns over Schroeder's exact formulation of the promotion relation (and particularly over the idea that there is some kind of existential status quo which will be realised if a person does 'nothing'), the over-determination hypothesis about moral reasons seems to be particularly dubious on the surface. This is because it seems highly implausible to claim, say, that helping a stranger could raise the probability of almost any desire that I might hold being satisfied. What if my sole desire were to avoid helping strangers?

Perhaps the idea is more plausible if we assume that agents with a single desire could not exist. Perhaps an agent must always have some network of desires, where helping a stranger in need will always raise the probability that at least one desire in any possible network will be satisfied. However, this claim is still relatively controversial and in need of significant argumentative support.

Schroeder does offer some wiggle room here, claiming that a spectrum of views about moral reasons is afforded by his over-determination hypothesis. The limits of this spectrum represent the different extents to which there can be actions which promote (in Schroeder's weak sense) the satisfaction of *any* possible desire. Thus at one extreme it might be that moral reasons are Kantian in scope: moral actions promote the satisfaction of almost any possible desire such that reasons to perform such actions apply to all agents, regardless of what they desire. Alternatively, it may be that moral reasons are Aristotelian in scope: moral actions promote the satisfaction of any desire that a human agent could hold, such that reasons to perform such actions apply to all human agents (it is assumed that, on an Aristotelian view, human agents are constrained to hold certain desires). Finally, it may be that moral reasons are only limited to the class of agents who have certain specific concerns. This would be a more conventionally Humean position whereby agent-neutral reason ascriptions (including those concerning morality) 'have to be understood either as false, or as having restricted scope—implying only that the reason is a reason for all of *us*, around *here*' (ibid: 118).

Schroeder is optimistic about the prospects for the 'Kantian' view of universality. As stated above, I have reservations about this. However, I will set these aside as the most interesting and important aspect of Schroeder's view, for present purposes, is his account of the weight of reasons.

According to Schroeder, the weight of reasons should be understood in terms of a 'weightier than' relation which applies to sets of reasons. This relation is ultimately cashed out in terms of two elements, a weight base and a weight recursion. Schroeder characterises these as follows:

Weight Base	One way for set of reasons *A* to be weightier than set of reasons *B* is for set of reasons *B* to be empty, but *A* non-empty.
Weight Recursion	The other way for set of reasons *A* to be weightier than set of reasons *B* is for the set of all the (right kind of) reasons to place more weight on *A* to be weightier than the set of all the (right kind of) reasons to place more weight on *B*. (ibid: p138).

Applying these criteria allows Schroeder to divorce his version of Humeanism from a proportionalist account of the weight of reasons, according to which the weight of a reason is proportional to: (*i*) the strength of the desire to which it is attached and (*ii*) the degree to which the relevant action will promote the object of that desire (or else the degree to which it is believed that it will promote it). Schroeder finds this view unattractive and, in any case, it is incompatible with his claim that moral reasons apply to everyone with equal weight. If the strength of a reason is tied to the strength of the desire whose object the relevant action will promote, together with the degree to which it will promote it, then whether any agent-neutral reasons (which exist because some action will promote the satisfaction of almost any possible desire) will be equally weighty for all agents becomes a matter of contingency. Thus, on proportionalism, moral reasons need not have equal weight for everyone.

Schroeder's proposed weightier than relation is intended to allow for agent-neutral reasons which are equally weighty for everyone, and non-coincidently so. The basic suggestion is that an agent's reasons to *A* will be weightier than her reasons to *B* either if she has reasons (of the right kind) to *A* but no reasons to *B*, or if her reasons (of the right kind) for placing weight on her reasons to *A* are weightier than her reasons (of the right kind) for placing weight on her reasons to *B*.[61] Here the relative weight of reasons to place weight on reasons is, ultimately, determined by whether a set of reasons to place weight on a set of reasons is empty or not (the weight base).

This suggestion may seem somewhat obscure in the abstract. Schroeder illustrates it using an epistemic example. In this example you see Tom Grabit come out of the library and remove a book from under his shirt. This is a reason to believe that he stole the book. However, you then learn that Tom has a twin brother, Tim. This is a reason to place less weight on the fact that you saw someone who looked like Tom stealing a book when considering whether or not he did so. Then you learn that Tom's brother Tim was abroad at the time. This is a reason not to place weight on the reason not to place weight on your reason to think that Tom stole the book. Then you learn that Tom and Tim have a third identical sibling, Tam, who was in the country at the time. This is a reason... and so on.

The eventual idea is that *at some point* there will be reason(s) to place weight on a reason and no reason(s) not to. This will act as a base from which the weightier than calculus can operate, running through the whole sequence of reasons to place weight on reasons until one reason turns out to be weightier than another.

---

[61] According to Schroeder, the 'right kind of reasons' problem arises whenever an activity is governed by some standard of correctness (Schroeder, 2007: 133). The problem is to explain why some reasons are the right kind to take account of and some are not, in determining what is correct. For example, that a chess move will increase one's chances of winning the game is the right kind of reason to make it; that it will allow one to line one's pieces up in height order is not. The question is, why not? Schroeder offers an answer to this, which is discussed in the text with respect to the activity of deliberation.

This account of reasons' weight allows Schroeder to claim that agent-neutral reasons, including moral reasons, will be equally weighty for everyone. This is because, according to Schroeder, the right kind of reasons to place weight on in deliberation just are the agent-neutral reasons, *viz.* 'reasons that are shared by everyone engaged in the activity of doing *A* [deliberating], such that the fact that they are engaged in doing *A* is sufficient to explain why these are reasons for them' (ibid: 135). Agent-neutral reasons are the kinds of reasons that any deliberator will have, given that such reasons exist in relation to (almost) any possible desire. So, if the right kinds of reasons to place weight on in deliberation just are those shared by anyone engaged in the practice of deliberation, then agent-neutral reasons will be the right kind of reasons to place weight on in deliberation. Assuming that this is right, it is agent-neutral reasons which govern the weight of reasons (that is Schroeder's suggestion, in any case). This suggestion allows all agents to have equally weighty moral reasons, as these are reasons that everyone shares (i.e. agent-neutral reasons), given that they are massively over-determined with respect to promoting the satisfaction of agents' desires.

Schroeder's is an extremely controversial account of the weight of reasons. Among other things, it involves accepting that the strength of desires and the (degree of belief in the) extent to which an action will promote their satisfaction have no direct bearing on the weight of reasons to perform an action.[62] This is an exceptionally controversial suggestion in the context of a Humean approach to practical reasons. For a start, it seems to entail that practical reasoning does not involve calculating how to maximise desire satisfaction at all. Rather, it involves establishing which reasons one has most reason to place weight on in deliberation, given the various massively over-determined (agent-neutral) reasons that one has for placing weight on each of one's reasons. This has implications for many disciplines and threatens to undermine any theory which relies on a decision-theoretic account of

[62] Proportionalists can either claim that the weight of a reason is proportional to the actual degree to which performing some action will promote the satisfaction of a desire, or to the degree of belief that the relevant agent has that performing some action will promote the satisfaction of a desire. For purposes of simplicity, I discuss Schroeder's argument by counter-example against proportionalism in terms of degrees of promotion rather than degrees of belief.

practical rationality (including Davidson's account of radical interpretation, as it stands). This suggests that Schroeder's rejection of proportionalism requires significant support, if it is to be taken seriously. It also suggests that his own account of practical reasons' weight, and of how considerations of weight feature in deliberation, needs to be well developed and defended.

However, Schroeder's arguments against proportionalism are less than convincing. He offers one argument by counter-example, in which Aunt Margaret's strongest desire is to recreate a scene from a *Martha Stewart Living* catalogue on Mars (ibid: ch.5). This desire gives her a reason to build a Mars-bound spaceship (a necessary means to get there, given that no-one will give her such a ship). However, Schroeder assumes that this reason must be relatively weak, given that building such a ship is a crazy thing to do. If Aunt Margaret's reason to build a Mars-bound spaceship is relatively weak, despite its being a necessary means to the satisfaction of her strongest desire, then proportionalism (on which such a reason, it is presumed, must be relatively strong) is false.

This argument is extremely problematic. First, it assumes that Aunt Margaret's reason to build a Mars-bound spaceship *is* relatively weak. This assumption may seem intuitive, but I think that it needs to be argued for. Second, supposing that we grant this assumption, it is also assumed that on a proportionalist account of reasons' weight, Aunt Margaret's reason to build a Mars-bound spaceship must be relatively strong. This assumption seems to fall out of the description of the case, which seems to involve the assumption that Aunt Margaret is *capable* of building a Mars-bound spaceship. However, if this assumption is false (i.e. if Aunt Margaret is incapable of building such a ship) then, assuming that to have a reason to do something one must be able to do it, she has no reason to build a Mars-bound spaceship after all.

Alternatively, suppose that Aunt Margaret is able to build a Mars-bound spaceship. In that case it is still doubtful whether doing so will significantly promote the satisfaction of her strongest desire, for perhaps she is unable to pilot such a ship to Mars, or has no place to launch it from, or has no space-agency to act as mission-control, or whatever. On

Schroeder's view of promotion, an action promotes the satisfaction of a desire if it raises the probability of that desire's being satisfied, relative to some background level. Assuming that there are many other obstacles which are likely to prevent Aunt Margaret from actually getting to Mars, building a Mars-bound spaceship does not significantly raise the probability of her desire being satisfied. Thus, on a proportionalist account, and given Schroeder's own view of promotion, Aunt Margaret's reason to build a Mars-bound spaceship may not be very weighty.

Finally, suppose both that Aunt Margaret is able to build a Mars-bound spaceship, and that all of the other necessary conditions for Aunt Margaret to get to Mars are likely to obtain (perhaps she is a rich rocket scientist and astronaut with fantastic contacts in the space-agency, etc.). In that case, perhaps her reason to build a Mars-bound spaceship is relatively weighty after all (recreating her chosen scene on Mars would certainly be a lot of fun!).

Although Schroeder assumes that Aunt Margaret's reason to build a Mars-bound spaceship must be relatively weak, I see no reason to make this assumption. Although intuitions might conflict at this point, it is at least open for the proportionalist to maintain that if Aunt Margaret was able to build a Mars-bound spaceship, and if she was relatively likely to get it to Mars, then she has a relatively weighty reason for doing so, given her strongest desire. In the absence of some argument against this claim, I take Schroeder's argument by counter-example to beg the question.

Schroeder's other argument is that proportionalism does not offer a good analysis of what it is for reasons to have weight. He suggests that weight is a normative feature of reasons, which has to do with the correctness of placing weight on reasons in deliberation (ibid: 100-1). Strengths of desires and degrees of belief/degrees of desire-promotion are not normative and, therefore, cannot explain why it is correct to place a certain degree of weight on a reason when deliberating. Thus a proportionalist account of what it is for reasons to have weight is on the wrong track.

I think this argument is too quick. This is because the proportionalist does not need to explain the weight of reasons simply in terms of strengths of desire and degrees of belief/degrees of promotion of desire satisfaction. Rather, she can draw on her account of practical reasons' normativity, using this, in combination with strengths of desire and degrees of belief/desire-promotion, to explain reasons' weight. For example, on my interpretivist view the normativity of practical reasons is explained in terms of its being a constitutive requirement of interpretable functioning that agents (generally) form certain intentions to act, given their beliefs and desires. On this account, reasons' weight is proportional to strengths of desire and degrees of belief just because interpretable functioning (generally) involves forming intentions to act in ways which one believes will maximise overall desire satisfaction, given one's means-end beliefs. Thus it is correct to place weight on reasons, *qua* belief desire-pairs, in proportion to strengths of desire and degrees of belief just because doing so is a constitutive requirement of interpretable functioning.

Setting my own view of practical reasons' normativity aside, I think that it is open for proportionalists in general to account for the correctness of placing weight on one's reasons in proportion to strengths of desire and degrees of belief/desire-promotion by invoking their general account of the normativity of reasons. In the absence of arguments against this strategy, I take Schroeder's allegation that proportionalism fails to account for practical reasons' weight to fall short.

I now turn to Schroeder's own account of practical reasons' weight. A troubling objection to this account comes from Heuer (unpublished), who suggests that, among other things, Schroeder's account of practical reasons will generate agent-neutral reasons to do all kinds of zany things which, because of his account of weight, will nevertheless be relatively weighty.[63] Suppose, for instance, that building a Mars-bound spacecraft is a way to promote the satisfaction of many different desires (fulfilling a childhood dream, teaching one's

[63] Shafer-Landau (2011: §4) makes a similar point about the proliferation of agent-neutral (and, therefore, weighty) reasons given Schroeder's view of reasons' weight.

children about science, having something to do, etc). If so, then building a Mars-bound spacecraft is something that everyone potentially has some reason to do, given Schroeder's weak promotion relation. This seems to generate absurd results. As Heuer puts it, there is 'potential for an uncontrollable explosion of reasons for all of us to do all kinds of zany things, and the fact that *everyone* has the same reason means that it is effectively agent-neutral and thus the "right kind of reason" to place weight on in deliberation' (ibid: 11, emphasis in original). On Schroeder's view, perhaps *everyone* has a relatively weighty reason to build a Mars-bound spaceship!

