TWO SOURCES OF MORAL REASONS

Iain Ezra Macdonald

A Thesis Submitted for the Degree of PhD
at the
University of St. Andrews

2010

Two Sources of Moral Reasons

Iain Ezra Macdonald

A thesis submitted for the degree of Doctor of Philosophy

24th August 2009

Abstract

One of the core questions in contemporary metaethics concerns the nature and status of moral claims. However, this question presupposes that morality is unified, and that a single metaethical account will suffice. This thesis aims to challenge that presupposition. In particular, I argue that there is a substantial theoretical payoff to be had from combining two distinct metaethical theories – realism, on the one hand, and constructivism, on the other – whilst limiting the scope of each. In the realist case, the discourse aims to describe a particular feature of reality; in the constructivist case, the discourse aims to solve some of the coordination problems faced by people as social beings. We have, therefore, two distinct sources of moral reasons.

The resulting 'hybrid' theory is appealing at the metaethical level, but also yields an attractive picture at the applied level. Specifically, it retains the core intuition underlying utilitarianism, whilst incorporating a broadly contractarian account of morality. On this account, our reasons for not harming other persons are at least the same as our reasons for not harming animals – but we have additional reasons to refrain from harming persons.

Chapter One establishes a moderate presumption in favour of moral realism, understood as the claim that moral discourse aims to represent the world, deals in objective truths, and yields statements capable of truth or falsity. Chapter Two addresses arguments for moral antirealism: these arguments can be met by restricting the scope of moral realism. Chapter Three explores the content of the resultant moral realism: specifically, realism about the intrinsic value of hedonic states. Chapter Four deals with that part of morality which is unaccounted for by restricted moral realism, and offers an outline form of contractarian constructivism. Chapter Five investigates the consequences of the hybrid metaethical theory for applied ethics.

I, Iain Ezra Macdonald, hereby certify that this thesis, which is approximately 78,000 words in length, has been written by me, that it is the record of work carried out by me and that it has not been submitted in any previous application for a higher degree.

I was admitted as a candidate for the degree of Doctor of Philosophy in September 2005, the higher study for which this is a record was carried out in the University of St Andrews between 2005 and 2009.

Date: 12th April 2010. Signature of candidate ……….........................

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of Doctor of Philosophy in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

Date: 12th April 2010. Signature of supervisor ………........................

In submitting this thesis to the University of St Andrews we understand that we are giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. We also understand that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that my thesis will be electronically accessible for personal or research use unless exempt by award of an embargo as requested below, and that the library has the right to migrate my thesis into new electronic forms as required to ensure continued access to the thesis. We have obtained any third-party copyright permissions that may be required in order to allow such access and migration.

The following is an agreed request by candidate and supervisor regarding the electronic publication of this thesis:

Access to printed copy and electronic publication of thesis through the University of St Andrews.

Date: 12thApril 2010.

Signature of candidate ....……...........

Signature of supervisor ……….....................

For Mum & Dad

Acknowledgements

I would like to thank all of those who have been involved – however indirectly – with the process which led to this thesis. Most directly, my many supervisors: John Skorupski, Daniel Nolan, Tim Mulgan, Neil Sinclair, and Antti Kauppinen; but especially Berys Gaut, who held the whole project together. I could not have hoped for a more patient or helpful helmsman.

Less directly, the staff and graduate community at St Andrews, who provided a bright and engaging environment for study, and who tolerated my unsteady philosophy, within and without the Departments; and, of course, at the Barbers. The fellow graduates with whom I have shared interesting times, and those who made those times interesting. Those in the past who have made the enterprise of academic philosophy seem good and worthwhile, particularly my undergraduate Director of Studies, Hallvard Lillehammer. And all of those with whom I have shared good times in St Andrews, who will remain nameless – you know who you are. With one exception, of course: Clairelouise Anderson, for her continual support and almost-limitless patience during the final phase, for putting up with me, and for innumerable good qualities. I couldn’t have done it without you.

The first three years of the project were supported by a Doctoral Research Grant from the Arts and Humanities Research Council, gratefully acknowledged.

Most of all, though, I’d like to thank my parents, who have supported me and encouraged me for a good quarter of a century, provided me with all manner of opportunities, and put up with my many foibles along the way. I am very lucky to have two such lovely people in my life. They are super.

Contents

CHAPTER ONE

Moral Realism I

1. Introduction

Taken at face value, moral practice involves judgements capable of truth or falsity, and possessed of motive force; these judgements concern some feature of the world, and are thereby made true or false. The moral status of an action doesn't depend on the whims of the agent, nor are the demands of morality merely optional. We have reason to care about morality. Moral knowledge is possible, and is, at least in some cases, the object of moral inquiry. Moral beliefs are, at least sometimes, supported by reasons, and a fit subject for rational argument and debate. And there is, at least sometimes, a fact of the matter for us to discover. Consequently, sometimes our most cherished moral beliefs turn out to be wrong.

At least, that is how it seems at first glance. The truth about morality may turn out to be otherwise; certainly, a good number of philosophers hold that these appearances are deceptive. Worse, there are compelling reasons to think that the doubters are correct. But the burden of proof lies with them; that things seem to be a certain way is reason to think that they *are* a certain way, and to doubt the appearances requires justification.[1] The central claim of my thesis is straightforward: in some cases these reasons are compelling, and the truth about morality runs contrary to the appearances; in others, they are not, and things are just as they seem to be. Furthermore, a little investigation yields reasons beyond mere appearances to think that some of the appearances are veridical. The metaethical story is, in this sense, disunified, and for this reason I refer to my position as a 'hybrid' theory.

This chapter deals with the metaethical account which I label *realist*: it begins with an attempt to define some characteristics of the realist position, and examines a set of objections. In some cases, these objections are compelling; the realist story simply will not work for those cases. But in other cases the objections can be met; for these cases, absent good reason to think that realism is unworkable, and given good reason to think that realism in these cases is true, I take it that the (realist) appearances are veridical. This forms the distinction central to my

[1] C.f. Aristotle 1925: 160, 'We must, as in all other cases, set the apparent facts before us and, after first discussing the difficulties, go on to prove, if possible, the truth of all the common opinions about these affections of the mind, or, failing this, of the greater number and the most authoritative; for if we both resolve the difficulties and leave the common opinions undisturbed, we shall have proved the case sufficiently.'

thesis – between that part of ethics for which a realist account is workable, and that part of ethics for which it is not.

The account that will be developed offers a novel answer to the vexed question of whether the ends of action justify the means, or *vice versa*. For some philosophers – notably the older Christian philosophers – morality consisted primarily of a set of commandments. There were some actions which simply should not be done, regardless of whatever good might ensue. Kant, to ignore entirely the subtleties of his account, also held this line. For others – Bentham, Mill, and their philosophical descendents – the good was prior to the right: morality was a matter of identifying the bearer of value (namely, utility), and then enjoining people to maximise this value. W. D. Ross, writing in 1930, proposed a plurality of *prima facie* duties *and* goods, appearing to stand in some more-or-less systematic relation to each other, but independently knowable and not interreducible.

This account locates itself alongside Ross, holding that we should keep distinct our theory of value (axiology), and of what we ought to do (deontology). I will argue that we should adopt a limited axiological realism – that is, we should be realists about intrinsic value – but be constructivists about other moral reasons. Such a position has both metaphysical and first-order ethical advantages. It is metaethically preferable because it offers the best explanation of the ethical phenomena. It is ethically advantageous because it provides a plausible (and well-grounded) account of what is intrinsically valuable, capturing the core insights underlying utilitarianism, whilst including a further source of moral reasons which accommodates otherwise-problematic ethical phenomena. I address the substance of realist morality in Chapter Three; the substance of constructivist morality in Chapter Four; and provide an examination of the applications of the theory in Chapter Five.

2. Methodology

The strategy of this chapter is to cite certain features of moral discourse and practice, and then to claim that these are best explained by moral realism. So, for instance, one appearance of moral discourse is that moral claims may sometimes be wrong – and this is explained by the suggestion that moral claims have mind-independent truthmakers. Of course, this by itself does not establish realism (the anti-realist might claim, for instance, that claims are true if and only if the speaker would want them to be true after full deliberation, thereby making space for the possibility of error). But the overall picture is, I argue, best suited to a realist explanation.

Nonetheless, this might seem unsatisfactory. Why should we feel constrained to take these appearances seriously in the first place? Absent Catholic inclinations, we do not take the pronouncements of the Vatican to give *prima facie* reason to believe in the existence of theological (or other supernatural) properties. But morality differs in its claims: religion makes claims which are, if not exactly testable, at least amenable to some empirically based investigation. We take the apparent causal closure of the physical universe to provide evidence against an interventionist God; considerations of personal identity cast doubt on the possibility of an afterlife; and the proposed combination of benevolence, omnipotence, and omniscience is held to be implausible by a sizeable number of people. And a key feature of *prima facie* reasons is, of course, that once overridden they cease to exist; for those who take the reasons to count overwhelmingly against the existence of an interventionist God, at least *some* of the reasons in favour cease to be reasons at all. Prior to the investigation, however, the various facts about the discourse and practice *did* give reason to think that there was something to be investigated. That we no longer feel required to take the appearances seriously does not bear on the *ex ante* relevance of those appearances.

Nonetheless, there is much of religious discourse which we do take at face value. Most people these days, whatever their views on the existence or otherwise of God (or of gods, or of some other supernatural entity) take religious talk to be truth-apt, purporting to describe some fact of the matter, and so on. For most people – professional philosophers included – the metatheological debate is between realism and error theory, with perhaps a few fictionalists at the margins. The error-theoretic positions take the surface grammar to be indicative of how the discourse ought to be understood (as truth-apt, and purporting to describe some realm of mind-independent fact), whilst denying the existence of the posited supernatural properties or entities. So the question is not simply whether we should save the appearances, but *which* appearances (if any) we should attempt to save.

To further complicate matters, sound metaethical theorising is also constrained by factors independent of moral discourse and practice. Mark Timmons calls this a project of 'external accommodation' (in contrast to the project of 'internal accommodation', which attempts to make sense of the commitments of the subject matter on its own terms).[2] One of the most appealing constraints is metaphysical naturalism, according to which the only things which exist are natural entities – meaning the kinds of things which are talked about by the natural sciences, usually taken to include physics, chemistry, and perhaps biology. Someone who subscribes to this constraint will be moved to give an error-theoretic or debunking account of

[2] Timmons 1999, esp. Ch 1.

any appearances which commit to the existence of non-natural entities or properties. Whether or not metaphysical naturalism is true, what does need to be borne in mind is that the sorts of appearances which we are inclined to save will depend on our extra-moral inclinations. Put differently, the project of internal accommodation is informed and regulated by the project of external accommodation.