Heuer proposes two possible solutions to this problem. The first is to strengthen the right kind of reasons filter, such that the problematic kinds of agent-neutral reason are no longer the right kind of reason to place weight on in deliberation. The second is to strengthen the promotion relation, such that there are fewer agent-neutral reasons. Heuer does not make an explicit suggestion as to how the promotion relation might be strengthened. Given that Schroeder provides a good case against two stronger versions of the promotion relation, I set this proposal aside (Schroeder, 2007: 110-13). Heuer does, however, suggest how the right kind of reasons filter might be strengthened.

The proposal here is that Schroeder could accept that deliberating involves more than simply weighing up one's reasons. Rather, perhaps deliberating necessarily involves having some goal or other, such that the reasons which any agent has by virtue of being a deliberator are those which all agents share because of their relation to that goal. If this is right, then the right kinds of reasons to place weight on in deliberation will be those which relate to the constitutive goal of deliberation, rather than those which any agent has because some action will promote the satisfaction of any possible desire.

This proposal means that Schroeder would have to defend the idea that there is some constitutive goal of deliberation (over and above deciding what to do). However, this seems to conflict with Schroeder's attack on Velleman's proposal that there are reasons that all agents share because they attach to desires that are constitutive of being an agent (Schroeder, 2007: 107; Velleman, 2000: chs. 1 & 8). Nevertheless, Heuer proposes that the

two approaches may not be as opposed as Schroeder supposes, in that Schroeder's view seems to involve the idea that deliberation 'stands in for something like Velleman's "constitutive aim of action" as being inextricably linked with all actions, and as the source of our agenthood, [the explanation of] our susceptibility to reasons for action' (Heuer, unpublished: 18). If Schroeder were prepared to restrict the aims of deliberation in some way then he might be able to restrict his class of agent-neutral reasons possessing deliberative weight to the non-zany kinds. However, the problem of demonstrating a constitutive goal of deliberation would then represent a significant challenge for the revised Schroederian view. Schroeder's reasons overdetermination hypothesis, together with his existing account of reasons' weight are already controversial and problematic. Defending the additional proposal that deliberation has a constitutive aim, where the posited aim must allow for an improvement in Schroeder's existing view of reasons' weight, would only add to these problems.

As can be seen, Schroeder's attempt to secure moral reasons which are agent-neutral and equally weighty for everyone is highly contentious and somewhat problematic. Although Schroeder's attempt to offer a Humean account of moral reasons with the scope traditionally associated with Kantianism is both novel and interesting, I do not think that it is ultimately successful. Therefore, I shall not attempt to invoke Schroeder's account of moral reasons in defence of the charge that my own account of practical reasons cannot accommodate the possibility of moral reasons for action.[64] In the remainder of this chapter I consider some more conventional (and less ambitious) attempts to explain moral reasons in terms of agents' desires.

[64] There are also many incompatibilities between Schroeder's general account of practical reasons and my own. I shall not discuss or attempt to reconcile these here, given that I set Schroeder's account of moral reasons aside.

3. Foot: Morality as a System of Hypothetical Imperatives

Before adopting an Aristotelian naturalist approach to moral reasons (discussed in chapter 1), Foot claimed that morality can be best understood as a system of hypothetical imperatives (Foot, 1972). Her starting point was to question the claim that moral imperatives are categorical, asking what the notion of a categorical imperative might amount to. Finding no promising way of understanding this notion, she suggested that morality might be like other sets of imperatives, such as those of etiquette, which are expressed in categorical terms but are nonetheless hypothetical. As Foot explains, 'considerations of etiquette [which are expressed categorically] do not have any automatic reason-giving force, and a man might be right if he denied that he had any reason to do "what's done" '(ibid: 309).

On Foot's view, the demands of etiquette are hypothetical in that they only provide reasons to behave 'correctly' if a person desires to do 'the done thing' (ibid: 309-10). Thus a person can lack any reason whatsoever to do what is 'correct', as etiquette has no desire-independent grounding to underpin a sense in which 'the done thing' is *to-be-done*. Likewise, Foot contended that there is no desire-independent grounding to morality to underpin a sense in which the moral thing is *to-be-done*. Categorical statements about what is morally required are categorical in form but not in content.

> It seems, then, that in so far as it is backed up by statements to the effect that the moral *is* inescapable, or that we *do* have to do what is morally required of us, it is uncertain whether the doctrine of the categorical imperative even makes sense.
>
> The conclusion we should draw is that moral judgements have no better claim to be categorical imperatives than do statements about matters of etiquette. People may indeed follow either morality or etiquette without asking why they should do so, but equally well they may not. They may ask for reasons and may reasonably refuse to follow either if reasons are not to be found (ibid: 312, emphasis in original).

So morality, like etiquette, should be regarded as a system of hypothetical imperatives. Moral imperatives are typically expressed in categorical terms but, lacking any desire-independent grounding, have no claim to be categorical in a more fundamental sense.

Foot's early account of moral imperatives seems to place the demands of morality on a par with the somewhat specious demands of etiquette. Both sets of demands only apply on the condition that a person has certain relevant desires, despite any intuitions we may have that the demands of morality have a special importance lacked by those of etiquette. Further, given an assumption that agents are primarily self-interested, agents might often have reasons to flout morality rather than to uphold it. Thus morality seems to lose its distinct and special status altogether if moral imperatives are hypothetical.

Foot's response to these problems is to set aside the self-interested picture of human nature which, she suggests, motivates the adoption (by Kant in particular) of a categorical approach to morality. Foot replaces this view with one according to which 'a man may care about the suffering of others, having a sense of identification with them, and wanting to help if he can' (ibid: 313). This, along with other suitable desires (such as for truth, liberty, and to treat others with respect) accounts for the importance of moral considerations in human life. Thus, for the early Foot, moral reasons are conditional upon desires, but the desires that they are conditional upon are a deep and central feature of human nature. This explains why so many are 'prepared to fight so hard for moral ends—for example liberty and justice' (ibid: 314). By contrast, 'one could hardly be devoted to behaving *comme il faut*' (ibid: 314).

I agree with much of this picture. Specifically, I agree that many of our moral reasons are conditional upon certain deeply held desires that we have concerning the well-being of others, and the ways in which we relate with them. However, Foot does not consider or address the problem central of how to account for hypothetical imperatives (as introduced in chapter 1, section 3). That is, she does not discuss or discharge the worry that, for there to be hypothetical reasons for action, desire satisfaction must be 'to-be-pursued' in some

sense. This threatens to undermine her account of morality as a system of hypothetical imperatives; if hypothetical imperatives are subject to the same concerns about prescriptive normativity as categorical imperatives, little is gained by claiming that moral imperatives are hypothetical rather than categorical.

Nevertheless, I think that the account of reasons provided in this thesis can overcome (or at least avoid) the problem of how hypothetical reasons are to be explained. Desire satisfaction is not to-be-pursued in some fundamental sense, but our desires do entail that certain courses of action make more or less sense than others (i.e. make us more or less interpretable). Practical reasons are to be understood in terms of patterns of relations between agents' actions and their desires/beliefs which support interpretability. On an interpretation-based account of this kind, practical reasons are essentially hypothetical (i.e. conditional upon agents' desires) because sense-making essentially involves establishing relations between agents' desires and their actions.

Supposing that this account of practical reasons is plausible, we can plug Foot's account of moral reasons in to get an account of moral reasons that is hypothetical in kind. This is, in the main, the sort of account of moral reasons that I wish to endorse.

However, issues over the nature of hypothetical reasons aside, I am not convinced that Foot has done enough to show that her hypothetical account of moral reasons is viable. This is because she does not give enough detail on the kinds of desires that generate moral reasons to convincingly claim that such desires generate reasons sufficiently in line with the moral reasons that we normally take ourselves to have. I suspect that this lack of detail is a direct consequence of Foot's method of approaching the problem. This involves taking the content of moral imperatives for granted (as with the content of etiquette), and then asking what kinds of reasons we might have to abide by these imperatives. This suggests a two tier structure, whereby moral imperatives are taken to have a fixed content which we may or

may not have reasons to follow. Thus moral reasons, on Foot's account, seem to be reasons to do 'the moral thing' in a *de dicto* sense, rather than a *de re* sense.[65]

Perhaps the idea that the content of moral imperatives is independent of the desires which generate reasons to 'be moral' gains credibility from Foot's analogy with etiquette. Etiquette appears to consist of a somewhat arbitrary set of conventions that agents may or may not have reasons to follow, depending on the extent of their desire to do 'the done thing'. However, it seems that there is a disanalogy between etiquette and morality here. Plausibly reasons of etiquette are reasons to do 'the done thing', explained by a desire to behave 'properly'. Etiquette is largely a matter of convention after all; it is reasonable to suppose that, on the whole, it is a desire to act according to social convention which generates reasons of etiquette. As such, agents' reasons for doing 'the done thing' do not necessarily have any bearing on its content.[66]

With morality things seem to be different. Reasons to be moral do not, in general, seem to be reasons to adhere to moral conventions *qua* conventions. As Foot suggests, it is a concern for others which largely accounts for our moral reasons for action (and not a concern to follow moral convention). As such, the content of morality is not generally a matter of convention and moral reasons do not seem to be reasons to do 'the moral thing' in a *de dicto* sense.[67] Rather, on a plausible Foot-style account, certain of our (other-

---

[65] Foot does not explicitly defend this two tier structure (according to which the content of morality is independent of the desires which give us reasons to be moral). Nevertheless, Foot's assertion that we can reasonably ask for reasons to follow the demands of morality does suggest that she considers the content of morality to be fixed independently of our reasons for being moral.

[66] An explanation of the content of etiquette is needed, of course. Presumably, on a hypothetical account, there is some connection between the content of etiquette and the desires of those agents who existed at the time that its conventions were established. However, that is not a matter for discussion here.

[67] One might suggest that there can be reasons to do 'the moral thing' in a *de dicto* sense where the content of morality is not taken to be conventional. However, it is difficult to find an explanation of

regarding) desires must fix the content of our moral reasons. Although there may be well-established moral principles, these are always up for grabs; what is morally conventional is not necessarily what we have reason to do. This is so for all but the moral fetishist, who does act morally in order to adhere to moral convention (at least on a hypotheticalist account of practical reasons, such as Foot's). It can reasonably be suggested that the moral fetishist is morally lacking because of this.[68]

The problem for Foot is that her methodology seems to conflict with the need to offer a full explanation of the content of morality in terms of our desires. Thus although she suggests that moral reasons are generated by certain of our (other-regarding) desires, she also makes a tacit methodological assumption to the effect that the content of morality is fixed prior to our reasons for being moral. This involves taking the content of morality for granted. Consequently, Foot fails to offer sufficient demonstration that our desires do generate reasons with the appropriate moral content.

Foot's account cannot be accepted as viable until: (i) the (tacit) assumption that the content of moral imperatives is fixed independently of our desires is removed; and (ii) the claim that our desires *do* give us reasons with the appropriate moral content is adequately supported. I attempt to provide some support for this latter claim towards the end of this chapter—enough support to show that it is plausible, if not true.

I also believe that we can go further than Foot, by claiming not just that humans have certain other-regarding desires which give them reasons to be moral, but that these desires are core to agency in the sense of being part of the desire-set that agents must, on the

---

the content of morality which is consistent with this claim, given a hypothetical view of reasons for action.

[68] On moral fetishism see Smith (1994: 76); Lillehammer (1997). Smith (1994: 77-84) takes Foot's account of morality to fail because it makes moral action a kind of fetish. As discussed, I think that Foot's discussion has this implication in places, although I think that her emphasis on our other regarding concerns also shows that her view can be developed in a non-fetishising way.

whole, have if they are to be interpretable. At the end of section 6 I suggest that some degree of support for this claim can be provided, although I do not give it very much attention as it is not crucial in rebutting the objection from moral reasons.

4. Gauthier: Morals by Agreement

In contrast to Foot, Gauthier does not pick out any particular desires which generate moral reasons. Rather, he attempts to explain moral reasons in terms of the structure of rational agency. Gauthier claims that the maximal satisfaction of an agent's desires, whatever these are, sometimes requires that she accepts constraints on her maximising behaviour as, where these constraints are mutually accepted within a social group, they can lead to net gains in desire satisfaction for everyone concerned (Gauthier, 1986; 2003). For example, if everyone agreed not to murder each other then, assuming that everyone complied with this agreement, they would all benefit by avoiding the threat of being murdered. If avoiding this threat brought with it a higher expected utility than the expected utility of being free to commit murder, then it would be rational to accept a constraint against murder.

Not that morality is a question of actual agreements. Rather, Gauthier claims that if one were drawing up a social contract from within a pre-social position (a position in which no-one can expect others to constrain their maximising behaviour, in which no agreements or bargains already exist, and in which no person or group can exercise social dominance or coercion), it would be rational to include certain constraints on agents' maximising behaviour as these would lead to mutual benefits for all concerned. To the extent that a tendency to adhere to the kinds of constraints that one *would* agree to under these circumstances increases one's own expected utility, one has reasons to be moral.

Gauthier's moral theory stems from a recognition of the role of others in allowing/enabling us to maximally satisfy our own desires. If we ignore other people, and their interests, we run the risk that they will harm us in the pursuit of their ends, while we miss out on certain benefits to be gained from cooperative/coordinated action. Mutually cooperative action

holds the promise of security and increased individual prosperity. The logic of this situation is summarised by Gauthier as follows:

> No one, of course, can have reason to accept any unilateral constraint on her maximising behaviour; each benefits from, and only from, the constraint accepted by her fellows. But if one benefits more from a constraint on others than one loses by being constrained oneself, one may have reason to accept a practice requiring everyone, including oneself, to exhibit such a constraint (ibid, 2003: 98-9).