The strategy, therefore, will be to first get clear on how morality seems to be, and then attempt to save only those appearances which are unobjectionable from the external viewpoint. These appearances are, variously, syntactic, metaphysical, and epistemological. I will adopt, for the purposes of this project, a thoroughgoing naturalism, according to which we ought not to admit the existence of non-natural entities or properties without very good reason. The particular kinds of non-natural properties posited by the theologian are particularly objectionable because they are taken to interact with, but be somehow apart from, the physical world. Divine intervention is an obvious instance where such objectionable properties would be required, and I will assume that any commitment to the existence of such properties, or to related epistemic faculties (such as a *sui generis* faculty of moral intuition), can happily be rejected. This is not to say that we cannot allow the existence of properties which are naturalistically respectable, but nevertheless irreducible to the terms of the natural sciences; nor, indeed, is it to say that we cannot allow for non-natural properties, provided that they can coexist with our best scientific understanding of the world (the properties involved in divine intervention are problematic largely because they are held to be non-natural, but still interact with the natural world, and hence require a violation of the laws of physics).[3] But it is not clear that non-natural moral properties *could* coexist with our best scientific understanding of the world: if they are to be knowable, then they must be capable of causing some change in our beliefs, and hence physical changes. This seems to be just as problematic as ‘divine intervention’. The problem extends to any account which takes moral properties to be natural, but causally impotent, properties: if they are capable of causing a change in our mental states, then they must be causally potent.[4]

There are at least two ways that the non-naturalist might attempt to tackle this problem. One is to posit a *sui generis* faculty of moral intuition. This move is unpalatable, for various reasons (which I will discuss in some detail in the next chapter). The second is to maintain that our epistemic access to moral properties is not a causal matter. For instance, it does seem

[3] O’Connor 1994 provides a helpful discussion of naturalistically respectable properties which are irreducible to terms of the natural sciences.

[4] A parallel argument applies to the issue of qualia, occurring in Fox 1989 and discussed in Dennett 1991: 400-404.

that we can come to have knowledge about numbers, arithmetic truths, properties of sets, and so on, but it is not obvious that numbers (or arithmetic truths, or properties of sets) are themselves causally potent. Trivially, we might think that only concrete objects are causally potent, and that numbers, as abstract objects, lack this potency. On the other hand, numbers do feature in naturalistic explanations. But it also seems that numerical *properties* can be causally potent: the key turned in the lock because it depressed a certain number of tumblers, for instance – and this is a perfectly naturalistic explanation. I will therefore pursue a thoroughgoing naturalism, and leave discussion of non-natural properties to one side.

As regards the surface appearances of the moral discourse, I take it that the syntactic appearances ought to be taken very seriously indeed. This is partly a function of the fact that speakers' intentions must have a large part to play in what they accomplish by their utterances. It would be surprising if speakers were guilty of a systematic failure to make assertions, exclamations, or somesuch, in spite of their apparent competency with the relevant language. Furthermore, since the surface grammar of a language determines, to a large extent, what we can do with that language, and is in that sense partially constitutive of the language, it is hard to see how it could be entirely misleading (after all, what is 'really going on' in the language just is a matter of its constitutive grammar and semantics). If a discourse appears to be truth-apt, then the burden of proof falls on those who deny that this is the case.

At a first attempt, then, we ought to place a high degree of credence in the syntactic appearances of a discourse; a moderate degree of credence in the epistemological appearances (insofar as they do not conflict directly with our best scientific understanding of the world); and a relatively low degree of credence in the metaphysical appearances. Our obligations regarding the appearances follow this pattern on account of the varying levels of effort involved in explaining the corresponding errors. It is one thing to provide a debunking account of certain metaphysical intuitions; quite another to do the same for surface grammar.

But not everyone thinks that the appearances ought to be saved. Expressivists, who hold that the point of moral discourse and practice is to co-ordinate actions (in a broad sense), have tended to claim that only *some* of the appearances need to be saved, viz., those that matter – those that are essential to the point of the practice.[5] If some of the appearances are merely incidental, then the theoretical cost of failing to save these appearances is minimal: we perhaps need to provide a genealogical or historical explanation of why these appearances have come about, but our final theory doesn't need to preserve them. I think that this is

[5] See e.g. Gibbard 1990, 2003; Hare 1991, 1997 (Hare's account is, strictly speaking, prescriptivist, rather than expressivist).

entirely correct: were it true that morality, for instance, was essentially a device for co-ordinating attitudes, then it is this function which we would need to account for. But as should become clear in due course, I hold that not only does morality seem to be descriptive in function, we can also make good sense of this appearance. We could, if we wished, adopt a revisionary theory – according to which the appearances are deceptive, or to be best made sense of via some circuitous route - but this is, I will argue, unmotivated.

Neil Sinclair has suggested that the difference between metatheological and metaethical theorising - at least with respect to our motivation to save the appearance - depends on what he calls 'the pragmatic assumption'; the assumption that there is something peculiarly useful about morality, and hence that we have good reason to carry on engaging in moral practice.[6] We therefore have reason to look for a metaethical theory which vindicates the core elements of our moral practice. But importantly for Sinclair and his fellow expressivists, we can retain the core elements of moral practice - those that render it useful - without committing ourselves to moral realism. What we are looking for is a vindicatory account of the 'pragmatically important forms of moral practice and the pragmatically important assumptions of those who engage in it.'[7] Something very similar seems to underwrite other expressivisms - Blackburn's quasi-realism, Gibbard's 'norm-expressivism', among others - as well as their close irrealist cousins, such as Kalderon's moral fictionalism.[8] Our moral practice matters to us, and we consequently have a vested interest in avoiding debunking it.

I think that Sinclair has accurately identified a thought which guides much contemporary anti-realist metaethics. I also suspect this thought of being misplaced. The thought is misplaced, simply because the pragmatic issue - whether we have an interest in retaining our current moral practices - has no bearing on the truth of the metaethical theory. The mere fact that the existence of moral properties would be somehow conducive to our wellbeing gives us no reason to believe that moral realism is true, any more than the putative benefits of the existence of God make it more likely that God exists.[9]

Even if the pragmatic issue *did* relate to the theoretical issue, it is not clear that that metaethical theorising has a significant practical upshot. Were it the case that morality as

[6] Sinclair 2009.

[7] Sinclair 2009: 7.

[8] See Blackburn 1984, 1993, 1998, 2002; Gibbard, op. cit.; Kalderon 2005.

[9] Although there may be cases where such a connection obtains in virtue of the content of moral properties – as there would be were it analytic that 'goodness' and 'conduciveness to wellbeing' were identical. But it is nonetheless true that the potential extrinsic benefits of putative moral properties do not give us reason to believe in their existence. There may, of course, be pragmatic reasons to form such beliefs, but these are distinct from the theoretical reasons under discussion.

practiced by the folk is dependent on metaethics as practiced by Anglophone philosophers, then there would perhaps be some reason to be guided by the 'pragmatic assumption'. But this is, at the very least, tendentious. Absent some (admittedly significant) exceptions, such as the Lockean reference to 'inalienable rights' enshrined in the American constitution, academic metaethics does not seem to have any great impact on the wider world.

Furthermore, there are at least two ways in which morality may have a 'practical upshot' - two kinds of reasons why we might want to preserve the discourse and practice. The first is impartial: we simply think that it is a good thing that there exist moral norms which regulate behaviour. The second is self-interested: it is in my own interest that there exist moral norms which regulate behaviour.[10] Each of these is problematic. If our reasons for wanting to preserve morality are moral reasons, then there is a worry about circularity. Either this circularity is vicious, or it is not. If it is, then we have a problem; if it is not, then it is hard to see how we can avoid admitting (for instance) religious reasons for wanting to preserve religious discourse. If, on the other hand, it is self-interest that provides reason for wanting to vindicate (most of) morality, then the 'pragmatic assumption' starts to look implausible. After all, whilst it may be in my interest that other people behave morally, it does not follow from this that it is in my interests to do so. Quite plausibly, my interests are best served by being a lone, sensible knave in a sea of hapless moralists.

There are, then, three distinct questions: what we ought to believe; what beliefs we have reason to promote in the general populace; and whether or not we can make any difference to what the general populace thinks or does. I have already suggested that the impact of academic metaethics on folk morality is insignificant - but even if I am wrong about this, the determining factor is not what academic philosophers believe, but what the dominant (folk) metaethical theory is. Our metaethical theory could, therefore, be radically revisionary (we could hold, for instance, that morality is composed of a set of necessarily false claims, used by the weak in order to trick the strong into giving up resources), without carrying any pragmatic cost.

The attempt to guide our inquiry by *practical* concerns is, therefore, mistaken.[11] Rather, the original suggestion was correct: begin by citing appearances, and then proceed to see whether or not we can make sense of these appearances - and, if so, how. Ascriptions of error at the

[10] 'Norms' here is meant in a purely descriptive sense.

[11] At least, in the narrow sense which the expressivists want to employ. I do not deny that theoretical reason may be (quite properly, perhaps necessarily) guided by practical reason.

level of the object theory are to be assessed in terms of the associated explanatory cost, and not in terms of the practical consequences of holding such a view.[12]

3. Preliminaries

I have already indicated that I wish to talk of my position as partly realist. The question of what is meant by this is, of course, a good one. However, it should be possible to formulate an answer fairly concisely. I wish, therefore, to begin by clarifying some terms, and – in particular – the location of my version of realism along some philosophical axes. I should emphasize that my use of the term 'realism' is *not* intended to specify allegiance to any particular philosophers.

There are a number of conditions which are generally accepted as characterising realist discourse.[13] They are:

a. **Truth-Aptness**: a statement is truth-apt if and only if it is able to be true or false. This is an apparently simple condition, but difficult to analyse. What one takes truth-aptness to consist in will depend on one's theory of truth. Minimalists (e.g. Blackburn, Wright, Horwich) take truth-aptness to be a property of sentences of any discourse which is possessed of the appropriate surface syntax, and subject to standards of discipline governing warranted assertion.[14]
b. **Cognitivism:** a discourse is cognitivist if basic statements within that discourse serve to express beliefs. Cognitivism is therefore related to truth-aptness (it entails truth-aptness), but distinct. Non-cognitivism, in contrast, holds that basic statements within that discourse serve primarily to express non-cognitive states, such as desires, pro-attitudes, etc.
c. **Objectivity:** At a first pass, we might say: a statement is objective if its truth is independent of the perspective (or situation) of whoever utters it.[15] So the

[12] As Hume notes, 'There is no method of reasoning more common, and yet none more blameable, than, in philosophical disputes, to endeavour the refutation of any hypothesis, by a pretence of its dangerous consequences to religion and morality.' (Hume 1975: 96).

[13] C.f. Timmons 1999: 35, 'Moral realists . . . hold that moral properties and facts exist and that their existence and nature are conceptually and metaphysically independent of our moral beliefs and theories, including our warranted or even ideally warranted beliefs and theories . . .'.

[14] Lenman 2003.