Crucially, Gauthier claims that it can be rational for agents to adhere to constraints on their maximising behaviour, even where this involves forgoing certain benefits which could apparently be gained by free-riding on others' acceptance of such constraints (i.e. playing Hobbes' Fool—Hobbes, 1996: 96).

In explaining this outcome, Gauthier characterises a straightforward maximiser (SM) as 'a person who seeks to maximise his utility given the strategies of those with whom he interacts' (Gauthier, 1986: 167). He characterises a constrained maximiser (CM) as:

> (i) someone who is conditionally disposed to base her actions on a joint strategy or practice should the utility she expects were everyone so to base his action be no less than what she would expect were everyone to employ individual strategies...; (ii) someone who actually acts on this conditional disposition should her expected utility be greater than what she would expect were everyone to employ individual strategies (ibid: 167).

SMs choose any expected-utility-maximising action; CMs choose any co-operative action with higher expected utility than that of everyone acting individually, conditional on the expectation that others will so co-operate.

If it is assumed that agents are transparent (i.e. that it is clear whether an agent has a tendency towards straightforward or constrained maximisation), then CMs will do better

than SMs. This is because, even though SMs would be able to benefit from free-riding on CMs if the opportunity arose, this opportunity never arises. If CMs are always able to correctly identify SMs, they will always adopt a strategy of straightforward maximisation when interacting with them. But, when interacting with other CMs, they will adopt a strategy of constrained maximisation (if and when this brings with it the prospect of greater expected utility than each could achieve on her own), thus reaping benefits not available to SMs. So, supposing rationality is to be understood in terms of expected-utility maximising tendencies, those with a tendency towards constrained maximisation will be more rational than those with a tendency towards straightforward maximisation (if the transparency assumption holds).

Unfortunately, transparency does not hold. CMs will sometimes fail to recognise each other, while SMs will sometimes successfully convince CMs that they are of like disposition. This leads Gauthier to relax the transparency assumption in favour of an assumption of translucency—that agents are able to recognise each others' tendencies with some degree of success (ibid: 174).

This complicates things somewhat, in that CMs can do better than SMs, but not in all circumstances. Specifically, whether CMs will do better than SMs depends upon three factors: (a) the degree of translucency (how effective agents are at identifying each others' tendencies); (b) the ratio of gains from defection to gains from cooperation; (c) the proportion of CMs and SMs in the population.

(a) The worse agents are at detecting each other's dispositions, the better SMs will do. The more opaque agents are, the more opportunities there are for SMs to profit at CMs' expense and the fewer opportunities there are for CMs to successfully cooperate.
(b) The lower the ratio of gains from defection to gains from cooperation, the better CMs will do. This is because they will need to successfully cooperate fewer times to compensate for each time that they are unwittingly exploited.

(c) The more SMs there are in a population, the more likely it is that a CM will mistakenly take an SM to be a CM and get exploited (assuming some fixed degree of translucency), and the less likely she is to find other CMs with whom she can successfully cooperate.

These three factors mean that a straightforward case cannot be given for the rationality of a disposition towards constrained maximisation. Nevertheless, Gauthier shows that circumstances under which a tendency towards constrained maximisation is rational can, plausibly, exist. For instance, if the ratio of gains from defection to gains from cooperation is 2-1, and the balance of CMs to SMs in the population 50-50, then CMs must manage to successfully cooperate in 2/3 of their encounters with each other and avoid exploitation in 4/5 of their encounters with SMs to do better (ibid: 174-7).

No direct argument is given to show that the conditions in which we actually exist make constrained maximisation rational. Rather, it is shown that a disposition towards constrained maximisation is rational 'if persons are sufficiently translucent and enough are like-minded', given certain assumptions about the relative payoffs of cooperation and defection (ibid: 177).

However, an indirect argument for the rationality of being a CM can be taken from Gauthier's comparison of his account of agent rationality to Trivers' evolutionary theory of reciprocal altruism (Trivers, 1971). On Trivers' theory, both egoistic and reciprocally altruistic populations are genetically stable, in that people with the opposing tendency in each population would die out as they did worse with respect to their peers (supposing that survival pressures have lead altruists to be able to reliably detect egoists within their population). However, although both types of population are genetically stable, populations of reciprocal altruists can expect to do better in absolute terms (*ceteris paribus*) than populations of egoists as 'the benefits of co-operation ensure that, in any given set of circumstances, each member of a group of reciprocal altruists should do better than a corresponding member of a group of egoists' (Gauthier, 1986: 188). Thus, on Trivers'

account, we have an evolutionary explanation of the success of populations with a tendency towards reciprocal altruism.

This leads Gauthier to propose that 'if human beings are so disposed [towards utility maximisation], then we may conclude that the disposition to constrained maximization increases genetic fitness' (ibid: 189). That reciprocal altruism is evolutionarily advantageous is evidence for the claim that the conditions necessary for constrained maximisation to be rational are generally realised among human beings.

Supposing this to be correct, Gauthier's contractarian moral theory is an appealing hypotheticalist account of moral reasons. Moral reasons are indexed to a (rational) tendency to adhere to the kinds of constraints on maximising behaviour that agents would accept if drawing up a social contract from within a pre-social position. This theory fits nicely into a plausible account of agent rationality. It also fits nicely into an evolutionary account of the development of morality. Finally, the theory appears to successfully capture much of the content of morality: agreements to help others, to refrain from harming them, to keep commitments, and to maintain just social outcomes all seem to be the kinds of things that would be included, to some degree, in Gauthier's hypothetical social contract. A (limited) tendency towards each of these can reasonably be supposed to increase expected utility, when generalised throughout the population, in comparison with its absence.

Despite its advantages, Gauthier's theory has some significant limitations. The first limitation is that the purely formal nature of Gauthier's theory is insufficient to generate specific moral content. Gauthier's claim is that morality can arise given certain purely formal considerations. As long as situations arise which meet certain structural conditions (specifically, that an agent can do better by having her peers constrained than she can by being unconstrained herself) then morality gets off the ground. Thus for Gauthier, it does not matter what agents desire; morality is a consequence of the structure of certain social choice settings and not of their content.

However, it seems that Gauthier's theory does require some substantive constraints on desire. For instance, it could not be that agents desire more than any other possible outcome to avoid being constrained; if they did then the expected utility of accepting a constraint would always be lower than the expected utility of avoiding it. Such paradoxes aside, perhaps Gauthier can claim that the mere structural possibility of doing better by accepting certain behavioural constraints than by rejecting them is enough to generate moral reasons. However, it seems obvious that for Gauthier's account to generate moral reasons with any specific content, it must take into account the desires that agents do have rather than simply invoking the structure of certain social choice settings. People do best by refraining from stealing because they value property. If people did not value property then accepting a constraint against theft would have no rational basis. So, on Gauthier's theory, the content of morality cannot simply be a function of certain formal features of social choice; it must also be a function of the particular desires of the agents that are plugged into the relevant social choice settings. Formal conditions alone are insufficient to generate the distinctive constraints on behaviour that make up the contents of morality.

Perhaps this problem can be accommodated by accepting that Gauthier's theory does not provide the content of morality by itself, but simply the rationale for moral behaviour. Thus it can be argued that the existence of moral reasons depends on the occurrence of certain social choice settings in which agents can mutually benefit from a tendency to adhere to certain constraints on their behaviour. It can also be argued that the content of these constraints within any particular society is determined by the desires of its members. On such an account, the rationale for moral behaviour is purely formal, while the content of morality is contingent on the desires of actual agents. I shall assume that Gauthier's account of moral reasons must be supplemented by an account of what agents (generally desire to explain the specific moral reasons which we (take ourselves to) have.

A second limitation, which Gauthier acknowledges, is that his theory does not yield reasons to behave morally towards those who are vulnerable, or towards animals. Hence:

> We may agree that moral constraints arising from what are, in the fullest sense, conditions of mutual advantage, do not correspond in every respect to the 'plain duties' of conventional morality. Animals, the unborn, the congenitally handicapped and defective, fall beyond the pale of a morality tied to mutuality. The disposition to comply with moral constraints...may be rationally defended only within the scope of expected benefit (ibid: 268).

Barry finds this aspect of Gauthier's theory totally undermining, in that 'justice as mutual advantage fails egregiously to do one thing that we normally expect a conception of justice to do, and that is provide some moral basis for the claims of the relatively powerless' (Barry, 1995: 46). We might agree that Gauthier's theory fails to account for some of the more significant cases in which we take moral reasons to apply. This is a major short-coming.

A third limitation is that Gauthier's theory cannot account for moral reasons to be more generous than it would be rational to agree to be when drawing up a social contract from within a presocial position. For example, it seems reasonable to suppose that I have moral reasons (if not obligations) to give a great deal of my wealth to the destitute. These reasons could be explained in terms of a Foot-style concern for the well-being of my fellow humans, but not in terms of the fact that under Gauthier's pre-social conditions it would be rational to agree to give up much of my wealth for the sake of those in need (which, presumably, it would not). So, Gauthier's account leaves out moral reasons to act in certain morally generous ways, assuming that the general acceptance of a commitment to perform such actions would lower agents' expected utility.

This is perhaps less of a worry, as one might wonder whether we really do have moral reasons to act in these ways. Perhaps we can have reasons to perform acts of extreme generosity, but perhaps such reasons are personal rather than moral.

However, the structure of Gauthier's theory does not give any indication of the degree of constraint that it would be rational for people to accept. This is an important, if opaque issue. If expected utility is maximised by a tendency to accept only relatively liberal

constraints on one's actions, then many traditional moral reasons will not be supplied by Gauthier's theory. If the constraints that it would be rational to accept are more stringent, then more of the scope of traditional morality will be covered.

A final, theoretical limitation, is that the account of moral reasons provided by Gauthier depends upon the success of his account of agent rationality in terms of the payoffs of certain tendencies (or dispositions) rather than the payoffs of particular actions. If rational agency is to be understood in terms of the expected utility of certain tendencies, rather than of certain actions, then we can have reasons to adhere to the kinds of constraints that we would adopt in the presocial position. But if rational agency is simply a matter of performing actions which have the highest degree of expected utility, then adhering to the kinds of constraints that it would be rational to agree to in the presocial position is not necessarily rational.

I will not attempt to discuss the merits of a dispositional account of agent rationality here. Rather, the conclusions that I draw concerning Gauthier's account are provisional on the success of his view of agent rationality.[69] These conclusions are: (i) that although the logic of Gauthier-type social choice settings can explain some (and perhaps a large number) of our moral reasons for action, it cannot explain the entire scope of morality, including certain key cases such as reasons for helping/refraining from harming the vulnerable; (ii) even where Gauthier-type considerations do explain moral reasons for action, they only explain their practical logic and not their content. As such, there is a very important role for the content of human desires in accounting for the content of moral reasons for action *viz.* the content of the agreements that it would be rational for us to make in the presocial position.

---

[69] As mentioned in footnote 3, above, Gauthier's account of moral reasons is incompatible with my account of practical reasons, as it stands. The former treats the subject matter of practical rationality as dispositions to perform certain act-types, the latter treats it as intentions to perform certain act-tokens. I see no reason in principle why this incompatibility cannot be remedied, although it would require a considerable (and presumably complex) reworking of my account of practical reasons.

Finally, it is worth noting that moral reasons for action might be over-determined, in that we might have moral reasons both because of the rationality of having certain cooperative tendencies, and because the content of our specific concerns entail that it is rational for us to act morally. As Gauthier mentions, Trivers offers an evolutionary explanation of guilt as a mechanism by which altruistic tendencies are re-inforced (Gauthier, 1986: 187-8). As will be seen, such affective mechanisms are capable of generating reasons to act morally independently of the rationality of any cooperative tendencies that they tend to reinforce.

5. Railton: Moral Reasons as Reasons to do what is Socially Rational

So far we have seen two hypotheticalist attempts at avoiding the troubling idea that, if practical reasons are indexed to the contents of agents' desires, agents' may have no moral reasons for action. According to the first approach, people's normal human concern for others provides reasons for them to behave morally. According to the second approach, the logic of certain social choice settings makes it rational for agents to accept moral constraints on their actions in order to maximally satisfy their desires. Railton proposes a third approach, which involves defining a notion of what is 'socially rational' (where this is seen as equivalent to what is morally right) and then suggesting that this is something that matters to people enough for them to have reasons to promote it (Railton, 1986). This approach differs from Foot's in that it is not a direct concern for others which generates moral reasons; it is a concern for what is rational from an impartial perspective.

Before discussing the existence, or otherwise, of reasons to do what is rational from a social perspective, it will be instructive to explain this notion, and to understand some general features of Railton's account of practical rationality.

Railton's notion of social rationality makes essential reference to his more basic notion of a person's 'non-moral good'. According to Railton, it is possible to give an account of what is good for a person in objective terms. This is 'roughly, what he would want himself to seek if he knew what he were doing' (ibid: 12); that is, what he would want himself to seek if he were fully and vividly informed on all relevant matters and able to deliberate correctly. This

differs from an individual's subjective interest—what he actually wants—in that a person can want something from a position of ignorance about its nature and/or consequences, or because she wrongly believes that it will satisfy some want of hers. Railton gives the example of a homesick and dehydrated traveller, Lonnie. Lonnie wants to drink a glass of milk to settle his stomach, unaware that it will be difficult to digest and thereby worsen his stomach ache and dehydration. However, were Lonnie fully informed and rational (i.e. aware that drinking milk would worsen his dehydration and stomach ache, able to reason properly from means to ends, and so on), he would want himself (his less informed and rational self, that is) to drink clear fluids as opposed to milk. Plausibly, we might say that although drinking milk will satisfy Lonnie's desire for something familiar and comforting, it is not good for him.

There are certain problems with this account of a person's non-moral good. For example, Rosati points out that there are certain cases in which what we would want for ourself under conditions of full information and rationality may not, intuitively, be what is good for us (Rosati, 1995). For instance, a control freak (Sandy) who is deciding whether to undergo therapy to become less uptight may desire most to remain in control, such that her fully informed and rational self would not want her to undergo the therapy. Even so, we might think that therapy is good for her. The problem here is that Sandy's fully informed and rational self would, on Railton's model, share her existing non-instrumental desires. The worry is that some of these desires could be directed at outcomes which cannot properly be considered good for her. Nevertheless, let us suppose for the sake of argument that Railton's account of an individual's non-moral good, or something closely approximating to this account, is correct.

Railton's notion of social rationality is an extension of his concept of individual rationality into the social realm. Thus social rationality is 'what would be rationally approved of were the interests of all potentially affected individuals counted equally under circumstances of

full and vivid information' (ibid: 22). That is, a socially rational course of action is one which will maximally promote the aggregate of individuals' non-moral goods.[70]

According to Railton, social rationality is 'a recognizable and intuitively plausible – if hardly uncontroversial – criterion of moral rightness'. Suppose that we grant this criterion. This entails that moral rightness is to be understood in consequentialist terms, with the relevant theory of the good being provided by the notion of the aggregate of individuals' non-moral goods. Although controversial, Railton's account of the right and the good is by no means outlandish. Further, what is important so far as this section is concerned is not so much whether Railton offers a plausible notion of what is morally right and/or good, but whether he offers a plausible defence of the claim that we have reasons to do what, on his account, is morally right. In short, Railton needs to connect social rationality to hypothetical reasons for action.[71]

As things stand it is possible that there are agents who have no reason to do the moral thing, as this may not correspond to their own objective interest. What is required is an appropriate desire on which to tie such reasons. Railton turns to the idea that acting in a

[70] Railton's account of what is rational from an individual and from a social standpoint conflicts with my own account of practical rationality. This is because he introduces the notion of a fully-informed and rational self in explaining what it is rational for agents to do. On my account, it is the desires and beliefs that agents actually have which determine what it is rational for them to do. Nevertheless, I think that either view could be modified to make the two approaches compatible. Thus an account of what is individually rational framed in terms of agents' actual beliefs and desires could be adopted. The ensuing account of social rationality would then be framed in terms of what would be approved of, were the actual concerns of all potentially affected agents counted equally. Alternatively, an account of practical reasons framed in terms of how it would be rational for our idealised counterparts to intend to act, given *their* beliefs and desires (i.e. what it would make sense for our idealised selves to do), could be adopted.

[71] Railton is an instrumentalist about practical rationality, such that moral rightness is not taken to be reason-giving in its own right (Railton, 1986: 6).

way that is rational from an impartial point of view is something that matters to people. Hence:

> [I]n public discourse and private reflection we are often concerned with whether our conduct is justifiable from a general rather than merely a personal standpoint, it therefore is far from arbitrary that we attach so much importance to morality as a standard of criticism and self-criticism' (ibid: 31).

Because we care about acting in ways which are rational from an impartial perspective, we have reasons to be moral.

The idea that we are concerned about acting in ways which are rational from a social perspective, and not just from an individual one, is the linchpin upon which Railton's account of moral reasons depends. It is at this point that I am least convinced by his account. This is because I am not convinced that people are generally and robustly concerned about acting in ways which are rational from an impartial perspective; not, at least, in Railton's sense. Without such a concern, reasons to promote Railton's moral good are lacking.

As will be seen in section 6, I am largely convinced by the idea that, in general, people care about promoting what they perceive to be just or fair. But this is very different to caring about maximising the aggregate of individuals' non-moral goods. Notions of justice and fairness are more Kantian than consequentialist. The concern is that each person gets what she deserves, not that aggregate social welfare is maximised. Both involve impartiality, but not in the same way. A desire for justice involves an impartial concern for the distribution of outcomes according to desert. A desire for what is socially rational involves an impartial concern for goodness to be maximised, regardless of whose goodness it is. Thus although Railton might be right that we are often concerned with what can be justified from a general

standpoint, this claim is compatible with the idea that pursuing justice, rather than some aggregate notion of the good, is what we consider justified from a general standpoint.[72]

One response here would be to claim that deontological concerns, such as justice, can be included within a consequentialist theory. For instance, it might be suggested that the good includes just outcomes, such that maximising justice is part of maximising aggregate social good. However, Railton's own theory of aggregate welfare, which is framed in terms of maximally satisfying the aggregate of everyone's fully informed and rational desires, does not include concerns such as justice (not in a fundamental sense, at least).

In section 6 I suggest that, among other things, people are generally concerned with the well-being of others. This supposed concern is closely aligned with a concern for Railtonian social rationality. However, there is an important difference. The caring principle discussed in section 6 states that we must always show a concern for others, but not that we must always show a concern for the aggregate of everyone's welfare. It is possible to care about the welfare of each and every person without caring particularly about the aggregate of people's welfare. Thus although it is plausible that we want everyone's welfare to be maximised, this does not entail that we want the aggregate of everyone's welfare to be maximised. I am sure that some people, including many consequentialists, care about the maximisation of aggregate welfare. But I think that it is implausible to suppose that most

[72] Perhaps one might object to a deontic account of what people consider justified from a general standpoint by invoking, for example, the paradox of deontology (Scheffler, 1985: 409). However, this would be a *non-sequitur* as people can have certain ends or consider certain things justified even if those ends cannot be pursued in an entirely consistent manner. Perhaps there is something awry with the very notion that certain outcomes can be deserved at all, just as there is something wrong with the idea that eating Marmite is the sole moral good. In both cases, even if the conception of what is morally justified is wrong, it is still a conception that people can have and can pursue (insofar as it gives clear answers to moral questions). At the limit, adverting to a principle of justice may be unhelpful in resolving whether to commit an injustice in order to prevent a greater injustice. But this does not show that the notion of justice does not operate in people's everyday attempts to act in ways which are justifiable from a social point of view.

people care strongly about the aggregate of people's goods, such that this concern can be seen as the driving force behind morality.[73]

It might be objected that caring about everyone's welfare is equivalent to caring about the aggregate of welfare across society. However, this is false. This can be seen by considering Taurek's controversial claim that in cases where we can save either the many or the few, we should toss a coin to give each person an equal chance of being saved (Taurek, 1977). The decision to toss a coin shows an equal concern for the well-being of each and every person in the supposed scenario; each person gets an equal chance of being saved, as each person's well-being matters equally. Nevertheless, Taurek has no concern for the aggregate social good; he is perfectly prepared for the many to die for the sake of the few, as his concern is with each person's well-being (i.e. with everyone's well-being as individuals) but not with their sum.

It might seem strange to invoke Taurek's unusual and highly counter-intuitive position at this point. Given that most people think that it is obvious that we should save the many and not the few, it seems that most people do care for the aggregate of well-being across society and not simply for the well-being of everyone individually. However, the purpose of citing Taurek's position is simply to show that there is a distinction to be made between caring about everyone's well-being individually and caring about the aggregate of well-being across society; these are not equivalent concerns. With this distinction in mind, it can then be argued that while most people are generally concerned about the well-being of other people, they do not really care about the aggregate of people's well-being, or not very much. Social aggregates are too far removed from the tangible reality of people's daily experience to be the kind of thing that they reliably care about or are motivated by.

This claim is compatible with the fact that most people would choose to save the many over the few simply because we can reasonably suppose that most people would rather save more of those who they care about than less. Although Taurek (ibid: 306-7) finds this

[73] This is an intuition about an empirical hypothesis, of course, and is subject to dispute.

motivation unpalatable, because it involves treating people like cherished objects rather than valuing them as subjects in their own right, it is not psychologically unrealistic to expect people to operate in this way. What is unrealistic, at least as far as I can fathom, is to expect people in general to have a strong concern with maximising a social aggregate, such that this can be seen to be *the* driving force behind our tendency to act morally.

However, perhaps I have been too quick here. Although people may not be directly concerned with maximising the aggregate of non-moral goodness across society, a concern for each person's welfare, combined with the desire for more of those who we care about to do better than less, may amount to a desire by proxy for aggregate non-moral goodness to be maximised. Perhaps we *do* want aggregate welfare to be maximised, though not because we care about it in and of itself. Rather, perhaps we desire the welfare of as many of those who we care about as possible to be maximised, where we just so happen to care about *everyone*.

This seems possible, though it also seems somewhat unlikely that most people are concerned about everyone's well-being. If this were the case, then humans would generally have reasons to do what is socially rational, such that Railton's view of moral reasons could explain at least some (and perhaps even many) of our reasons to be moral (assuming that Railton's notion of social rationality, or something akin to it, is viable).

Nevertheless, Railton's account of moral reasons cannot explain all moral reasons. Specifically, it cannot explain reasons to, say, promote justice or keep commitments when doing so conflicts with what is socially rational. That we have social concerns other than maximising aggregate welfare, such as for justice and faithfulness to commitments, suggests that there are reasons of justice and commitment which exist even when these conflict with the Railtonian good (at least on a hypotheticalist account, where reasons are indexed to desires). To the extent that such reasons exist, Railton's account does not explain the whole of morality. However, Railton can furnish reasons to perform many moral actions, on the assumptions that a reasonable notion of the social good can be provided and that we have sufficient concern with it. This does not imply that Railton's account of moral reasons, if at

all successful, is in competition with other hypotheticalist accounts. Rather, as suggested at the end of the previous section, the over-determination of moral reasons is perfectly acceptable where there is overlap between accounts.

A final point: despite my claim that the maximisation of aggregate social welfare is too abstract a concern for people to have in general, there is some psychological evidence to suggest that people *are* concerned with this (Koenigs *et al*, 2007; Greene, 2007). This evidence shows that patients with damage to the ventromedial prefrontal cortex (VMPFC) area of the brain (which is associated with the experience of moral emotions) are more likely to make utilitarian moral judgements than normal subjects. This finding is consistent with the claim that people do have a concern for aggregate welfare, where this concern is tempered to some degree by affect-based moral reasoning in normal persons. As such, a hypotheticalist account of moral reasons may need to include agents' utilitarian concerns at some level.

However, Moll and de Oliveira-Souza question this inference, suggesting that 'that VMPFC patients make more prosocial choices (from a utilitarian perspective) is a reminder of the gulf that divides observable behaviors and internal motivations. The apparently 'prosocial' choices of VMPFC patients might reflect a lack of prosocial feelings' (Moll and de Oliveira-Souza, 2007). That is, perhaps it is not that we generally care about aggregate goodness at all. Perhaps, rather, when presented with certain kinds of scenario we choose a utilitarian outcome (which seems more rational) to the extent that we lack any of the emotions or concerns which would lead us to make some other moral choice. It might simply be that utilitarian choices result from generic rational processes, when plugged into certain scenarios in which we must select between competing moral actions, rather than from a specific concern for people's welfare.

This suggests a cautious agnosticism about the role of Railton's account in explaining moral reasons for action. Such an account can potentially explain many moral reasons for action, but only if we have an appropriate concern with the aggregate good (either directly, or because we care about everyone individually and want more of the people who we care

about to do well). It is not clear whether or not we have this concern (at least in a robust enough sense to explain moral reasons).

6. Moral Reasons as Empathy-Based Hypothetical Reasons

In this section I present my preferred response to the objection from moral reasons. This response is grounded in our capacity for empathic affect. The main purpose of the section is to offer support for two claims: (i) that human-beings, in general, have desires directed towards the well-being of others; (ii) that these desires generate reasons to act in ways which we would typically call moral. This involves two tasks. First, to provide psychological evidence that human-beings are, in general, motivated by a concern for each other's well-being. Second, to show that this kind of motivation translates into reasons which correspond, more or less, with the central content of morality.

I will also attempt to provide some motivation for thinking that the kinds of other-regarding desires attributed to humans are part of the desire-set core to interpretable agency. However, my thesis does not depend upon this claim and a full defence of it goes beyond what I can offer here.

One way of supporting the claim that human beings are, in general, concerned about each other's well-being is to invoke their capacity for empathic affect. Empathic affect can be defined as 'an affective response more appropriate to another's situation than one's own' (Hoffman, 2000: 4). In what follows, I discuss the proposal that a capacity for empathic affect can generate the kinds of other-regarding motivations required to support a hypothetical account of moral reasons. I use Hoffman's book *Empathy and Moral Development*, as the primary basis for the psychological claims made in this section (ibid). In this book Hoffman draws on both his own, and others' research on our capacity for empathic affect and its role in motivating pro-social behaviour. He attempts to offer a systematic account of the role of empathy in pro-social behaviour. This account is well supported by the available psychological evidence and is presented in a specifically moral

context. This makes it ideal for the purposes of illustrating how a hypothetical account of moral reasons might be grounded in a capacity for empathic affect.

Hoffman's account (and the preceding work on empathy that it draws upon) is, of course, subject to criticism and debate (see e.g. Batson *et al*, 1987). I attempt to avoid engaging in matters of substantive psychological debate here. Such matters are important when it comes to the final nature of an account of moral reasons framed in terms of our other-regarding motivations. However, contentious issues can, by and large, be treated as matters of detail so far as the illustrative concerns of this section go.