[15] c.f. Nagel 1986: 4-5, 'Objectivity is a method of understanding . . . only derivatively do we call objective the truths that can be arrived at in this way . . . A view or form of thought is more objective than another if it relies less on the specifics of the individual's makeup and position in the world, or on the character of the particular type of creature he is.'

statement 'this tastes good' is subjective, in that our best understanding of the statement is as elliptical for 'this tastes good to me', and the truth or falsity of that statement is relative to the utterer.[16] Conversely, the statement 'that food tastes good to IEM', on the other hand, is objective. But the notion of objectivity is difficult to make precise. I suspect that this is partly due to the fact that it is, as J L Austin would put it, a 'trouser-word'.[17] That is, we understand what it is for something to be objective by contrasting it with what it is for something to *fail* to be objective – in other words, to be subjective. Whilst 'objective' is often used in relation to facts, truths, and so on, 'subjective' is not.[18] So we think that the word 'objective' is used not to modify terms such as 'fact' or 'mind-independent truth', but rather to underline them. In a similar vein, David Brink claims that the relevant question about objectivity in ethics is, '[i]s ethics or can it be objective in the way that other disciplines, such as the natural sciences, are . . .', and suggests that the commonsense question of objectivity is best understood as a question about realism.[19]

d. **Mind-independence**: a statement is mind-independent if and only if its truth is not, or is only accidentally, a function of some person(s) mental states, actual or otherwise. There are therefore at least two senses of 'mind-dependence' available, one where the relevant mental states are actual, one where they are counterfactual. In order to preserve a contrast with objectivity, mind-independence (in the sense in which I am interested) requires both actual and counterfactual independence. So the statement 'x is red' is objective, in that its truth does not depend on the perspective or situation of the utterer, but mind-dependent, in that colour terms are response- (and hence mind-) dependent. Similarly, we might suppose that there are properties which are constituted by what fully informed, ideally rational agents would agree. Trivially, the property of being what fully informed, ideally rational

[16] C.f. McDowell 1998: 113, 'A subjective property, in [Mackie's] sense, is one such that no adequate conception of what it is for a thing to possess it is available except in terms of how the thing would, in suitable circumstances, affect a subject – a sentient being.' McDowell is charging Mackie with adopting a notion of objectivity which contrasts with *this* understanding. But the relevant notion of objectivity should align with the notion of a property being 'in the world'. Response-dependent concepts, such as 'is red' and 'is good', are, in the Mackian sense, subjective – but nonetheless are in the world, in McDowell's sense, being independent of our vicissitudes.

[17] Austin 1964: 70-71. In the case of 'objective', it is word 'subjective' which 'wears the trousers'.

[18] C.f. 'It is an objective, mind-independent truth that . . .', 'It is a matter of objective fact that . . .', with 'It is a subjective truth that . . .', 'It is a matter of subjective fact that . . .'.

[19] Brink 1989: 5.

agents would agree on (read *de dicto*) is one such property.[20] Such properties, dependent on counterfactual claims about various persons' mental states, are mind-dependent.

e. **Partial verity**: this is the condition that at least some of the claims within the discourse are true. This condition is primarily relevant in the case of moral realism, which needs to be distinguished from an error-theoretic account, on which statements within moral discourse are truth-apt, expressive of belief, and objective (or mind-independent), but false.

Traditionally, the debates over the truth-aptness of moral discourse, on the one hand, and over cognitivism, on the other, have been run together. However, it has been argued that non-cognitivist positions can allow for truth-aptness. Kalderon has put forward a position which he calls 'moral fictionalism', which combines the claim that moral judgements express propositions (and are therefore truth-apt), with the claim that the attitudes involved in sincere acceptance of a moral sentence are not beliefs (non-cognitivism).[21] Gibbard, Blackburn and others have combined a minimalist account of truth with some form of non-cognitivism. So it doesn't seem that truth-aptness *per se* is going to do the work of distinguishing between cognitivist and non-cognitivist positions (nor, by extension, between realist and irrealist positions). Wright notes that, once we grant a minimalist conception of truth (according to which truth-aptness is a common property of discourses), moral discourse turns out to be truth-apt –but goes on to argue that there are a number of different ways in which discourses may turn out to be realist.[22] These are to be distinguished by the answers they give to a set of test questions. Wright suggests that the distinction can be drawn in (amongst others) the following terms:

f. **Superassertibility:** a statement is superassertible if and only if 'it is, or can be, warranted and some warrant for it would survive arbitrarily close scrutiny of its pedigree and arbitrarily extensive increments to or other forms of improvement of our information.'[23] One *anti-realist* claim is that superassertibility (within that discourse) serves as a truth-predicate.

[20] Read *de re,* this would leave open the possibility that fully informed, ideally rational agents would agree on some claim P just because P is true, and not because the truth of P is constituted by their agreement. The *de dicto* reading entails necessary coextensivity, and implies property identity (although the jury is out on whether necessarily coextensive properties are identical: philosophers tend to be wary of the suggestion that "has precisely three sides" and "has precisely three angles" pick out the same property when predicated of two-dimensional shapes, for instance).

[21] Kalderon 2005.

[22] Wright 1992.

[23] Wright 1992: 48.

g. **The Euthyphro Contrast**: suppose we accept the biconditional that an act is pious if and only if it is loved by the gods. This then admits of two readings: i) acts are pious *because* they are loved by the gods, and ii) the gods love acts *because* they are pious. These readings are *anti-realist* and *realist* respectively, with the distinction being drawn in terms of the direction of explanation.

h. **Cognitive Command:** a discourse exerts Cognitive Command if and only if 'It is a priori that differences of opinion formulated within the discourse, unless excusable as a result of vagueness in a disputed statement, or in the standards of acceptability, or variation in personal evidence thresholds . . will involve something which may properly be regarded as a cognitive shortcoming'.[24] A 'cognitive shortcoming', here, may be a matter of an error in reasoning, shortfall of information, or similar defect. Discourses which exert cognitive command are realist, and those which do not are anti-realist.

i. **Width of Cosmological Role:** *width of cosmological role* is the extent to citing the kinds of states of affairs with which a discourse deals is 'potentially contributive to the explanation of things *other than*, or *other than via*, our being in attitudinal states which take such states of affairs as object.'[25]

I cite these potential constraints for two reasons: firstly, to demonstrate that 'realism' need not be seen as a unified position, and secondly in order that I may help myself to the terminology in due course. I will be arguing, at least initially, that the appearances of moral discourse cohere with 'realism' in the sense obtained by conjoining **a, b, c,** and **d.** That is, moral discourse and practice is realist insofar as it purports to describe a set of objective, mind-independent facts. Furthermore, we should also subscribe to **e.** It is certainly not clear that moral discourse fails to display Cognitive Command; I discuss this question in some depth in Chapter Two. The position which will be developed over the course of this thesis is that within this construal of moral discourse and practice, there is a further subdivision to be had, between that part of morality wherein superassertibility counts as a truth-predicate and the Euthyphro contrast runs in the 'irrealist' direction, and that part where the converse is true. That is, the appearances are roughly veridical, but not entirely: we should be in part moral realists, and in part moral irrealists. However, the irrealism in question will be modest: retaining cognitivism, descriptivism, objectivism, and truth-aptness, whilst denying strong mind-independence. But more on this in due course.

[24] Ibid. 144.
[25] Ibid. 196.

4. Phenomenology

The first set of arguments for moral realism draws on the nature of our moral experience. The process of forming moral judgements (whether about acts, intentions, or outcomes) has the appearance of a response to certain (moral) features of the world. It does not seem that we are merely expressing sentiments: from the point of view of someone involved in moral deliberation, the process is one of considering reasons and choosing to act accordingly, rather than simply being drawn along by the strongest relevant desire.[26] From the point of view of persons engaged in moral disagreement, the disagreement is genuine (as opposed to a mere contrast of tastes), and can perhaps be resolved by adducing and weighing reasons. When we form moral judgements about particular cases, the moral quality of an act seems to be properly located in the act, rather than in our response to it. There is, as Smith points out, a marked contrast between ascriptions of moral wrongness, on the one hand, and ascriptions of being nauseating, on the other: when we take an act to be morally wrong, it is the act itself which is salient; when we take an object to be nauseating, it is our inner state which is salient. And this parallel with perception can be extended: our moral judgements, whether of particular cases or of general principles, seem often to be immediate (i.e. non-inferential).[27] Persons may be better- or worse- placed to form correct moral judgments: partiality and bias detract from the warrant of any particular moral judgement, while sensitivity and a capacity to see things from many points of view are conditions which support warranted moral judgement.[28]

These considerations constitute a (modest) argument for moral realism, because these features of our moral experience are (I claim) best explained by moral realism. These features cluster together into two groups: the first group provides support for a perceptual model of moral epistemology; the second simply provides support for moral realism. The perceptual model, if true, gives direct support for moral realism, since perception just is a matter of direct sensitivity to properties or objects which are in the world, ready to be perceived.

There are features common to both moral judgement and ordinary, vanilla perception. The first of these is immediacy: although perception is causally mediated, the experience of perception itself is immediate, in that it need not involve inference. The moral case is

[26] Haldane & Wright 1993: 250.

[27] C.f. Mackie 1977: 31-5, McDowell 1998: 131.

[28] One worrisome case is that of the moral status of the family: it is a commonly-held view that we are morally permitted, if not obliged, to give preferential treatment in many situations to family members. But our view on what is morally permissible as regards our own family members is certainly coloured by our relation to them; indeed, the thought is precisely that this partiality is appropriate. Nonetheless, discussion of this thought need not itself be biased or partial.

supposed to be, at least in some cases, similar in this respect.[29] There are moral judgements which do not depend for their warrant on some further belief; the regress of justification ends with a foundational claim, one which lacks further justification, but whose status seems to be that of a datum for moral theorising. The fact that we suppose this not to destroy warrant suggests that the perceptual analogy is appropriate: if we can view our moral judgements as, in at least some cases, a matter of sensitivity to the moral facts, then it is straightforward to make sense of such moral judgements as warranted in the absence of further justification.

The second feature is the qualitative nature of moral judgement. Just as my experience of seeing a bottle of beer has a certain subjective phenomenal character, or quale, so too with moral experience – at least in some cases. I am not suggesting that all moral judgements have a distinctive phenomenal character. There may be dispassionate judgements about morality, just as there are dispassionate judgements about science. But when it comes to judgements about particular cases, it often seems that the moral quality is located in the object, rather than in ourselves, and our experience of the object is coloured by this quality.[30] There is something that it is like to see an instance of wanton cruelty as wicked, just as there is something that it is like to see a painting, or detect a particular taste – and this quality has objective pretensions.

Of course, this may be misleading. 'Take any action allow'd to be vicious,' Hume says, ' . . . [e]xamine it in all lights, and see if you can find that matter of fact, or real existence, which you call *vice*. In which-ever way you take it, you find only certain passions, motives, volitions, and thoughts . . . The vice entirely escapes you, as long as you consider the object. You can never find it, till you turn your reflection into your own breast, and find a sentiment of disapprobation . . .'[31] Similarly, Blackburn takes the apparent moral quality of actions to be projections of our own sentiments, results of our thoughts and sentiments, rather than perceptions proper. But the claim here is that the appearances are realist: if they are to be explained away, we need a positive argument against realism combined with some form of explanation.

[29] One might think that there is a direct analogy between perceptual justification and moral justification. Perceptual beliefs are warranted when gained by competent judges in suitable circumstances; their justification results from direct contact with their objects under felicitous conditions. Some moral beliefs, similarly, are warranted when formed by competent judges under correct circumstances, and involve direct contact between judge and object. In the cat-burning case, for instance, we can apprehend the wrongness of the deed directly – and we may derive a general belief from observing multiple instances of this. So here it seems that a belief concerning the general connection between gratuitous harm and moral wrongness derives its warrant from the particular judgements concerning wrongful, gratuitous harm – and we may take these to be instances of direct (immediate) perception. But I am not convinced by this line of thought, for reasons which will be discussed shortly.