Hoffman identifies five mechanisms through which empathic affect can be aroused. Three of these mechanisms (motor mimicry, classical conditioning and direct association) are automatic and involuntary; two are 'higher-order cognitive modes' (mediated association and role-taking).[74] As humans mature, their capacity for empathic arousal develops in sophistication from the simple, automatic modes to include the complex, higher-order ones.

This range of empathy-arousal modes means that empathic affect is reliably induced across a very broad range of victim-distress situations. Although children under a certain age may be unaware, for example, that the well-being of a joyful child can be compromised if, say, she is in dire poverty, mature adults *are* able (through role-taking) to empathise with, and consequently show concern for people even when they are not in any immediate distress (Hoffman, 2000: 90-91). As Hoffman summarises: 'the importance of many modes of empathic arousal is that they enable observers to respond empathically to whatever distress cues are available' (ibid: 59).

As well as being reliably aroused by a range of distress cues, there is strong evidence that empathic affect functions as a prosocial motive (a motive to act in ways which promote others' interests but not necessarily one's own). Regarding innocent bystander cases, Hoffman summarises this evidence under three categories: evidence that empathic distress

[74] For full explanation and discussion of these modes of arousal, see Hoffman (2000: ch2).

is associated with helping; evidence that empathic distress precedes helping; and evidence that observers feel better after helping (ibid: 30-36). The preponderance of evidence in each case suggests that empathic distress is a strongly prosocial motive.[75]

Of course this does not mean that an innocent bystander will always help. Self-interest may prevail, or people may even take measures to avoid empathising (blaming the victim, looking away etc.). There is also debate over whether empathic affect is an altruistic or an egoistic motive (see, e.g. Batson, 1991). The issue here is whether people help in order to make themselves feel better, or whether they help because they want others to feel better (or to otherwise benefit from their actions). However, even if helping behaviours are not entirely altruistic, Hoffman claims that empathy-based helping is prosocial 'because it is instigated by another's distress, not one's own, its primary aim is to help another, and one feels good only if the victim is helped [and not simply because one has tried to help]'

[75] Batson *et al* (1987) have criticised the idea that empathic affect is essentially a kind of distress. Rather, they suggest that empathy and personal distress are different vicarious emotions, with different motivational consequences. Empathy, they suggest, is associated with an altruistic concern for others' welfare whereas distress is associated with an egoistic concern to end one's discomfort. However, Hoffman and Batson *et al* may be talking past each other somewhat here. Hoffman focuses on empathic affect as a *capacity* to feel emotions which are more appropriate to others' situations than one's own; Batson *et al* focus on empathy as *a particular kind of vicarious emotion* (one involving other-regarding concern). To the extent that a capacity to empathise with others in specifically moral contexts might often involve 'feeling' their distress (or even feeling distress 'on their behalf'), Hoffman's characterisation of empathic affect as often involving experiences of empathic distress does not seem to be misplaced, nor does his suggestion that such distress can be a prosocial motive.

Nevertheless, the suggestion that it is feelings of empathic distress, rather than other kinds of empathic emotion, which always motivate prosocial behaviour does conflict with the findings of Batson *et al*, who show that a concern for others (what they call empathy, and what might otherwise be called sympathy) is a prosocial motive independently of distress. Hoffman characterises sympathy as sympathetic distress. This seems to be wrong, given the findings of Batson *et al*. Hoffman's failure to distinguish between sympathy and distress should be kept in mind where Hoffman's claims about the prosocial role of sympathetic distress are concerned.

(Hoffman, 2000: 35). Thus even if the pro-social motivations which result from empathic affect are egoistic, in the sense that they reflect a concern to reduce one's own distress, this does not prevent them from generating moral reasons—reasons to, say, help others because one finds their situation concerning. Further, if Batson *et al* (1987) are right that empathy (sympathy, in Hoffman's terms) and distress are distinct emotions, and that the motivations engendered by empathy/sympathy are more purely altruistic, then feelings of empathy/sympathy are consistent with agents having more directly moral concerns.

Having established that empathic affect functions as a prosocial motive, Hoffman goes on to outline a number of different types of empathic distress. For innocent bystanders these are: sympathetic distress (where the cause of suffering is accidental/beyond the victim's control); empathic anger (where the suffering is a result of malice); bystander guilt (where the bystander could have helped/prevented the suffering but did not); and an empathic feeling of injustice (where it is felt that the suffering was undeserved). Each of these feelings is associated with different motivations. Specifically, sympathy is associated with a motive to help; anger is associated with a motive to punish; bystander guilt is associated with a motive to change one's future behaviour; and empathic feelings of injustice are associated with a motive to right the wrong (Hoffman, 2000: ch4.).

For transgressors (those who harm another in some way) the primary empathic affect is guilt over having caused such harm (ibid: ch5.). Unlike the empathic affects associated with innocent bystanders, there is evidence that transgression guilt needs reinforcing during childhood through parental inductions if it is to be consistently experienced. A tendency to focus on their own needs can 'blind them [children] to the harm done and override their empathic tendencies' (ibid: 135). Hoffman suggests that children need their attention bringing to any harmful effects of their actions on others in order for transgression guilt to be aroused. Where this routinely happens, children internalise moral norms against harming

others (beginning to anticipate that such harm will lead to a feeling of guilt), and eventually develop a tendency to avoid transgressions voluntarily.[76]

Finally, Hoffman introduces the idea that moral principles are connected to empathic affect, serving to place incidents where empathy is aroused within a wider cognitive context (ibid: ch.9). Such principles help to limit the influence of certain empathic failures, such as empathic over- or under-arousal and empathic bias. They also help to resolve conflicts in complex cases.

Two general principles are examined by Hoffman and linked to empathy: the caring principle and the justice principle. The caring principle states that we must always consider others' well-being. This principle is directly and straightforwardly connected to our capacity to feel empathic distress at others' suffering. When we are motivated by empathic distress to help another, this motivation represents a concern for their well-being. The caring principle generalises this concern; it is a generalised psychological edict to show concern for others' well-being.

The justice principle is more complex, both to state and to connect to empathic affect. Roughly the idea is that reward should correlate with desert, where desert can be understood in a number of ways including merit (effort, competence, productivity), need, and equality. Hoffman cites evidence which suggests that empathy is positively correlated with a preference for need-based justice (ibid: 230). Often we do not simply feel bad for those in need, we feel that they do not deserve to be in such a position. Abstracting and generalising these feelings yields a need-based principle of justice.

Things are more complicated than this, however. Hoffman points out that different justice principles are not necessarily inconsistent with each other and that they can be combined

[76] This raises an issue (discussed below) over the contingency of reasons not to harm others on one's degree of socialisation.

such that one principle is primary with others acting as constraints. For example, merit can be treated as primary, but constrained by certain need- or equality-based parameters.

This leads Hoffman to an interesting suggestion regarding Rawls theory of justice, which centres on the 'difference principle' (Rawls, 1971: 303). Hoffman characterises this as a merit- (productivity) based system of justice, constrained by need, whereby the 'merit-based distribution of society's resources is acceptable only if the resulting economic inequalities operate to the greatest benefit of society's least advantaged' (Hoffman, 2000: 232). Although Rawls attempts to derive the difference principle from pure rational self-interest, operating under conditions of ignorance, Hoffman suggests that the difference principle can also be extracted from our empathic tendencies. Thus he suggests that 'empathy and the veil of ignorance...are functionally equivalent regarding matters of justice: They both constrain self-interest, though in different contexts' (ibid: 235). The veil of ignorance constrains us within a particular theoretical context such that we must consider every position in society in drawing up a principle of distributive justice. Empathy constrains us in the social domain such that we must consider others in deciding on a fair way to act. This leads Hoffman to claim that 'only empathy can provide the internal motive basis for acting in accord with the [difference] principle and promoting institutions that embody it' (ibid: 236).

Whether or not Hoffman specifically establishes that the difference principle is an appropriate theoretical correlate to our empathic motivations, he does offer significant support for the general idea that moral principles act as cognitive generalisations of our empathic motivations. We do not simply respond to empathic distress as stimulated by others' distress cues, or anticipated from past experiences, we also form generalisations about how to act which regulate our behaviour. Not that this process is entirely original to each person. 'The child does not construct a moral code anew...but is active nonetheless in reconstructing and understanding moral rules on the basis of information obtained from adults, peers, the media, and his or her own experience' (ibid: 260).

Given the role of empathy in generating prosocial actions, both directly through empathic affect and indirectly through the acceptance of moral principles, it seems reasonable to conclude both that human-beings are concerned about each others' welfare, and that these concerns can be aligned with reasons to act in ways which would typically be considered moral.

In the introduction to this chapter I listed some typically moral actions. These were: (i) helping those in need; (ii) refraining from harming others; (iii) carrying out our commitments towards others; and (iv) supporting/upholding just social outcomes. On the model of practical reasons offered in this thesis, reasons to perform each of these act-types follow directly from the motivations generated by our capacity for empathic affect. Taking each in turn:

(i) The sympathetic distress caused by witnessing the suffering of another generates a prosocial motivation to help them; this motivation (or its associated desire) is, on my account of reasons, partly constitutive of a reason to help that person (together with certain associated *means-end beliefs*).

(ii) The anticipated guilt of acting harmfully towards another person generates a prosocial motivation to refrain from doing so; this motivation (or its associated desire) is partly constitutive of a reason to refrain from acting harmfully towards them (together with certain associated means-end beliefs).

(iii) The anticipated guilt of failing to fulfil a commitment, which could result in difficulty or suffering for other people, generates a prosocial motivation to prevent them from experiencing this difficulty or suffering; this motivation (or its associated desire) is partly constitutive of a reason to keep the commitment to those people (together with certain associated means-end beliefs).

(iv) An empathic feeling of injustice, which arises where we observe a social outcome that we believe is undeserved (i.e. not in accordance with need, merit,

or equality), generates a prosocial motive to bring about the deserved outcome; this motivation (or its associated desire) is partly constitutive of a reason to promote/uphold the just social outcome (together with certain associated means-end beliefs).

In addition, reasons to perform each of the four act-types listed above can also be derived from the moral principles of caring and justice, where we are motivated to uphold these principles because they are 'affectively charged' through their association with our empathic responses to others. Taking justice as an example, if suitably internalised a principle of justice functions as what Hoffman calls a 'hot cognition', or an 'affectively charged cognition' (ibid: 239-41). Such a principle is 'activated' when an appropriate empathy-inducing situation is experienced.

For example, Hoffman cites the example of a white southern schoolboy (in the US), who witnessed a black schoolboy being repeatedly victimised by his white peers (ibid: 239; full details, ch.4). In this case, the schoolboy's empathic distress at what he witnessed activated his justice principle, and 'gave that principle motive force' such that he set aside the prevailing, racist social conventions of the time to become in favour of ' "an end to the whole lousy business of segregation" ' (ibid: 240). In this case a desire to see a peer treated fairly led to a generalised motivation to see a whole race of people treated fairly. The motivation for upholding a moral principle was thus caused by the activation of the schoolboy's capacity for empathic affect.

Because of their generalised content, moral principles (where suitably internalised and, thereby, affectively charged) can extend the range of reasons generated by empathic responses. We do not need to feel empathy for specific individuals in order to have reasons to help them, to avoid harming them, to carry out any commitments that we may have or to promote just social outcomes. If we are motivated to uphold moral principles in any of these areas then we have reasons to act in moral ways regardless of whether we are currently empathising with somebody in distress. This capacity for holding affectively charged moral principles therefore suggests that much, if not all, of traditional morality can be

accommodated by the model of practical reasons that I propose. Given this potential, it seems reasonable to maintain that a hypothetical account of moral reasons can be given which identifies moral reasons as those associated with our empathic responses to the situations of others, and with the prosocial motivations tied to these responses.

However, there are problems. Much of the empathic story that I have given relies upon socialisation, which suggests that certain moral reasons might be limited to those who have been appropriately socialised. Specifically, according to Hoffman, significant feelings of guilt over transgressions towards others tend to occur only where individuals are socialised in such a way that they pay sufficient attention to the harmful effects of their actions on others. Likewise, Hoffman proposes that moral principles in general are internalised largely through a process of socialisation (ibid: ch10.). If this is right then I might be forced to accept that anti-social persons have no *moral* reason to refrain from harming others or, in general, to act in the interests of others and/or justice. The only residual moral reasons applying to anti-social persons would be reasons to help specific persons in distress, if and when such distress was sufficiently registered to produce an empathic response, together with its corresponding prosocial motivations.

This is a troubling problem. If I am proposing to attach moral reasons to desires for others' well-being, then I must accept that no such reasons exist where the relevant desires are lacking. At the same time, it seems wildly counter-intuitive to suggest that anti-social individuals have no moral reason not to harm others, or to be kind and just in general. The only strategy for response that I can think of is to bite the bullet, while trying to show that it is not as threatening to my approach to moral reasons as it would initially seem.

It is quite rare for there to be individuals who receive no degree of moral socialisation. Most children are disciplined many times by their parents/guardians, and by other adults who are responsible for their care at some stage or other (teachers, childminders, relatives etc.). According to evidence cited by Hoffman, 'children in the 2- to 10-year age range experience parental pressure to change their behavior every 6 to 9 minutes on average' (ibid: 141). The main issue, then, is not over a lack of disciplinary exchanges between parents/other carers

and children. Such exchanges are practically unavoidable. Rather, the issue is over the content of these exchanges. Children often develop anti-social tendencies not because of a lack of disciplinary input, but because the way they are disciplined does not 'highlight the other's perspective, point up the other's distress, and make it clear that the child's action caused it' (ibid: 143). Thus:

> For inductions to work, their message must get through to the child despite the child's involvement in pursuing his or her own goals and the emotionality of the situation. This requires a certain amount of external pressure – enough to get the child to stop what he or she is doing, attend, and process the induction but not enough to arouse undue anger and fear, which can disrupt the processing (ibid: 144).

For children to internalise moral norms, they must be disciplined in the right way: a way that points out the consequences of their actions but not so strongly as to prevent them from inferring the intended moral message.

Clearly the potential for children not to receive *appropriate* disciplinary intervention is high. Even if most children are disciplined in such a way that they internalise moral norms and principles, there is good reason to suppose that there are many who are not. Of these, some will become anti-social adults—adults who fail to show appropriate concern for the well-being of others and for just social outcomes. The problem for my approach to moral reasons is that it seems to involve accepting that such individuals lack moral reasons to show any concern for others.

One way of mitigating this outcome is to suggest that although socialisation through discipline is very important in the internalisation of moral norms and principles, there are other ways in which such norms and principles can be internalised. For example, as Hoffman accepts, peer exchanges in which children have to learn how to get along, despite their conflicting wants, can play a key role in moral internalisation (ibid: 257). However, he rejects Piaget's claim that parents corrupt children's moral internalisation because of their ability to wield power over them (Hoffman, 2000: 256-7; Piaget, 1932). Rather, Hoffman suggests that

'it may only be the parents, that is, nonabusing, inductive parents, in a sort of "coaching" role...who can set the stage and get children ready for the unique benefits of peer interaction' (Hoffman, 2000: 257).

Debates over the extent to which children can internalise moral norms and principles without appropriate inductive intervention aside, there is no avoiding the implication that *for those who genuinely do not care a fig about others*, there are no moral reasons, duties or constraints which apply to their behaviour on the proposed account. For some, the suggestion that individuals of this kind are relatively few and far between will be far too weak to overcome the intuition that anyone, anti-social or not, has moral reasons for action.

To respond to such complainants I must draw on support from outside the present account of moral reasons. One source of support is Gauthier's moral theory, according to which even those who do not care about others can have reasons to be moral, where a tendency to abide by moral constraints is associated with higher expected utility than a tendency not to. Thus Gauthier's theory can provide a sort of moral safety net that explains why even those who fall short of the normal moral concerns of appropriately socialised individuals have reasons to adhere to many (though not all) of the standard moral constraints. In short, having such a tendency is in their interest.

A second problem with my approach to moral reasons is that, given my account of practical reasons, it indexes the strength of our moral reasons to the degree of concern we have for the well-being of others, and for upholding caring and justice principles. There may only be a few anti-social agents who lack any concern for these things whatsoever, but perhaps there are many more people who do not care about them very much. Moral reasons, on my view, are not universally weighty and it is possible that, for a great many people, they have little weight at all.

I must bite this bullet. The only mitigation I can offer is to point out that a great many people place a great deal of importance on the moral domain. As such, many people have strong reasons to behave morally. If this, in conjunction with the moral safety net offered by

Gauthier's account, is not enough to overcome the intuition that moral reasons must be significantly (and perhaps even equally) weighty for everyone, then all I can do is to point out that I have offered a plausible account of reasons that many people have for behaving morally. This is the best that I can do, given my theoretical commitments.

I now turn to the claim that desires for the well-being of others form part of the desire-set core to interpretable agency. This claim might look particularly indefensible, given the preceding discussion. If some humans can lack this desire, how can it be core to interpretable agency itself? The answer is that not every agent must hold all of the desires which are in the desire-set core to agency. The account of interpretable agency is holistic; agents must, by and large, have desires which fall within a certain set. They might lack some of these desires while remaining interpretable, but they cannot stray too far without becoming unintelligible.

With respect to human agents, we seem able to make sense people who lack any concern for the well-being of others (we can make sense of psychopaths, for example). It also seems that we can make sense of people who lack any direct concern for their own well-being, but who are concerned about the well-being of others (we can make sense of figures like Jesus, for example). What we cannot do is to make sense of people who are neither concerned about their own well-being, nor about that of others. People in this category (the clinically insane, for example) are unpredictable and beyond interpretation. What this is supposed to show is that, as long as a person has desires towards someone's well-being, we seem able to interpret them. So, it seems that well-being desires of *some kind* must feature in the desire-set core to interpretable agency (assuming, for Davidsonian reasons, that interpretability by humans is a test of interpretability in general—i.e. that if some creature could not be interpreted by a human they could not be interpreted at all, given that interpretation depends on substantive agreement such that agents inhabiting radically different conceptual schemes could not exist (Davidson, 1974b: 197)).

This does not suffice to show that a desire for others' well-being is core to agency, or that a desire for our own well-being is either. What it shows is that a desire for somebody's well-

being must be held by any interpretable agent. It requires further argument, which I will not attempt here, to show specifically that desiring others' well-being is core to agency. Such a desire is congruent with interpretability, while some well-being desire or other is (if I am right) required for a creature to be interpretable. But perhaps the desire core to agency is only the desire *that somebody's well-being be promoted*, where this desire is non-committal between one's own and other agents' well-being.

Furthermore, even if I could provide some further argument to show why we must, on the whole, desire other agents' well-being rather than simply *some* agent's well-being, there are certain issues with the argument as it has run up to now. Specifically, can we *really* make sense of a person who only desires others' well-being (or the well-being of some specific other person, for that matter)? Or must such a person be a myth? Is well-being even the type of concept that could be applied to all conceivable agents?

On this last issue, one might argue that any interpretable creature must have preferences, such that an account of well-being framed in terms of preference satisfaction would apply to any conceivable agent. But whether such an account of well-being is correct is a matter of contention (Griffin, 1986: esp. ch. 2; Parfit, 1987: 493-502). If it is not, then the concept of well-being as it applies to humans may not be applicable to agency in general, such that desiring well-being may not be core to agency in any sense at all.

I will not attempt to address these issues further here. All that I aim to have shown is that it is not totally implausible to regard desires for the well-being of others as part of the desire-set core to agency and, therefore, that moral reasons (at least of the hypothetical sort that I have described in this section) could conceivably be tied up with agency in general rather than with human agency in particular. However, further defence from the (possible) objection that my view ties moral reasons too closely to human psychology may be required, if that objection is to be convincingly set *aside*.

7. Conclusion

In this chapter I have attempted to show that a plausible hypotheticalist account of moral reasons can be given, such that my account of practical reasons is released from the charge that it excludes the possibility of moral reasons for action. Having set aside Schroeder's view of moral reasons in section 2, I have discussed four hypotheticalist alternatives, the last being in many ways a development of the first. I have suggested that reasons to perform the core types of moral actions stipulated at the beginning of the chapter can be derived from Gauthier's account and from the empathy-based account of moral reasons offered in section 6. Of these, the empathy-based account is preferred as it affords moral reasons to show significant levels of concern for others, and for abiding by moral principles, for normally socialised individuals.

The reasons supported by Gauthier's account extend only as far as the expected utility gained from a tendency to abide by moral constraints. Therefore, Gauthier's account is unable to capture the full scope of traditional morality. However, his account does offer a safety net of moral reasons for those who are insufficiently socialised to exhibit normal degrees of moral concern for others. Railton's account fails to provide reasons to keep commitments or promote just outcomes where these conflict with aggregate welfare. Further, it does not seem to be as well grounded psychologically as the empathy-based account, given that the extent of our concern for aggregate welfare is currently unclear. Nevertheless, to the extent that we do have such a concern, there is no reason why Railton's account cannot feature in a pluralistic hypotheticalism about moral reasons too. On such a view, moral reasons would derive from a number of different sources of other-regarding concern: mutual benefit; empathy-based concerns for the well-being of others; and (possibly) a general concern with aggregate welfare.

In any case, further development of the empathy-based view proposed in section 6 would require a thorough investigation of our concerns for others' well-being, and with upholding the kinds of moral principles which are associated with a capacity for empathic affect. The

aims of such an investigation would be: (i) to provide a fuller account of the kinds of emotional and motivational responses which arise in relation to our capacity for empathic affect; (ii) to establish how widespread, deeply rooted and strong our empathy-based moral concerns are; (iii) to establish a more detailed understanding of the kinds of moral reasons that we might have, given these concerns. This would allow for a more sophisticated account of the kinds of moral reasons that a capacity for empathic affect might generate, and of the significance of these reasons for those who they apply to.

In the absence of such research, I remain optimistic that a plausible hypothetical account of moral reasons can be provided, given the resources available to the hypotheticalist. As such, I am confident that an account of practical reasons which treats all such reasons as hypothetical is compatible with the existence of moral reasons for action.

Conclusion

The aim of this thesis was to offer an account of practical reasons which: (i) maintained that such things exist; (ii) conformed with a naturalistic worldview. My approach to meeting this aim was to develop an interpretivist account of practical reasons, as outlined in chapter 2. On this account, practical reasons are belief-desire subsets from which certain intentions to act and, derivatively, certain actions follow, given the constraints of interpretable functioning. As well as outlining the proposed account of practical reasons, chapter 2 explained its motivation, restricted the focus of the account to interpretable *mental* functioning, and put forward an accompanying, interpretivist account of practical reasons' normative force.

The motivation for focusing on interpretability was twofold. First, I suggested that interpretability plays an important role in our practical lives, both with respect to our interactions with others, and with respect to our own decision-making processes. Second, I suggested that constraints on interpretable functioning can be seen to pair attitudes with actions in much the same way that practical reasons are paired with actions by the relevant normative scheme. The ensuing proposal was that an account of practical reasons can be mapped on to an account of interpretable functioning, such that reasons can be regarded as sets of attitudes from which certain actions follow, given the constraints of interpretable functioning. Thus interpretability seemed to be a promising candidate for giving a reductive, naturalistic account of practical reasons.

Having motivated my interpretivist approach to practical reasons, I suggested that it is interpretable *mental* functioning which is the key element in such an approach, given that the causal elements of interpretability do not seem apt to feature in an explanation of what practical reasons are. Interpretable mental functioning was spelled-out in terms of rationality. This lead to the charge that, on a Davidsonian view of radical interpretation,

rationality is invoked *qua* independently normative scheme. Following Timothy Schroeder, I denied this charge.

I then set out to reductively explain the normative force of practical reasons in terms of agents having a constitutive commitment to function interpretably. Drawing on Korsgaard's suggestion, this commitment was understood in terms of two features: necessity and inescapability. First, it is a necessary feature of interpretable functioning that agents generally perform the actions which rationally follow from the attitudes that they hold. Second, interpretable functioning is inescapable in that, for anyone who is an interpretable creature, adhering to the constitutive constraints of interpretable functioning is unavoidable. Any attempts to behave uninterpretably are necessarily self-defeating (i.e. involve adhering to the constraints of interpretable functioning which the agent is trying to avoid). Given these two features, my proposal was that the normative force of practical reasons can be reduced to its being a necessary feature of our inescapable mode of functioning that we generally do what we have (most) reason to do.

Having set out the account of practical reasons in chapter 2 and clarified the nature of rational functioning in chapter 3, I considered two key objections to the proposed view. The first of these objections, discussed in chapter 4, was the suggestion that interpretation involves a commitment to irreducible normativity. This objection took two different, but related forms. The first was the objection that meaning is irreducibly normative. The second was the objection that preference is irreducibly normative.

Regarding the objection that meaning is irreducibly normative, I followed Glüer and Wikforss by suggesting that an account of semantic correctness in terms truth can be provided, where this need not be normative in kind. This seems compatible with a Davidsonian approach to language, on which meaning is explained in terms of truth. For Davidson, the truth conditions for sentences uttered are established by a process of charitable interpretation. Charity in interpretation involves maximising an interpreted speaker's degree of truth and consistency, given one's own beliefs and notion of consistency. The claim that charity is itself a normative constraint on interpretation was

resisted. I suggested that charity is a constitutive constraint on interpretation which can be spelled out in non-normative terms. That is, finding speakers to be largely true and consistent can be regarded as a necessary feature of interpretation, where truth can be given a Tarskian treatment and consistency can be characterised in formal, logical terms.

Regarding the objection from the normativity of preference, I suggested *contra* Hurley that substantive constraints on the eligibility of preferences need not take the form of objective values. Rather, such constraints might be regarded as a brute feature of interpretability. That is, it might be a brute fact about interpretation that it involves attributing certain kinds of preferences over others. One potentially troubling objection to this approach was that it is incompatible with there being certain essentially normative distinctions (such as between fairness and unfairness) which are nevertheless eligible to feature in agents' preferences. My response here was to suggest that if the descriptive component of such distinctions can be separated from their evaluative component, the former can be treated as determining preference independently of the latter. This response depends on a non-centralist conception of value, on which thin evaluative kinds (such as goodness) are to be explained in terms of their relation to thick ones (such as fairness).

Perhaps the brute eligibility proposal may seem unappealing. On this proposal interpretation simply involves attributing preferences for certain kinds of things rather than for others, where no explanation of why this is so can be given. However, the alternative appears to involve accepting that value *is an irreducible* normative constraint on preference. In turn, this seems to involve accepting that the normative force of value is itself a brute feature of reality. I take this option to be less appealing, especially given my naturalistic bias. Science can tolerate primitives, but not primitive normative relations (at least not where these relations are taken to explain behaviour).