[30] C.f. Haldane & Wright 1993: 246, McNaughton 1988: 56.

[31] Hume 2000: 301.

In a similar vein, Nagel claims that 'reasons . . . usually present themselves with some pretensions of objectivity to begin with, just as perceptual appearances do. When two things look the same size to me, they look at least initially as if they *are* the same size. And when I want to take aspirin because it will cure my headache, I believe at least initially that this *is* a reason for me to take aspirin.'[32] The claim here is not that all desires entail beliefs about reasons; this would commit us to thinking that cats, badgers, and other creatures have beliefs about reasons, and that is implausible. Rather, when my desires come with justificatory clauses, I believe that the content of that clause relates to something which actually is a reason.

At this point, one might worry that the presumptive argument for realism has gone too far: these considerations motivate not only realism about morality, but also realism about the beautiful, the comic, the sexually appealing, and so on. There are three points to be made in response. Firstly, the argument here is only intended to establish a weak presumption. This is entirely compatible with finding conclusive reason to adopt antirealism with respect to a discourse. Secondly, each domain of discourse needs to be considered independently. Thirdly, in cases where realism would be particularly implausible – the case of the comic, for example – it is unclear that our experience strikes us as sensitivity to some (distinctively comic) aspect of the world. Put loosely, the laughter comes first, and then the ascription of comic quality. It is entirely congruent with the phenomenology of the comic to suppose that the practice deals primarily with our responses to the world, rather than with features of the world itself.[33]

It is also worth noting that the perceptual model is well able to accommodate ethical particularism. Particularism is the claim that 'the possibility of moral thought and judgement does not depend on the provision of a suitable supply of moral principles.'[34] The core argument for particularism depends on the thought that a feature which provides a reason for action in one situation may provide no reason, or an opposite reason, in another: this is *reasons holism*. Given this, Dancy suggests that we cannot generate true exceptionless ethical principles; we cannot claim that features invariantly provide reasons, since that would involve the denial of holism. Any such invariance would be a 'cosmic accident'. Hence, ethical thought must take particular cases as prior.

[32] Nagel 1986: 149.

[33] Comic discourse may still be primarily descriptive, but the content of the descriptions would deal either with individuals' responses, or with what ideal observers would respond to in certain ways.

[34] Dancy 2004: 74.

Although I am not convinced that Dancy is entirely correct, one important insight contained within his claim ought, I think, to be retained – and that is that our judgements about individual cases are often epistemically prior to any principles which attempt to systematise these intuitions.[35] We should accept that, in at least some cases, individual moral judgements have priority.

Lastly, McNaughton offers an argument from the 'authority' of moral demands.[36] It appears (to us) that the demands of morality are independent of our desires: we ought to do such-and-such because it is the right thing to do, not because we wish to do so. Moral demands have a kind of authority. But if moral judgements are (for instance) merely expressions of sentiment, then they would not have this authority. Were this true, it would pose a problem for expressivist theories, at least. It is now generally accepted, however, that this kind of authority can be explained by irrealist and realist theories alike. What would be problematic is for moral demands to be expressions of mere whims, or immediate desires. However, there are many other options: moral demands could be construed as functions of what agents would desire under ideal circumstances, of what agents' second-order desires require, of what agents would feel guilty about, and so on. Whilst it is often true that moral demands are experienced as external to us, we should not treat this as anything more than the already-discussed pretension to objectivity.

There are some general worries about phenomenological arguments. Bloomfield claims that any argument from moral experience to moral reality fails, because '[particular moral] experiences can be fully explained in terms of moral attitudes and moral judgements we may make, and (therefore) positing moral reality based on them is more than is needed to explain the data.'[37] Furthermore, the existence of intractable moral disagreement should undermine any confidence we have in our deeply held moral beliefs – and hence make us wary of 'making an inference to moral reality from the contents of a data set that is so easily confounded.'[38] Loeb argues that the relevant phenomenology supports, at most, the claim that morality is objective – a claim which can be agreed upon by Kantians, error theorists, etc.[39]

[35] See, in particular, Lance & Little 2005. We should be wary of the claim that generalists are committed to the existence of invariable, exceptionless moral principles. Furthermore, generalism about value is compatible with particularism about what we ought to do, all things considered.

[36] McNaughton 1988: 48. A similar view is put forward in Shafer-Landau 2003: 29.

[37] Bloomfield 2001: 7. C.f. Hume 1978: 468-69, Harman 1988.

[38] Bloomfield 2001: 8.

[39] Loeb 2007, esp. pp. 470-1. Curiously, Loeb claims that Mackie's error theory is 'anti-objectivist'. This is implausible: Mackie's position was precisely that moral discourse and practice is committed to the existence of mind-independent, intrinsic normative properties – which, it turns out, don't exist, leaving our moral theory systematically in error. Loeb also characterises Lillehammer as an error theorist. He isn't (Lillehammer, in discussion).

Loeb has two core worries about arguments from experience, relating to explanation and justification respectively. The explanation worry runs as follows: if the argument from experience is supposed to be an argument to the best explanation, we need some account of *how* moral realism is supposed to explain moral experience. But it is far from clear that moral facts can explain anything: we are therefore not entitled to use this claim to ground a presumptive argument for moral realism. Arguments which turn on the claim that moral realism saves some key appearance of moral discourse need to explain why the appearance is likely to be veridical whilst avoiding a collapse into the argument to the best explanation. More generally, we don't have any reason to think that the moral appearances come with automatic warrant, any more than our moral beliefs do: we should not think that we are entitled to cleave to appearances unless given reason to do otherwise. Worse still, we don't even have reason to think that the appearances are constant: philosophers and laymen differ, for instance, in what they take to be the (metaethical) case.[40]

I will discuss the issue of moral explanation at some length later in this chapter, in Section 9. For the time being, it will suffice to note two points. Firstly, explanation need not be ineliminable to count as good explanation; even if we could provide an alternative explanation of moral experience in terms of particular moral attitudes and judgements, this would not entail that the straightforward moral explanation is illegitimate. Secondly the presumptive argument on offer is weak: the claim is not that moral experience requires explanation in terms of moral reality, but that it lends itself quite naturally to explanation in terms of moral reality (or something fairly close). I deny that moral disagreement need be, or indeed is often, intractable – but even were it intractable, Wedgewood (forthcoming) has argued persuasively that we may still retain confidence in our deeply-held moral beliefs, since (roughly) we can use those same beliefs in evaluating the credibility of our interlocutors, and hence be unworried by the existence of disagreement. I should treat the existence of disagreement as undermining the level of justification of a disputed belief to just the same extent as I view my interlocutor as my epistemic peer; but in the case of moral disagreement, the fact that my interlocutor (for instance) thinks that apartheid is morally acceptable is enough to disqualify him from counting, from my point of view, as my epistemic peer. So the fact that this disagreement exists need not lead me to doubt my own deeply held moral beliefs. And in any case, the relevant datum is not the exact content of our moral beliefs or judgements, but rather their nature – the pretensions to objectivity, phenomenal character, etc.

[40] Loeb 2007: 478, '[it is] inappropriate to claim that morality *seems* objective to the average person.' Again, this seems poorly motivated. Goodwin & Darley 2008, an empirical study of 'folk' metaethics, indicates that laypeople do generally treat morality as objective, although the degree of objectivity attributed to various ethical beliefs varies with their content.

5. Semantics

5.1. The Frege-Geach Problem

The second case for realism attends to moral discourse itself. Moral discourse has a group of features which characteristically accompany descriptive discourse. Specifically, moral statements are grammatically apt for truth (that is, appending ‘it is true that . . .’ to a moral statement yields a syntactically complete sentence); we can generate moral arguments which appear valid; moral statements can be negated, and embedded in unasserted contexts without a change in meaning.

There are a number of distinct issues here, but the initial worry is that raised by Geach – the problem of embedding. As originally posed, the problem is that treating moral language as anything other than descriptive poses problems for analyses of claims such as, ‘If gambling is bad, inviting people to gamble is bad.’[41] The problem also arises for other discourses for which we might want to attempt an expressivist analysis – discourse about the comic, the beautiful, and so on – but since we are interested in the case of morality, I will restrict the discussion to moral discourse. The upshot – that the Frege-Geach problem is a genuine problem for expressivists – remains the same in all cases, although the theoretical pressures in favour of anti-realism may vary. Returning to the case at hand, then, the problem is that we cannot analyse ‘gambling is bad’ as ‘Boo! for gambling’, because the conditional then turns out to mean:

FG: If Boo! for gambling then Boo! for inviting people to gamble.

Which is grammatically unsound. Nor could we make sense of the putative fact that the two occurrences of ‘gambling is bad’ mean the same. Geach concludes that we must analyse calling something “P” in terms of predicating “P” of that thing: moral condemnation of gambling must be understood in terms of predicating ‘is bad’ of gambling, rather than *vice versa*. Call this the *embedding* problem.

[41] Geach 1960: 224. Strictly speaking, this is an argument for descriptivism, rather than realism – but it forms part of the argument for realism.

This challenge can be developed further. Closely related to the embedding problem is the problem of validity.[42] We can generate moral *modus ponens* arguments, such as:

A_1: If gambling is bad, inviting people to gamble is bad.
A_2: Gambling is bad.
Hence
A_3: Inviting people to gamble is bad.

Such arguments are apparently valid – but (so the challenge goes) our best understanding of validity requires that the premisses and conclusion be truth-apt, and this entails the denial of non-cognitivism. Call this the *validity* challenge. The stock expressivist response is to treat the issue of validity as secondary to the issue of inference. That is, what we ought to be concerned with (if the expressivist is correct) is the thought that someone who accepts the premisses, but denies the conclusion, of a valid moral *modus ponens* has made a mistake. And we can do this by talking of inconsistent attitudes, plans which are not jointly realisable, etc. [43] In *Spreading The Word*, for instance, Blackburn accounts for the force of **FG** in terms of an attitude which involves disapproval of the conjunction (it is wrong to tell lies) and (it is not wrong to get your little brother to tell lies).[44] Against this move, it has been argued that this fails to account for the fact that we see instances of moral *modus ponens* as instances of *modus ponens*. [45] But for the quasi-realist, we *do* have an instance of modus ponens. If Blackburn's quasi-realism can be made to work, then not only can we earn the right to talk of moral facts, properties, and suchlike – but also all of the talk of truth, logical relations, and so on, that the moral realist wishes to retain. We begin by identifying a functional structure of attitudes which is isomorphic to the logical structure of arguments, conditionals, and so on: we can then project this functional structure onto the discourse.[46]

[42] Because validity is a matter of relations between propositions, and for the proposition embedded in A_1 to be the same as the proposition which occurs in A_2, the clause 'gambling is bad' must mean the same in A_1 as in A_2. Constancy of meaning is required for constancy of the propositions expressed, and hence for validity.
[43] E.g. Blackburn 1984, 1998, 2002, Gibbard 1990, 2003.
[44] Blackburn 1984.
[45] Schueler 1988.
[46] However, on Blackburn's account, the validity of a moral *modus ponens* argument derives from the projection, or from the functional structure which underlies it. But an important feature of valid arguments is that they are truth-preserving *in virtue of features of the arguments themselves.* On Blackburn's account, we have the standard sense of validity, and another sense – call it validity* - whose force is derivative, rather than intrinsic.