In chapter 5 I considered the objection that my account of practical reasons excludes the possibility of moral reasons for action. I addressed this objection by considering four available accounts on which moral reasons are hypothetical. Having rejected Mark

Schroeder's view of moral reasons, I accepted that an account of moral reasons as categorical is not available on my view.

The first three proposals were from Foot, Gauthier and Railton respectively. I suggested that each of these views has merits but claimed that none of them can completely account for our moral reasons for action. Specifically, it was suggested that Foot's view of morality as a system of hypothetical imperatives was underdeveloped, in that it failed to explain why our desires generate reasons with the appropriate moral content. Gauthier's view of morality as a kind of mutual advantage was seen to have serious limitations, particularly in terms of accounting for moral reasons to help those who are vulnerable. Railton's view of morality as what is socially rational depended on a desire to maximise aggregate welfare in order to supply moral reasons for action. It was argued that this kind of desire is unlikely to provide a general explanation of agents' moral reasons, in that many people may lack such a desire, at least in any robust sense.

I proposed a fourth species of hypothetical view, on which moral reasons are explained by desires associated with our capacity for empathic affect. This capacity is: a natural feature of human functioning; one which can be reliably triggered in a number of different situations and in a number of different ways; one which can produce a range of different prosocial motivations, depending on the *circumstances*. Further, empathic responses seem able to explain a motivation to uphold moral principles, such as Hoffman's caring and justice principles. This suggests that empathic motivations may be able to explain a large number of our moral reasons for action.

However, a limitation to the empathy-based account of moral reasons is that people's capacity to respond empathically in certain kinds of situations depends on their degree of socialisation. Thus it is possible that individuals who have not received appropriate inductive interventions when young may not feel (sufficiently) motivated to avoid causing others distress, for example. The threat is then that on the proposed view, such individuals may lack reasons to avoid harming others.

This is a bullet which I had to bite, given my account of moral reasons. However, I do not think that it necessarily undermines the proposed account; we can doubt the intuition that our own judgements about others' moral reasons necessarily reflect what they have reasons to do. In any case, I proposed that a moral safety net, of sorts, could be provided by Gauthier's theory of morality as mutual advantage. For anti-social individuals, reasons to behave morally can still exist where a tendency to do so is of benefit to them (although not when dealing with vulnerable individuals).

It seems inevitable that an account of moral reasons as hypothetical in kind will be unable to provide the sorts of universal moral reasons that some may demand. Nevertheless, I think that it is too demanding for a theory of practical reasons to be expected to allow for the existence of reasons of this kind. So long as some plausible account of reasons to be moral can be provided, I take the objection from moral reasons to be dissolved.

Having summarised the view of practical reasons proposed in this thesis, and explained my responses to a couple of major objections to it, I now wish briefly to discuss the prospects for an interpretivist view of practical reasons, and the outcomes of this thesis. I hope to have shown that an interpretivist approach to practical reasons is a plausible naturalistic option. The approach, at least as I have developed it, is heavily dependent on Davidsonian views about interpretability, intentionality, rationality and action. To the extent that these views are contentious, the view of practical reasons that I have developed is also controversial. However, for philosophers of a Davidsonian bent, I hope that the proposed view has some appeal.

I also hope that some of the more general points about practical reasons that I have made will be taken to hold, independently of the viability of the proposed view of reasons. For instance, I hope that the distinction between mind-dependent and mind-independent accounts of normative force (drawn in chapter 1 section 4) will be seen as relevant to any account of practical reasons. I hope that the contention that reductivist versions of the constitutivist approach to normativity are immune to Enoch's schmagency worry will be accepted (as discussed in chapter 2, section 5). I hope that my claims about the

psychological priority of continence over maximisation, and about the lack of any clear *rational* priority between these two principles (as made in chapter 3) will be agreed to. I hope that the claim that substantive constraints on the eligibility of preferences need not be normative will be accepted (chapter 4 section 3). And I hope that the account of moral reasons suggested in chapter 5 section 6 will be considered to be an interesting and plausible view of such reasons, independently of the general account of practical reasons that I defend.

More importantly, I hope that it will be agreed that, given the general Davidsonian approach to language, mind and action that I adopt, the account of practical reasons that I offer is promising. Developing this view more fully would involve (among other things): (a) further consideration of how plausible it is to adopt a non-normative reading of Davidson's theory of radical interpretation; (b) further consideration of the possibility of offering an account of normative force in terms of agents' having a constitutive commitment to function in accordance with the requirements of interpretability; (c) further consideration of whether a conceptually or merely ontologically reductivist approach is best for the interpretivist about practical reasons. Despite having not fully addressed these issues, I hope that it will be agreed that an interpretivist approach to practical reasons is worthy of consideration and that, if successful, it promises to avoid the worry that practical reasons cannot feature in a naturalistic worldview.

Bibliography

Audi, R. (1979) 'Weakness of Will and Practical Judgment' in *Noûs*, 13 (2): 173-196.

Audi, R. (1990) 'Weakness of Will and Rational Action' in *Australasian Journal of Philosophy* 68 (3): 270-281.

Baier, A. (1985) 'Rhyme and Reason: Reflections on Davidson's Version of Having Reasons' in LePore, E. and McLaughlin, B. P. (eds.) *Actions and Events: Perspectives on the Philosophy of Donald Davidson*, Oxford: Blackwell: 116-129.

Barry, B. (1995) *Justice as Impartiality*, Oxford: Clarendon Press.

Batson, D.C., Fultz, J. and Schoenrade, P. (1987) 'Distress and Empathy: Two Qualitatively Distinct Vicarious Emotions with Different Motivational Consequences' in *Journal of Personality*, 55 (1): 19-39.

Batson, C. D. (1991) *The Altruism Question: Toward a Social-Psychological Answer*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Beardman, S. (2007) 'The Special Status of Instrumental Reasons' in *Philosophical Studies*, 134 (2): 255-287.

Bentham, J. (1988) *The Principles of Morals and Legislation*, New York: Prometheus Books.

Bermudez, J. L. (2009) *Decision Theory and Rationality*, Oxford: Oxford University Press.

Blackburn, S. (1984) *Spreading the Word*, Oxford, Clarendon Press.

Blackburn, S. (1992) 'Through Thick and Thin' in *Aristotelian Society Supplementary Volume*, 66: 285-299.

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

Boghossian, P. A. (1989) 'The Rule-Following Considerations' in *Mind*, 98 (392): 507-549.

Bradley, R. and List, C. (2009) 'Desire-as-Belief Revisited' in *Analysis*, 69 (1):681-697.

Brandt, R. (1972) 'Rationality, Egoism, and Morality' in *The Journal of Philosophy*, 69 (20): 681-697.

Bratman, M. E. (1999) *Intention, Plans and Practical Reason*, Cambridge, MA: Harvard University Press.

Brink, D. O. (2001) 'Realism, Naturalism, and Moral Semantics' in *Social Philosophy and Policy*, 18 (2): 154-176.

Broome, J. (1991) 'Desire, Belief and Expectation' in *Mind*, 100 (2): 265-267.

Broome, J. (1999) 'Normative Requirements' in *Ratio*, 12 (4): 398-419.

Broome, J. (2002) 'Practical Reasoning' in Bermudez, J. and Millar, A. (eds.) *Reason and Nature: Essays in the Theory of Rationality*, Oxford: Oxford University Press: 85–111.

Broome, J. (2007) 'Wide or Narrow Scope?' in *Mind*, 116 (462): 359–70.

Burton, S. (1992) ' "Thick" Concepts Revised' in *Analysis*, 52 (1): 28-32.

Charlton, W. (1988) *Weakness of Will*, Oxford: Basil Blackwell.

Child, W. (1994) 'On the Dualism of Scheme and Content' in *Proceedings of the Aristotelian Society*, 94 (1): 53-71.

Dancy, J. (2000) *Practical Reality*, Oxford: Oxford University Press.

Daskal, S. (2010) 'Absolute Value as Belief' in *Philosophical Studies*, 148 (2): 221-229.

Davidson, D. (1963) 'Actions, Reasons, and Causes' in Davidson, D. (2001) *Essays on Actions and Events*, Oxford: Oxford University Press: 3-19.

Davidson, D. (1967) 'Truth and Meaning' in Davidson, D. (2001) *Inquiries into Truth and Interpretation*, Oxford: Oxford University Press: 17-36.

Davidson, D. (1969) 'How is Weakness of the Will Possible?' in Davidson, D. (2001) *Essays on Actions and Events*, Oxford: Oxford University Press: 21-42.

Davidson, D. (1970) 'Mental Events' in Davidson, D. (2001) *Essays on Actions and Events*, Oxford: Oxford University Press: 207-227.

Davidson, D. (1973) 'Radical Interpretation' in Davidson, D. (2001) *Inquiries into Truth and Interpretation*, Oxford: Oxford University Press: 125-139.

Davidson, D. (1974a) 'Belief and the Basis of Meaning' in Davidson, D. (2001) *Inquiries into Truth and Interpretation*, Oxford: Oxford University Press: 141-154.

Davidson, D. (1974b) 'On the Very Idea of a Conceptual Scheme' in Davidson, D. (2001) *Inquiries into Truth and Interpretation*, Oxford: Oxford University Press: 183-198.

Davidson, D. (1975) 'Thought and Talk' in Davidson, D. (2001) *Inquiries into Truth and Interpretation*, Oxford: Oxford University Press: 155-170.

Davidson, D. (1978) 'Intending' in Davidson, D. (2001) *Essays on Actions and Events*, Oxford: Oxford University Press: 83-102.

Davidson, D. (1980) 'A Unified Theory of Thought, Meaning, and Action' in Davidson, D. (2004) *Problems of Rationality*, Oxford: Oxford University Press: 17-36.

Davidson, D. (1982) 'Paradoxes of Irrationality' in Davidson, D. (2004) *Problems of Rationality*, Oxford: Oxford University Press: 169-187.

Davidson, D. (1986) 'Deception and Division' in Davidson, D. (2004) *Problems of Rationality*, Oxford: Oxford University Press: 199-212.

Davidson, D. (1995) 'Could There Be a Science of Rationality?' in Davidson, D. (2004) *Problems of Rationality*, Oxford: Oxford University Press: 117-134.

Davidson, D. (2001) 'Comments on Karlovy Vary Papers' in Kotatko, P., Pagin, P. and Segal, G. (eds.) *Interpreting Davidson*, Stanford: CSLI Publications: 285-307.

Dennett, D. C. (1987) *The Intentional Stance*, Cambridge, MA: MIT Press.

Dreier, J. (1997) 'Humean Doubts about the Practical Justification of Morality' in Cullity, G. and Gaut, B. (eds.) *Ethics and Practical Reason*, Oxford: Clarendon Press: 81-99.

Elster, J. (1999) 'Davidson on Weakness of the Will and Self-Deception' in Hahn, L. E. (ed.) *The Philosophy of Donald Davidson*, Chicago and La Salle, IL: Open Court Publishing Company: 425-442.

Engel, P. (2001) 'Is Truth a Norm?' in Kotatko, P., Pagin, P. and Segal, G. (eds.) *Interpreting Davidson*, Stanford: CSLI Publications: 37-51.

Enoch, D. (2006) 'Agency, Schmagency: Why Normativity Won't Come From What is Constitutive of Action' in *Philosophical Review*, 115 (2): 169-198.

Finlay, S. (2010) 'Recent Work on Normativity' in *Analysis*, 70 (2): 331-346.

Føllesdal, D. (1985) 'Causation and Explanation: A Problem in Davidson's View on Action and Mind' in LePore, E. and McLaughlin, B. P. (eds.) *Actions and Events: Perspectives on the Philosophy of Donald Davidson*, Oxford: Blackwell: 311-323.

Foot, P. (1972) 'Morality as a System of Hypothetical Imperatives' in *Philosophical Studies*, 81 (3): 305-316.

Foot, P. (2001) *Natural Goodness*, Oxford: Clarendon Press.

Gauthier, D. (1986) *Morals by Agreement*, Oxford: Clarendon Press.

Gauthier, D. (2003) 'Why Contractariansim?' in Darwall, S. L. (ed.) Contractarianism/Contractualism, Oxford: Blackwell.

Glüer, K. (2001) 'Dreams and Nightmares: Conventions, Norms and Meaning in Davidson's Philosophy of Language' in Kotatko, P., Pagin, P. and Segal, G. (eds.) *Interpreting Davidson*, Stanford: CSLI Publications: 53-74.

Glüer, K. and Wikforss, Å. M. (2009) 'The Normativity of Meaning and Content' in Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy (Winter 2010 Edition)*, URL: <http://plato.stanford.edu/archives/win2010/entries/meaning-normativity/>.

Goldman, A. (2010) *Reasons from Within: Desires and Values*, Oxford: Oxford University Press.

Greene, J. D. (2007) 'Why are VMPFC Patients More Utilitarian?: A Dual-Process Theory of Moral Judgment Explains' in *Trends in Cognitive Sciences*, 11 (8): 322-3.

Griffin, J. (1986) *Well-Being: Its Meaning, Measurement, and Moral Importance*, Oxford: Clarendon Press.

Hacker, P. M. S. (1996) 'On Davidson's Idea of a Conceptual Scheme' in *The Philsophical Quarterly*, 46 (184): 289-307.