5.2. Minimalism

If this project can be successfully carried out, then the expressivist has laid the foundations for adopting what has come to be known as *minimalism* about truth. Minimalism about truth claims that any discourse which is possessed of the right syntax and 'standards of discipline' counts as truth-apt; thus, statements within comic discourse, aesthetic discourse, and moral discourse may all count as equally truth-apt provided that they satisfy certain minimal criteria.[47] Crucially, the notion of truth at work here is taken to imply nothing of particular metaphysical import: there is not held to be any one property which all true statements have in common, beyond their satisfying the minimal platitudes. The correspondence condition – that a statement is true if and only if it corresponds to the facts – is held to be a platitude. That is, it is trivially true, and adds nothing to the claim that a statement is true. In particular 'it is a fact that x' and 'it is true that x' are held to be equivalent: accepting that moral statements are true because they correspond to the moral facts does not commit one to 'moral facts' (in the realist sense) above and beyond what is specified in true moral statements.

Now it is not obvious that the correspondence platitude is quite as innocuous as minimalists have claimed. Wright attempts to establish this innocuity as follows:

> 'Let it be indeed a platitude that
>
> "P" is true if and only if "P" corresponds to the facts.
>
> Is anything lost by paraphrasing this as
>
> (CP) "P" is true if and only if things are as "P" says they are?
>
> Presumably not . . .'[48]

But whenever we can affirm that "P" says that P, we can affirm CP. And that, coupled with the Deflationary Schema, yields the correspondence platitude.[49] So the correspondence platitude collapses into the entirely metaphysically innocent '"P" if and only if P', and so is itself innocent.

[47] These standards govern the assertion of the various sentences of the discourse, and derive from the functional structure (in the moral case) of the underlying attitudes.

[48] Wright 1992: 25.

[49] "P" is T if and only if P.

Now one might, *contra* Wright, think that something is indeed lost by carrying out this paraphrase. On the one hand, it seems likely that correspondence does not, by itself, imply any interesting story about explanation (e.g. whether the fact that p *explains* the truth of ‘p’), since, as the phrase has it, correspondence is not causation – hence the correspondence relation may be accidental, and therefore need not entail an explanatory relation. But I will leave this issue to one side. What does seem clear, however, is that correspondence is a dyadic relation. To assert that the correspondence platitude holds of a statement is to assert a relation between two distinct entities: the content of the statement, and whatever it is that makes the statement true.[50] It is not clear, on the other hand, that CP carries the same commitment. An example might help to make this clearer. Contrast:

C “This joke is funny” is true if and only if “This joke is funny” corresponds to the facts.

with

C* “This joke is funny” is true if and only if things are as [the statement that] “this joke is funny” says they are.

C carries with it the implication that there is a fact of the matter as to whether or not this joke is funny; C*, on the other hand, is ambiguous. Suppose that we are antirealists about comic discourse; there is, at least in some cases, no fact of the matter about whether a certain joke is funny, and our best understanding of whether or not jokes are funny refers to our responses to those jokes, rather than to some mind-independent facts about jokes. In this case, at least, something *is* lost by carrying out Wright’s suggested paraphrase, namely the dyadic nature of the correspondence relation. That is, we have a semantic intuition about the nature of moral discourse: it has the appearance of serving a descriptive function, rather than merely an expressive function. The fact that we would hesitate to apply the correspondence scheme as a platitude for comic discourse, but not in the moral case, suggests that it carries some metaphysical import (a commitment to ‘the facts’, for one). If this is correct, then there is reason to take the correspondence schema as a marker of realism (or, more accurately, descriptivism together with objectivism of some sort), rather than as a platitude. Since the Correspondence Platitude is indeed platitudinous for everyday ethical theorising, this commits the ethical theorist to a form of descriptivism: the content of ethical statements is descriptive

[50] I owe this suggestion to John Skorupski (discussion).

insofar as it purports to describe some set of facts whose content is distinct from the theoretical practice.

Expressivists could attempt to make sense of the dyadic nature of the correspondence relation by treating the Correspondence Platitude as saying that moral sentences are true if and only if the world warrants their utterance.[51]

CP$_E$: A moral sentence, M, is true if and only if the world warrants the utterance of M.

One obvious difficulty with this suggestion is that warrant and truth are distinct: we might have some kind of practical warrant for talk of human rights, for instance, but this does not by itself entail that warranted sentences involving talk of human rights are literally true. By a similar token, warrant and correspondence are distinct notions, although correspondence may entail warrant in the case of descriptive sentences. So although this move does allow the expressivist to make some sense of the Correspondence Platitude – and to introduce a dyadic relation of roughly the right kind (that is, between moral and non-moral relata), it does not enable the expressivist to capture the whole sense of the Platitude. Expressivists therefore face a dilemma: either introduce dyadicity at the cost of the notion of correspondence (this is the revisionary interpretation of CP), or retain the notion of correspondence, in which case the commitment to dyadicity seems to imply that all (positive, atomic) moral sentences turn out to be false. Neither option is obviously appealing.

Again, the claim here is not that error theory about the subject discourse is to be ruled out – we have yet to see whether the denial of non-cognitivism yields a plausible position, or whether non-cognitivism itself is not sufficiently well-motivated to be our best metaethical theory – but rather that the adoption of expressivism, and hence denial of descriptivism, carries with it the implication that our moral practice is in error. The error, here, is a function of the fact that our discourse carries with it realist commitments, the commitment embedded in the Correspondence Platitude being one such example. The platitude, I have suggested, is not metaphysically innocent, and must therefore be either discarded or taken seriously. The descriptivist then faces the challenge of providing an account of the nature of the relevant facts, and of how we could get to know them: the plausibility of descriptivism as a thesis will then turn on the success of the metaphysical and epistemological projects. Importantly, I think that these projects can be made to work. Again, the claim is not that our commitment to the

[51] Sinclair (discussion).

Correspondence Platitude constitutes a particularly strong argument for realism (or descriptivism more broadly), but rather that the burden of proof lies with the expressivist.

5.3. The Frege-Geach Problem Again

Returning to the ‘validity’ challenge, we need now to consider whether the expressivist has successfully managed to account for the logical relations between moral sentences. Clearly, the simplest aspect of the validity challenge is the issue of inconsistency: we need to explain why a moral sentence and its negation are inconsistent. Unwin, amongst others, has argued[52] that this poses a serious problem for the expressivist, since the posited attitudes lack sufficient structure. Suppose, for instance, that the basic attitudes involved are approval and disapproval. Failing to approve of something is clearly distinct from disapproving of it – but how else are we to analyse the claim ‘it is not right that I do this?’. One solution, advanced by Mark Schroeder, is to introduce additional structure: rather than approval and disapproval, we may use ‘being for’, and take this to operate over attitudes towards deeds; ‘murder is wrong’ is then understood as expressing ‘being for blaming for murdering’, and so on.[53] But *that* move yields a further problem when we consider ‘mixed’ sentences, such as ‘murdering is wrong and grass is green’. Such sentences are inconsistent both with the negation of the moral component, and with the negation of the descriptive component.[54] So it looks as if we ought either to be across-the-board descriptivists, or across-the-board expressivists. But being an expressivist about belief would be a curious position – and, after all, one of the motivations behind expressivism relies on the distinctively practical element of moral judgement (most naturally explained via conative, rather than cognitive, states).[55]

The problems of embedding, validity, and negation, then, are genuine problems for the expressivist. They are problematic not because expressivists cannot provide solutions, but because of what these solutions commit the theory to. I think that Schroeder correctly identifies the core of the problem: the expressivist needs attitudes which have the right levels of structure. But we are not entitled to simply stipulate that the relevant attitudes behave in the correct ways – we need to *explain* the relevant behaviour. Descriptivists, of course, can already do this, and, better still, can do this in a way which is entirely contiguous with the explanations which exist for ordinary, non-evaluative discourse. In the non-evaluative case, the ‘natural’ interpretation of the discourse (i.e. as straightforwardly descriptive) leads to an

[52] Unwin 1999.
[53] Schroeder 2008
[54] Schroeder 2008: 26.
[55] Ibid.

explanation of the relevant phenomena. Believing both that P and that not-P is inconsistent, because P and not-P are propositions which cannot jointly be true. But if we are comfortable with the presumption of descriptivism in the non-evaluative case, we should also be comfortable with the presumption of descriptivism in the evaluative case.

There is a third element to the Frege-Geach problem: the problem of making sense of the thought that accepting the premisses of a valid *modus ponens* gives reason to accept the conclusion.[56] Whereas the validity challenge concerns the logical relations between sentences (or the propositions that they express), this challenge is a matter of relations between beliefs (or whichever attitudes are involved in the acceptance of a moral sentence). Call this the *inference* challenge. For the cognitivist, this challenge is easily met – but the non-cognitivist would need to show how the attitudes involved in sincere acceptance of a moral sentence (call them M-attitudes) can stand in the right kinds of relations. As with the issue of validity, Kalderon thinks that this challenge can be met: if the attitudes involved in accepting moral sentences are something like desires in the 'directed-attention' sense (that is, involve, in part, taking certain features of the world to be salient in deliberation), then accepting a moral sentence involves (inter alia) reconfiguring the reasons that we take ourselves to have.[57] So accepting a moral conditional involves reconfiguring our affective states such that we endorse certain combinations of attitudes: the acceptance of the conditional reshapes the (subjective) landscape of reasons.

6. (Humean) Psychology

One motivation for expressivism – which I will treat in more detail below - is the aim to connect sincere moral judgements with motivation. On the Humean model (which takes beliefs as motivationally inert, desires as intrinsically motivational, and the two to be 'distinct existences' with no necessary interconnections), this is done by viewing sincere moral judgements as expressive of desires, rather than beliefs.[58] But this does not square with the suggestion made immediately above, that the function of moral judgement is primarily descriptive, rather than expressive. Nor does it square with the appearance of moral judgements as beliefs, rather than desires.

[56] See Eklund 2007. Eklund raises this as a potential problem for moral fictionalists (of whom, more in Chapter Two).
[57] See Scanlon 1998: 39, 'A person has a desire in the directed attention sense that P if the thought of P keeps occurring to him or her in a favourable light . . . if the person's attention is directed insistently towards considerations that present themselves as counting in favour of P.'; also Kalderon 2008.
[58] Hume 2000: 400; see also Smith 1994: 92-130.

There are two distinct thoughts here. The first is that agents, when using moral discourse, see themselves (implicitly or otherwise) as primarily engaged in description, rather than expression of attitudes. Terence Cuneo, for instance, has suggested that agents engaged in moral discourse intend to assert moral propositions, robustly construed.[59] As already discussed, the possibility of taking an error-theoretic approach to this is (*if* the suggestion is correct) left open: if an agent intends to assert a fact about the will of God when using moral discourse, for instance, they are engaged in thoroughgoing error. The question of whether an error-theoretic stance has associated costs then arises. Sophisticated expressivists tend to claim that their position can retain all of the important features of morality: one can hold that moral utterances serve primarily as expressions of attitudes, rather than descriptions of facts, and still use moral discourse for the relevant purposes, viz., co-ordinating attitudes and actions, establishing reciprocity, etc. But this presupposes that the only relevant features of morality are those which do not depend on a commitment to descriptivism of some sort. And that is far from clear: the divine command theorist, for instance, may well think that morality depends for its authority on its reference to a (divine) lawgiver.

So there is an issue here of the normative import of one's metaethical theory. Just as divine command theorists frequently maintain that without God, nothing is forbidden, it has been argued that realism has an advantage over non-cognitivism in this respect; our fundamental normative commitments, construed realistically, are non-arbitrary if they are true, whereas the same commitments on a non-cognitivist construal turn out to lack justification, and are so arbitrary.[60] Basic moral commitments on a non-cognitivist construal are pro- or con- attitudes, rather than beliefs; on a crude understanding of non-cognitivism, they are desires. But the mere fact that we desire something does not give us reason to desire that thing. And since the process of moral justification involves, sooner or later, citing our basic moral commitments, the chain of justification terminates in citing a mere desire – one which itself lacks justification. So, concludes the realist, the entire justification collapses. Moral justification, in order to be successful, needs to terminate in claims which are somehow self-justifying.

There are at least three problems with this, which I will mention, rather than discuss in detail. Firstly, there are strong non-cognitivist rebuttals of this claim: the charge of arbitrariness has force when viewed from a standpoint internal to the moral practice, and from this standpoint many of our fundamental commitments are fully justified. Our moral practice sets standards for justification, just as it sets standards for behaviour, and commitments which meet these standards are (in Shafer-Landau's terminology) non-arbitrary. Secondly, it is not clear that

[59] Cuneo 2006.

[60] Shafer-Landau 2005: 28 – 30.

realism escapes the charge of arbitrariness; in the absence of some story which connects our moral beliefs to the (realist) facts, our fundamental moral commitments may still turn out to be arbitrary (again, in Shafer-Landau's terminology).[61] Thirdly, and most worryingly, it seems that there is a straightforward abuse of the English language at hand. Choices or decisions may be arbitrary; roughly, a choice is arbitrary if and only if one has no reason to choose that option over any other. But then desires (broadly construed) are not the kinds of things which can be arbitrary. To be grammatically sound, Shafer-Landau's claim would have to be that our *choosing* one set of fundamental normative commitments over another is arbitrary, and this supports both the objection that the truth of those commitments will not (and by itself, certainly, cannot) justify those choices, and the objection that from within our moral practice there *is* reason to cleave to some fundamental commitments, rather than others.[62]

There may be other ways in which one's metaethical view has first-order import – indeed, on the view which I wish to set forward, there are several.[63] I will leave a detailed treatment of this topic for further chapters: in brief, the thought is that we can provide good non-moral reason for taking certain features of the world seriously, and that for some part of ethics – namely, the part to be given a constructivist account – the epistemology collapses into the metaphysics. The metaphysical account, therefore, provides some insight into how we should engage in ethical theory. But more on this in due course.

The second thought which I wish to explore here is that our moral discourse and practice purports to deal in beliefs, rather than attitudes. This is partly evidenced by semantic considerations (if the discourse is assertoric in form, this gives reason to think that the function of the discourse is to express beliefs, rather than attitudes), but also partly by considerations of moral psychology. Now 'belief' is a somewhat underspecified notion. The standard distinction, drawn within the Humean framework, is that beliefs have 'mind-to-world' direction of fit, and desires 'world-to-mind'. That is, beliefs aim to fit the world, while desires aim to change it.[64] Note that these accounts are not merely descriptive: beliefs which do not have any tendency to 'fit' the world are still beliefs, and likewise (*mutatis mutandis*)

[61] See Kawall 2005.

[62] I owe this point to Sarah Broadie (discussion).

[63] Fantl (2006) discusses some trivial instances of this.

[64] I should add that this model seems woefully underequipped to deal with the range of possible psychological states that exist – in particular, the range of pro- and con- attitudes which the expressivist wants to use in her metaethical theory. The central objection here runs as follows: there are plenty of pro- and con- attitudes which do *not* aim to change the world, even once we include 'aiming to keep things as they are' in the definition: for instance, idle desires (it would be nice to be able to fly unaided, but that doesn't motivate me to acquire wings), mild states of disapproval (well, I'm not keen on football, but I'm happy to leave people to it), and so on.

for desires. So the characteristic explanation of 'directions of fit' in terms of tendencies to extinguish when confronted with their object is inadequate. Sincere moral judgements, problematically, appear to come with both directions of fit: for any moral judgement M, we aim to judge that M if and only if M, but also tend to be motivated to act accordingly. So moral judgements appear to be both desire-like and belief-like – they seem to be 'besires'.[65] This, in itself, may be problematic: perhaps there is something dubious about the notion of a state which has both directions of fit; perhaps it commits one to denying what may perfectly well happen, viz., that moral motivation may peel apart from the descriptive content (if any) of one's moral judgements; etc.[66] I will address this difficult issue later in this thesis. For the time being, all I wish to do is to give reason to think that moral judgements are *at least* like beliefs.

Here is a first attempt. There are at least two properties of everyday, vanilla beliefs which seem to be shared by moral judgements: amenability to reason, on the one hand, and 'aiming at truth', on the other. The two are closely related: our beliefs alter in the light of fresh information precisely because they aim at the truth, where this is understood as a descriptive matter. Now, our moral judgements are generally amenable to reason – that is, fit subjects for rational debate, apt to change in response to new information or arguments, and so on. When we enter into a moral debate with someone, the content of the debate extends beyond the mere appropriateness of our responses to an action (although that may well be a corollary of the debate). The debate proceeds by adducing reasons, and we form moral beliefs in response to these arguments and sometimes even against desires to the contrary. For instance, we might strongly want to believe that philosophers of mathematics are morally inferior, perhaps because it reassures us about our own status in the world, but be compelled to revise our views by some careful argument. Now this by itself does not show that moral judgement is a matter of belief formation; our desires, and emotions more generally, also admit of rational interrelation, assessment, and alteration. And it has been suggested that there is a distinct element of our psychology which has evolved to provide a means of systematic re-ordering of desires, broadly construed. On Gibbard's account, for instance, normative language is processed by a module which provides a means of adjusting and coordinating our attitudes; consequently, sincere moral judgement is a matter of adopting certain norms.[67] But there is a more telling, related point to be made here: we have reason to think of our moral judgements as beliefs in virtue of the fact that one of the norms governing our moral judgements is that of having 'mind-to-world' direction of fit: we ought to hold a moral judgement if and only if the

[65] See Altham 1986: 284.
[66] Ibid. See also Lewis 1988.
[67] E.g. Gibbard 1990: 56-64.

world is as the moral judgement represents it as being. That is, it is a corollary of the correspondence schema that we apply the same norm to moral judgements as we do to beliefs – namely, that of having the direction of fit cited. As has already been mentioned, this is not a matter of making a descriptive or empirical claim; it is not as if we had examined the brains of agents engaged in moral discourse and discovered that moral judgements happen to turn out to be beliefs. Insofar as it is a platitude of moral discourse that moral judgements aim at the truth, we should treat moral judgements as being (at least) beliefs.

The expressivist, in response, will agree that moral judgements aim at the truth (understood as a metaphysically innocent notion), represent truth-apt contents (given minimalism about truth), and hence may qualify as beliefs. This is worrisome, as it threatens to collapse the distinction between cognitivism and non-cognitivism. But we can recast the debate in terms which remove this difficulty. Let us grant that any state of mind which is capable of truth or falsity counts (in that respect) as a belief.[68] What is at stake, here, is not the terminological issue (whether to call moral judgements 'beliefs' or not), but a more substantive one, viz., whether the function of moral judgement is primarily to describe the world or to express (and co-ordinate, etc.) attitudes. And as argued earlier, the direction-of-fit distinction, partly in virtue of its relation to the correspondence schema, captures this rather neatly. So if attributions of moral judgements to agents carry with them the implication of mind-to-world direction of fit, as I have suggested they do, then *that* creates a moderate presumption against expressivist accounts, and hence, in the context of the arguments of this chapter, in favour of realism.

7. Intuitions

The fourth argument for realism centres around the claim that people's intuitions are metaphysically loaded. This, obviously, needs some clarification. The claim concerns common pre-philosophical intuitions, at least in the first instance. It is a general claim, and therefore not rebutted by the observation that there are some whose intuitions run in the opposite direction. It is also (clearly) empirical.

There are, in fact, two distinct claims here. The first is that most people, at least pre-philosophically, are naive realists of some form or another, and that realist commitments can

[68] There are some problem cases here, such as suspicions, or fears. I may fear that such-and-such is the case, and my fear may turn out to be true. But insofar as this is coherent, it shows that 'fearing that . . .' is a state with a belief-like component, and not that fears simply *are* beliefs.

be elucidated from them by appropriate questioning.[69] That is, they already believe that their moral judgements, where true, are true in virtue of correspondence to some moral fact; that the demands of morality bind universally and categorically, and so on. The second claim is that moral discourse is informed by broadly realist presuppositions, and has been covered in the preceding sections.

Before expanding on these claims, however, I want to defuse one immediate objection. The objection runs as follows: absent some reason to think that people's pre-philosophical intuitions are correct, we do not have reason to cleave to a theory which accommodates these intuitions. We know that people's intuitions about, for instance, the rate at which objects of different weights accelerate when dropped, are – or at least have been in the past – systematically flawed. So we need more than the assumption that people's moral intuitions are metaphysically loaded; we also need reason to take this metaphysical loading seriously.

Again, I should reiterate that the argument on offer is only a weak presumptive argument. The underlying thought is that the assumption of moral realism can explain various phenomena more naturally than the alternatives – consequently, it should be the default position. The general argument, then, has been along the following lines:

a. Our moral discourse has certain appearances (truth-aptness; descriptivism; partial verity; etc).
b. These appearances can best be saved by metaethical realism.
c. Failure to save these appearances entails an error theory.
d. There is a *prima facie* reason to avoid error theory (note that this is not a practical reason: it is not that adopting an error theory would have negative consequences).

(b) is, of course, problematic; expressivists generally think that they can save most, or all, of the appearances. I take Blackburn's quasi-realism to be an instance of an expressivism which attempts to save all of the appearances, and Gibbard's norm-expressivism to be an instance of an expressivism which attempts to save slightly less than all. Now the instance cited in the last case was a semantic one; I will discuss some irrealist attempts to accommodate these cases in due course.

[69] I take Divine Command Theory to be realist in terms of broad commitments, at least for present purposes.

The thought currently on offer, however, runs as follows. Most people, including the vast number of people with religious convictions, are realists about morality, just as they are realists about everyday objects.[70] They are realists in the sense that they think moral judgements have truth-makers 'out there in the world', and are truth-apt, etc. This establishes a presumptive case for moral realism (again, Divine Command Theory counts as a form of moral realism). Now this is not intended to establish a strong case for realism, but it *is* intended to shift the burden of proof onto the irrealist. That is, we replace 'certain appearances' in (a) above with 'appearances of realism', such that:

a*: Our moral discourse is informed by assumptions of realism (we take moral judgements to be descriptive of some distinctive subject matter, assume that their truth or falsity doesn't depend on our whims, etc.)

b*: These assumptions can best be saved by realism.

c*: Failure to save these appearances entails an error theory

d*: There is a *prima facie* reason to avoid error theory.