Hampton, J. (1996) *On Instrumental Rationality*, in Schneewind, J.B. (ed.) *Reason, Ethics and Society: Themes from Kurt Baier, with His Responses*, Chicago and La Salle, IL: Open Court.

Hampton, J. (1998) *The Authority of Reason*, Cambridge: Cambridge University Press.

Hájek, A. and Pettit, P. (2004) 'Desire Beyond Belief' in *Australasian Journal of Philosophy*, 82 (1): 77-92.

Heil, J. (1989) 'Minds Divided' in *Mind*, 98 (392): 571-583.

Heuer, K. (unpublished) 'Hypotheticalism and the Objectivity of Morality', URL: <http://www7.georgetown.edu/students/kwh6/home/Papers_files/Hypotheticalism%20and%20the%20Objectivity%20of%20Morality.pdf>

Hobbes, T. (1996) *Leviathan*, Oxford: Oxford University Press.

Hoffman, M. L. (2000) *Empathy and Moral Development*, Cambridge: Cambridge University Press.

Hollis, M. and Sugden, R. (1993) 'Rationality in Action' in *Mind*, 102 (405): 1-35.

Hubin, D.C. (1999) 'What's Special About Humeanism' in *Noûs*, 33 (1): 30-45.

Hume, D. (1969) *A Treatise of Human Nature*, London: Penguin Books.

Hurley, S. L. (1989) *Natural Reasons: Personality and Polity*, Oxford: Oxford University Press.

Hursthouse, R. (1999) *On Virtue Ethics*, Oxford: Clarendon Press.

Hussein, N. J. Z. and Shah, N. (2006) 'Misunderstanding Metaethics: Korsgaard's Rejection of Realism' in Shafer-Landau, R. (ed.) *Oxford Studies in Metaethics*, Vol 1., Oxford: Clarendon Press:: 265-294.

Jackson, F. and Pargetter, R. (1986) 'Oughts, Options, and Actualism' in *The Philosophical Review*, 95 (2): 233-255.

Jeffrey, R. C. (1983) *The Logic of Decision*, Chicago: University of Chicago Press.

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

Kalderon, M. E. (2005) *Moral Fictionalism*, Oxford: Oxford University Press.

Kavka, G. S. (1983) 'The Toxin Puzzle' in *Analysis*, 43 (1): 33-36.

Klein, P. D. (1986) 'Radical Interpretation and Global Skepticism' in LePore, E. (ed.) *Truth and Interpretation: Perspectives on the Philosophy of Donal Davidson*, Oxford: Blackwell: 369-386.

Koenigs, M. *et al.* (2007) 'Damage to the Prefrontal Cortex Increases Utilitarian Moral Judgements' in *Nature*, 446: 908-911.

Kolodny, N. (2005) 'Why Be Rational' in *Mind*, 114 (455): 509-563.

Korsgaard, C. M. (1996) *The Sources of Normativity*, Cambridge: Cambridge University Press.

Korsgaard, C. M. (1997) 'The Normativity of Instrumental Reason' in Cullity, G. and Gaut, B. (eds.) *Ethics and Practical Reason*, Oxford: Clarendon Press: 215-254.

Korsgaard, C. M. (2008) *The Constitution of Agency: Essays on Practical Reason and Moral Psychology*, Oxford: Oxford University Press.

Korsgaard, C.M. (2009) *Self Constitution*, Oxford: Oxford University Press.

Kripke, S. A. (1980) *Naming and Necessity*, Oxford: Blackwell.

Kripke, S. A. (1982) *Wittgenstein on Rules and Private Language*, Oxford: Blackwell.

Lazar, A. (1999) 'Akrasia and the Principle of Continence or What the Tortoise Would Say to Achilles' in Hahn, L. E. (ed.) *The Philosophy of Donald Davidson*, Chicago and La Salle, IL: Open Court Publishing Company: 381-401.

Lenman, J. (2008) 'Moral Naturalism' in Zalta, E. N. (ed.) *The Stanford Encyclopedia of Philosophy (Winter 2008 Edition)*, URL: <http://plato.stanford.edu/archives/win2008/entries/naturalism-moral/>.

LePore, E. and Ludwig, K. (2005) *Donald Davidson: Meaning, Truth, Language, and Reality*, Oxford: Clarendon Press.

LePore, E. and Ludwig, K. (2007) *Donald Davidson's Truth-Theoretic Semantics*, Oxford : Clarendon Press.

Levi, I. (1999) 'Representing Preferences: Donald Davidson on Rational Choice' in Hahn, L. E. (ed.) *The Philosophy of Donald Davidson*, Chicago and La Salle, IL: Open Court Publishing Company: 531-570.

Lewis, D. (1984) 'Putnam's Paradox' in *Australasian Journal of Philosophy*, 62 (3): 221-236.

Lewis, D. (1988) 'Desire as Belief' in *Mind*, 97 (418): 323-32.

Lewis, D. (1996) 'Desire as Belief II' in *Mind*, 105 (418): 303-13.

Lillehammer, H. (1997) 'Smith on Moral Fetishism' in *Analysis*, 57 (3): 187-195.

Ludwig, K. (ed.) (2003) *Donald Davidson*, New York: Cambridge University Press.

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

McClennen. E. F. (1990) *Rationality and Dynamic Choice: Foundational Explorations*, Cambridge: Cambridge University Press.

McDowell, J. (2001) 'Scheme-Content Dualism and Empiricism' in Kotatko, P., Pagin, P. and Segal, G. (eds.) *Interpreting Davidson*, Stanford: CSLI Publications: 143-154.

McGinn, C. (1986) 'Radical Interpretation and Epistemology' in LePore, E. (ed.) *Truth and Interpretation: Perspectives on the Philosophy of Donal Davidson*, Oxford: Blackwell: 356-368.

Mele, A. (1987) *Irrationality*, Oxford: Oxford University Press.

Millgram, E. (1995) 'Was Hume a Humean?' in *Hume Studies*, 21 (1): 75-94.

Millgram, E. (2009) 'Practical Reason and the Structure of Actions' in Zalta, E. N. (ed.) *The Stanford Encyclopedia of Philosophy (Summer 2009 Edition)*, URL: <http://plato.stanford.edu/archives/sum2009/entries/practical-reason-action/>.

Moll, J. and de Oliveira-Souza, R. (2007) 'Response to Greene: Moral Sentiments and Reason: Friends or Foes?' in *Trends in Cognitive Sciences*, 11 (8): 323-4.

Moore, G.E. (1903) *Principia Ethica*, Cambridge: Cambridge University Press.

Nolan, D. Restall, G. and West, C. (2005) 'Moral Fictionalism versus the rest' in *Australasian Journal of Philosophy*, 83 (3): 307-330.

Nozick, R. (1993) *The Nature of Rationality*, Princeton, NJ: Princeton University Press.

Nussbaum, M. (1995) 'Aristotle on Human Nature and the Foundations of Ethics' in Altham, J. E. J. and Harrison, R. (eds.) *World, Mind, and Ethics: Essays on the Ethical Philosophy of Bernard Williams*, Cambridge: Cambridge University Press: 86–131.

Parfit, D. (1987) *Reasons and Persons*, (reprinted with further corrections), Oxford: Clarendon Press.

Parfit, D. (1997) 'Reasons and Motivation' in *Aristotelian Society Supplementary Volume*, 71: 99-130.

Parfit, D. (2006) 'Normativity' in Shafer-Landau, R. (ed.) *Oxford Studies in Metaethics*, Vol 1., Oxford: Clarendon Press: 325-380.

Parfit, D. (forthcoming) *On What Matters*, Oxford: Oxford University Press. URL: <http://fas-philosophy.rutgers.edu/chang/Papers/OnWhatMatters1.pdf>.

Piaget, J. (1932) *The Moral Judgement of the Child*, New York: Harcourt.

Pink, T. L. M. (1991) 'Purposive Intending' in *Mind*, 100 (399): 343-359.

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

Quine, W. V. O. (1960) *Word and Object*, Cambridge, MA: MIT Press.

Railton, P. A. (1986) 'Moral Realism' in Railton, P. A. (2003) *Facts and Values: Essays Toward a Morality of Consequence*, Cambridge: Cambridge University Press: 3-42.

Railton, P. (2006) 'The Humean Theory of Practical Rationality' in Copp, D. (ed.) *The Oxford Handbook of Ethical Theory*, Oxford: Oxford University Press: 265-281.

Ramsey, F. P. (1926) 'Truth and Probability' in Ramsey, F. P. (1931) *The Foundations of Mathematics*, London: Routledge and Kegan Paul: 156-198.

Rawls, J. (1971) *A Theory of Justice*, Cambridge, MA: Belknap Press.

Rosati, C. (1995) 'Naturalism, Normativity and the Open Question Argument' in *Noûs*, 29 (1): 46-70.

Scanlon, T. M. (1998) *What We Owe to Each Other*, Cambridge, MA: Harvard University Press.

Scheffler, S. (1985) 'Agent-Centred Restrictions, Rationality, and the Virtues' in *Mind*, 94 (375): 409-419.

Schroeder, M. (2004) 'The Scope of Instrumental Reason' in Hawthorne, J. and Zimmerman, D. W. (eds.) *Ethics: Philosophical Perspectives*, 18, 337-364.

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

Schroeder, M. (unpublished) 'The Negative Reason Existential Fallacy', URL: <http://www-bcf.usc.edu/~maschroe/research/Schroeder_Negative_Reason_Existential_Fallacy.pdf accessed 2/1/2011>.

Schroeder, T. (2003) 'Donald Davidson's Theory of Mind Is Non-Normative' in *Philosophers' Imprint*, 3 (1): 1-14.

Shafer-Landau, R. (2011) 'Three problems for Schroeder's hypotheticalism' in *Philosophical Studies* (online only at time of citation), URL: < http://dx.doi.org/10.1007/s11098-010-9655-4>.

Sinclair, N. (2007) 'Propositional Clothing and Belief' in *The Philosophical Quarterly*, 57 (228): 342-362.

Smith, M. (1994) *The Moral Problem*, Oxford: Blackwell.

Smith, M. (2004) 'Instrumental Desires, Instrumental Rationality' in *Aristotelian Society Supplementary Volume*, 78 (1): 93-109.

Soles, D. H. (1999) 'Prefers True: Archimedean Point or Achilles' Heel?' in Hahn, L. E. (ed.) *The Philosophy of Donald Davidson*, Chicago and La Salle, IL: Open Court Publishing Company: 311-329.

Street, S. (2006) 'A Darwinian Dilemma for Realist Theories of Value' in *Philosophical Studies*, 127 (1): 109-166.

Street, S. (2008) 'Constructivism about Reasons' in Shafer-Landau, R. (ed.) *Oxford Studies in Metaethics*, Vol. 3, Oxford: Oxford University Press: 207-245.

Street, S. (2009) 'In Defense of Future Tuesday Indifference: Ideally Coherent Eccentrics and the Contingency of What Matters' in *Philosophical Issues*, 19 (1): 273-298.

Streumer, B. (forthcoming) 'Can We Believe the Error Theory?', forthcoming (with revisions) in *Journal of Philosophy*, URL: <http://www.personal.reading.ac.uk/~lds05bs/BelieveErrorTheory.pdf>

Sturgeon, N. (1988) 'Moral Explanations' in Sayre McCord, G. (ed.) *Essays on Moral Realism*, Ithaca: Cornell University Press.

Taurek, J. M. (1977) 'Should the Numbers Count?' in *Philosophy and Public Affairs*, 6 (4): 293-316.

Thomson, J. J. (1997), 'The Right and the Good' in *Journal of Philosophy*, 94 (6): 273–298.

Trivers, R.L. (1971) 'The Evolution of Reciprocal Altruism' in *Quarterly Review of Biology*, 46 (1): 35-57.

Velleman, J. D. (2000) *The Possibility of Practical Reason*, Oxford: Oxford University Press.

Watson, G. (1977) 'Skepticism About Weakness of Will' in *Philosophical Review*, 86 (3): 316-339.

Wedgwood, R. (2003) 'Choosing Rationally and Choosing Correctly' in Stroud, S. and Tappolet, C. (eds.) *Weakness of Will and Practical Irrationality*, Oxford: Oxford University Press: 201-250.

Wedgwood, R. (2007) *The Nature of Normativity*, Oxford: Clarendon Press.

Weintraub, R. (2007) 'Desire as Belief, Lewis Notwithstanding' in *Analysis*, 67 (294): 116-122.

Wikforss, A. M. (2001) 'Semantic Normativity' in *Philosophical Studies*, 102 (1): 203-226.

Williams, B.A.O. (1981) 'Internal and External Reasons' in his *Moral Luck*, Cambridge: Cambridge University Press: 101-113.

Williams, J. R. G. (2007) 'Eligibility and Inscrutability' in *Philosophical Review*, 116 (3): 361-339.

Wittgenstein, L. (2001) *Philosophical Investigations*, Oxford: Blackwell.

Woodard, C. (2009) 'What's Wrong with Possibilism?' in *Analysis*, 69 (2): 219-226.

Zimmerman, M. J. (1996) *The Concept of Moral Obligation*, Cambridge: Cambridge University Press.

Zheng, Y. (2001) 'Akrasia, Picoeconomics, and a Rational Reconstruction of Judgment Formation in Dynamic Choice' in *Philosophical Studies*, 104 (3): 227-251.