Now if a hermeneutic expressivism such as Blackburn's is correct, these assumptions are not, in fact, assumptions of realism. As far as folk metaethics is concerned, Blackburn takes folk intuitions to be neither determinately realist nor determinately quasi-realist in nature – since all talk of 'moral facts', 'being determined by the eternal moral truth', and so on, can be given an expressivist reading.

8. Explanation

There is a stock argument for realism – closely related to that given in Section 6 – which claims that we should believe that there are real moral properties, because of the role that these properties play in certain kinds of explanation. This debate is particularly important: it is not only that the ability of moral properties to feature in certain explanations gives us reason to be moral realists, but that this is a necessary condition for moral realism to be even moderately plausible. This is so, because the realist holds that some of our moral beliefs are true. If so, then they can either be true accidentally, or non-accidentally. If the former, then it would be a miracle were *any* of our moral beliefs to be true. Worse, the realist would have no reason to think that any of her moral beliefs were true (in the absence of any good reason to

[70] With some notable exceptions: first-year undergraduate philosophy students, for instance, often lean strongly towards some form of moral subjectivism, relativism, or nihilism. But I take it that these are deviant cases: for instance, first-year undergraduate philosophy students also display a greater than average inclination to doubt the existence of tables, chairs, and other medium-sized dry goods.

think that a miracle has occurred, we should presume that it has not). Hence the realist must suppose that some of our moral beliefs are non-accidentally true, and that entails that the moral facts must be able to play a part in the explanation of these beliefs.

As mentioned above, the suggestion is that moral facts are sometimes involved in perfectly good explanations of non-moral occurrences. We might explain the cessation of slavery by reference to its moral wrongness, the people's dislike of the tyrant by reference to his evil nature, and so on. [71] This is uncontroversial; what *is* controversial is what follows from this. There are at least two problems with the attempt to move from moral explanation to moral realism. The first is that the explanation in terms of some moral property may be rephrased in terms of non-moral properties.[72] We could rephrase the earlier explanations as follows: the downfall of slavery was due to the changing intellectual, political, and economic trends of the time, and perhaps to changing beliefs about the nature of other persons; the dislike of the tyrant was due to his infliction of harm on the populace coupled with the absence of facts which might have placated the people, etc. Given this, it is not clear that the moral properties being invoked are doing any real explanatory work. The second objection is that these properties possess a very narrow 'cosmological role'.[73] That is, the explanations in which moral properties feature deal with agents' psychological states, or only explain occurrences *via* agents' psychological states.

There are at least two kinds of explanation which trade in moral terms; those which explain our moral *beliefs* in terms of moral properties of events, actions, etc., and those which explain 'brute facts' (e.g. the downfall of slavery) in terms of moral properties. Consider a case where we see children pouring petrol onto a cat, and burning it for recreational purposes.[74] We form the belief that their action was wrong; ostensibly, the wrongness of their action explains our belief that it was wrong, just as the physicist's observation of a trail in a cloud chamber explains his belief that a proton was in the chamber. What are we to make of such explanations? The case of the proton is straightforward: had the proton not been there, the trail would not have formed. We can give a counterfactual account of the explanatory force of the physical event. In the moral case, however, things become more complicated. Suppose that the moral facts were different; what would our beliefs be in that case? This counterfactual is difficult to evaluate, for the simple reason that the moral facts (such as they are) supervene on the non-moral facts. So to imagine a situation in which the moral facts differ is to imagine

[71] See Sturgeon 1988.
[72] Harman 1997.
[73] Wright 1992: 195-197.
[74] Harman 1997.

a case in which the non-moral facts differ. The relevant counterfactual, in the burning-cat case, is:

(D) If setting fire to cats were not wrong, then we would not believe that it was.

But if burning cats for recreational purposes is morally impermissible, then the counterfactual turn out to be true, due to supervenience (if setting fire to cats was not wrong, then the non-moral facts would also be different, and hence our beliefs would be relevantly different, provided that our moral sensibilities remained constant). The moral property of wrongness does, therefore, pass a counterfactual test of explanatory relevance, and an explanation which cites it does indeed seem to be a genuine explanation.[75]

This counterfactual test, however, seems too lax. Sayre-McCord offers the following example. Suppose that your peers begin to engage in witch-theorising: they throw women into ponds and explain their failure to sink (or otherwise) in terms of their status as witches, etc. Suppose that you learn how to give these ‘explanations’; being philosophically inclined, you develop a story according to which witch-properties supervene on non-witch properties. Then witch-explanations do start to pass the counterfactual test for explanatory value: if she hadn’t been a witch, she wouldn’t have floated. But none of that commits you to believing in witches: rather, we have shown that the counterfactual test is not by itself a good indicator of explanatory value, because witch explanations, moral explanations, and straightforward physical explanations all pass the test, but differ in their explanatory worth.[76]

Returning to case of the burning cat, one might think that what best explains moral belief formation is the background moral theory with which the observer operates. So the reason we form the belief that what the children did was wrong is because we have a background moral theory which tells us that burning cats for recreational purposes is wrong. Against this, Sturgeon retorts that the physicist’s belief-formation is *also* informed by background theory: the reason why the physicist forms a belief that there was a proton in the chamber is because his theory tells him that photons cause vapour trails. Yet there still seems to be a difference between the two cases. In the moral case, it seems that the moral facts are not *required* to explain our reactions; but the physical facts are required to explain our observations of events. Moral facts are not required in the sense that we could provide an adequate explanation without them. But that won’t do: we can explain various macro-phenomena in terms of the

[75] Sturgeon 1988, esp. 249-252.
[76] Sayre-McCord 1988: 276-279.

microscopic parts involved, but that doesn't lead us to be irrealists concerning, for instance, tables, chairs, or earthquakes.

A second suggestion has already been touched on, and that is Wright's challenge concerning 'width of cosmological role' – the extent to which citing certain facts is potentially explanatory of events which do not involve our intentional attitudes. In response, I want to sketch two claims. The first is that the criteria for positing moral facts are not the same as those for positing physical facts. In the physical case, we are concerned with physical explanations, so a harsh restriction on the ability of posited facts or entities to provide physical explanations seems reasonable. In the moral case, however, we are not simply concerned with explaining physical events. That is simply not what moral practice and discourse is *for*. There are other grounds, some of which have already been covered, for positing moral facts, or properties, and which are not ruled out by the lack of width of cosmological role of these facts. However – and this is the second claim – one upshot of the substantive theory which I will develop is that moral properties *do* have a fairly wide cosmological role. The reason for this will be fairly straightforward: moral properties are non-accidentally connected (in some cases, at least) to states of affairs which are themselves causally efficacious in ways which can be generalised over in law-like ways. More specifically, part of morality (the part which I will label 'constructivist') deals, amongst other things, with co-ordinating human behaviour. In these cases, constructivist goodness is connected (non-accidentally) to successful solutions to co-ordination problems, and can therefore be used to explain the growth, success, etc., of a society, and its absence to explain other consequences. A case in point would be the moral character of slavery: failure to abide by rules which enable agents to co-ordinate their actions and attitudes will lead to certain predictable consequences, one of which is that the society will fail to function in various ways, and these failures will have far-reaching consequences. Put another way, moral goodness of a certain sort is necessary for the success of social groups, and can play a role in explaining their fate, even where there is no mediation via intentional attitudes.

9. The Amoralist

An *amoralist* is someone who is a competent user of moral language, and who makes competent moral judgements, but fails to be motivated accordingly. The thought that I now wish to pursue is that realism is uniquely well placed to take the amoralist possibility seriously. The issue is, of course, complex: there is much mileage to be had from considering the relation between morality, (normative) reasons for action, and motivation. I will turn to the issues of moral motivation and rationalism in depth later, but for the time being I wish to outline and consider two possibilities.[77]

9.1. Motivation

It certainly seems that there are people who are competent users of moral language, and who make perfectly reasonable moral judgements, but nonetheless fail to be motivated accordingly. Adina Roskies gives the example of people with a certain form of brain damage – specifically, damage to the ventro-medial prefrontal cortex (VM).[78] She uses a measure of galvanic skin response (GSR) to indicate the presence or absence of motivation, and notes that whilst VM patients retain the same moral beliefs as they held prior to the damage, they do not report any motivation, nor does GSR indicate any motivation to act in accordance with their moral judgements.[79] Furthermore, VM patients retain normal patterns of moral reasoning, often scoring in the highest category of moral reasoning allowed by the tests.

The worry is straightforward. If moral judgements are simply expressions of attitudes, then the amoralist cannot be making genuine moral judgments. They must, rather, be making 'inverted commas' moral judgements; that is, making claims about how others would judge a certain action. If, on the other hand, moral judgements are, as they seem to be, expressions of belief, then the possibility that moral judgement and motivation may peel apart is left open. That, at any rate, is the beginnings of an argument against expressivism, and for cognitivism.[80]

The most straightforward response, for the expressivist, is to deny that amoralists exist. For the expressivist, the existence of an appropriate attitude is a necessary condition of genuine moral judgement, so, *ex hypothesi*, those who appear to make engage in moral discourse but lack the corresponding motivations are schmoralising (or moralising in an 'inverted commas' sense) rather than moralising. Now it is not clear how to adjudicate over this issue in a non-question-begging fashion. One line of argument runs as follows: whatever else putative amoralists do, they certainly *appear* to make genuine moral judgements, and realism permits us to save this appearance. This response is weak for two reasons: it is far from clear that saving this particular appearance is a theoretical merit, and, secondly, the expressivist's claim

[77] I take 'rationalism' to be the claim that, if I ought morally to φ, then I have (normative) reason to φ.
[78] Roskies 2003.
[79] GSR is a measure of arousal; particularly, it increases when perspiration increases. It is reliably correlated with the presence of motivation in normal subjects.
[80] Note that even if we take Roskies' VM patients to be making genuine moral judgements, this does not preclude the possibility of adopting some other cognitivist metaethical theory – such as constructivism.

is precisely that the amoralist's behaviour, to someone who has a full understanding of the fine detail of moral practice, lacks one very important appearance of moral discourse - namely its connection with motivation.[81]

9.2. Brink and Rationalism

Brink raises a related challenge.[82] His argument, roughly, is as follows. If we suppose that the concept of a moral consideration is such that moral considerations provide reasons for action, then we cannot regard as coherent the amoralist (henceforth 'sceptical amoralist') who asks 'why should I care about moral considerations?'. Similarly with the demand for a justification of morality itself: 'why be good?'. Since such questions are good questions – or at least have a reputable philosophical pedigree – we should adopt a position which allows us to make sense of them as such. [83] As Brink puts it, '[t]aking the amoralist challenge seriously . . . commits us to [reasons-]externalism; if the rationality of moral considerations can be justified, *it is not merely in virtue of the concept of morality*'.[84] Externalist moral realism (and externalist forms of cognitivism more generally) can do this; that is, they can attempt to give a serious answer to the question 'why be moral?'.

However, it is not clear that taking amoralism seriously does commit us to externalism about motivating reasons. Internalists can, perhaps, also make sense of the amoralist challenge: for the internalist, the question 'why be moral?' is just the question 'why make moral judgements?'[85]. But the two questions are *not* the same; there is, as I argue in subsequent chapters, a gap between making moral judgements, and seeing the point of moral practice. In certain cases (in particular, the case of the actual world), it is possible to become a competent user of moral discourse who makes moral judgements without acquiring the relevant motivations, and thus without being accordingly motivated. But a thoroughgoing internalist – one who thinks that there is, for all individual moral judgements, an internal connection between judgement and motivation – cannot make sense of such a position. One might, however, posit a weaker form of internalism according to which the connection is a necessary but general connection. I discuss such 'weak' versions of internalism in Chapter Two, Section

[81] C.f. Gibbard 2003. Gibbard thinks that the possibility of 'hyperexternalism' – of moral judgement detaching entirely from motivation – is implausible, and motivates expressivism.

[82] Brink 1989: 32 '. . . no one thinks that merely believing or judging that one has a moral obligation to do x gives one reason to do x: one's moral belief or judgement may be wrong. And bad moral beliefs or judgements do not provide good reasons to act.'

[83] It is possible that the question "why be moral" is, in some sense, not a good question (perhaps it is one which admits of no good answer). But the question is at least a coherent one.

[84] Brink 1989: 33.

[85] Sinclair (discussion).

3.2; the treatment of the issue at this stage will therefore be brief. But if we are to admit the possibility of genuine amoralists, then we must allow that moral judgements themselves do not necessarily involve pro- or con- attitudes. Minimally, this entails that some purely descriptive states amount to genuine moral judgements. This being so, it becomes puzzling why *all* genuine moral judgements might not be purely descriptive states.

In any case, it seems that there is something of a dilemma lurking in the background. It runs as follows. There is reason to take the possibility of amoralists seriously. On the other hand, there is a worry about what Gibbard terms 'hyperexternalism'. Consider a set of possible worlds whose inhabitants appear to use moral language. In some of these worlds, people reliably pursue what they term 'bad' and avoid what they term 'good', praise what is 'bad' and condemn what is 'good', and so on. The natural thought is that they have confused the meanings of the terms 'bad' and 'good': their use of the term 'bad' corresponds to our use of the term 'good', and *vice versa*.[86] In another world, people use moral language but are never motivated accordingly; they are 'moral scientists'. Given that the connection with motivation impinges on the natural translation of the terms, should we think that the world of moral scientists is one where the inhabitants are competent users of moral language? The intuitive answer is 'no': it is an essential feature of morality that there is some connection with motivation. But this something which realism seems to omit; if we allow that morality and motivation can peel apart such that amoralists are possible, why not to the extent that a *world* of amoralists is possible?

Consequently, we are faced with the following problem: there seems to be some kind of conceptual connection between moral discourse and motivation, but whichever account we choose to adopt should avoid making the connection too tight. One of the standard arguments for expressivism is that it is uniquely well placed to account for the existence of a connection between discourse and motivation connection. The moral realist is faced with the problem that if moral judgements are beliefs, and beliefs are not intrinsically motivating, *and* there are no necessary connections between 'distinct existences' (here, beliefs and desires), then it is difficult to see how there could be a connection of the required sort between judging that an act has a certain moral status, and being motivated accordingly. If, on the other hand, moral judgements are expressions of attitudes, then it is easy to see how a connection with motivation should obtain – although it is less clear what we are to make of putative amoralists. Expressivists in general (even without thoroughgoing internalism) have a problem with the existence of genuine amoralism. If the function of moral discourse is to express and

[86] A similar example appears in Hare 1991, and is extended in Lenman 1999.

coordinate attitudes, it then becomes difficult to see how to make sense of the acceptance of the *truth* of a moral statement in the absence of any accompanying motivation. So although Brink's argument is inconclusive, I suggest that it is not obvious how a non-cognitivist about morality can give an principled account according to which there exists a necessary but defeasible connection between moral judgement and motivation, and this, in turn, makes it hard for them to make sense of amoralists, or to take the amoralist challenge seriously.

10. Summary

The overall structure of my argument, then, is as follows: there are various ways in which morality appears to be, and many of these appearances are realist. They are realist in two senses: they are realist in the sense that folk metaethics embeds realist assumptions, but – more importantly – they are realist in that they are most naturally explained by the assumption of metaethical realism. This constitutes a weak presumptive argument for moral realism. I adduced considerations of phenomenology (our moral experience strikes us as being a certain way, and a subset of our moral practice is apt for a perceptual analogy); semantics (moral discourse is realist in structure), psychology (sincere moral judgements behave, and are treated, as expression of beliefs, rather than attitudes), intuitions (our moral intuitions are metaphysically loaded), explanation (moral explanations provide weak motivation for moral realism, and do not in any case lead to irrealism) and amoralism (there appear to be amoralists, and this generates certain sceptical worries; in order to take these worries seriously, we need to be moral realists).

None of these considerations is intended to be conclusive; what they are intended to do, however, is to create a presumptive argument for moral realism, and to shift the burden of proof onto the anti-realist, of whichever stripe. If what I have said is correct, ethical anti-realism entails taking an error-theoretic stance towards a number of significant elements of moral theory. My argument is intended to motivate the following claim: to deny realism is to commit oneself to claiming that these appearances are misleading, and that is a theoretical demerit.

CHAPTER TWO

Moral Antirealism

1. Introduction

There is a number of forceful arguments against realism; the denial of various elements of moral realism yields various forms of moral antirealism. However, I will suggest that the success of these arguments is limited. This chapter aims to critically examine the reasons which one might have for denying moral realism (which, as matters stand, I take to be the default position). Some of these reasons will turn out to be conclusive, others not. This provides the groundwork for the hybrid theory under development: where the antirealist arguments fail, we should cleave to moral realism, and where they succeed, we need to offer a separate theory. Since there is a part of ethics for which the antirealist arguments succeed, and a part for which they do not, there is a partition within ethics at the metaethical level. Crucially, there is a pattern to the success or failure of the various antirealist arguments: where they succeed, they succeed together, and where they fail, they fail together. The resultant division forms the central element of my thesis.

An analysis of the various antirealist arguments will also tell us something about the *kind* of moral realism which is plausible, and about the variety of antirealism which is required. Different antirealist arguments attack different elements of moral realism, and the antirealism which we should adopt results from the rejection of those elements which need to be rejected. Since there is (as I argued in the last chapter) a moderate presumption in favour of moral realism, I aim to retain as many of the salient features of realism as possible. For some areas of ethics, I claim that this can be done wholesale; we can be straightforward moral realists, at least in part. For other areas of ethics, we should adopt the form of antirealism which differs minimally from moral realism.[1] In subsequent chapters, I will aim to develop the realist and antirealist theories respectively; this chapter is concerned primarily with motivating the denial of realism in parts, whilst arguing that a suitably developed hybrid theory can accommodate key antirealist insights without ceding too much ground. The hybrid theorist, that is, can have her cake and eat it.

I begin by critically discussing two distinct theories which hold that moral discourse is systematically flawed: the error theories of John Mackie and Richard Joyce. Each holds that

[1] That is, in as few respects as possible.

moral discourse is assertoric, expresses truth-apt propositions, purports to describe mind-independent facts, etc. – but that all of these propositions turn out to be, literally, false. In Mackie's case, this is because moral discourse is committed to the existence of non-existent moral properties; for Joyce, this is because moral discourse is committed to strong categorical imperatives, which cannot be made to answer to anything in the world. I suggest that restricting moral realism to claims about certain phenomenal states (along hedonistic utilitarian lines) can defuse these objections. I go on to discuss the suggestion that morality cannot deal in unknowable truths, and therefore must be mind-dependent, before turning to a discussion of the expressivist challenge. Expressivism is shown to have at least two underlying motivations – one ontological, and one concerning motivation. I argue that the connection between moral judgement and motivation is best conceived of as a form of weak social internalism, according to which it is necessarily the case that members of a moral community are generally motivated in accordance with their sincere moral judgements. This reduces the strength of the motivation motivation. But again, for each of these worries, restricting moral realism to a broadly utilitarian conception defangs the objection. This leaves open the question of what we are to make of claims about the moral status of lying, breaking promises, and so on – the 'remainder' of morality. I will go on to explore the relevant utilitarian conception in detail in Chapter Three, and will return to the question of the remainder of morality in Chapter Four.

2. Antirealism: Motivations and Varieties

There exist many different versions of moral antirealism, each arrived at by denying one, or several, of the claims of realism. The denial of the partial verity condition yields error theory; the denial of descriptivism yields expressivism; the denial of cognitivism combined with the acceptance of representationalism yields fictionalism; and so on, and so forth. Each form of antirealism comes with its own distinctive motivation: in the case of error theorists, the thought is that moral discourse comes with some unfulfilled (or unfulfillable) commitment to certain entitities or properties; expressivists tend to hold that expressivism best explains the supposedly internal connection between moral judgement and motivation; fictionalists share the expressivists' worry that the purpose of moral discourse is to coordinate actions and attitudes, but also wish to respect the surface syntax of moral discourse. And, of course, these distinctions are neither neat nor tidy: Joyce, for instance, combines an error theory about moral discourse with a fictionalist proposal.[2] But I am primarily interested in the motivations

[2] Joyce 2001.

behind the various antirealisms, rather than the taxonomical project – so it is to these that I now turn.

2.1. Error Theory: Mackie

Error theory accepts the (realist) claims that moral discourse is truth-apt, expresses beliefs which purport to represent mind-independent facts, and so on, but denies the 'partial verity' condition. Moral discourse and practice, on this model, is systematically flawed: it consists of statements which either are not, or cannot be, true. This systematic falsity is to be explained in terms of the truthmakers for moral claims, which, it turns out, fail to exist (or obtain). Contemporary philosophers, for instance, are error theorists about witches (the non-existence of witches means that all of our claims about witches turn out to be false) and about phlogiston (for the same reason).[3] We might adopt the same view as regards morality.

The canonical example of this is given by Mackie, in his *Ethics: Inventing Right and Wrong*.[4] Mackie claims that ordinary moral discourse contains metaphysical commitments, specifically to metaphysically 'queer', non-natural, ethical properties, and that there is overwhelming reason to doubt the existence of such properties. Hence we should conclude that moral discourse is systematically in error; the truth-conditions for moral statements simply do not obtain. Mackie offers two arguments for doubting the existence of ethical properties, the first being the argument from 'queerness', the second being the argument from relativity. I will deal with these in reverse order.

2.1.1. The argument from relativity

The argument from relativity runs as follows: different cultures have different fundamental ethical beliefs, and their disagreements over these fundamental matters are persistent, genuine, and (apparently) intractable. The best explanation of such disagreement, it is then suggested, is that there is no right answer available. In contrast, scientific disagreements are held to be amenable to argument and discussion, and scientific judgements tend to converge over time. That indicates that in science there is some fact of the matter which is being discussed: the truth of scientific claims explains our belief in them (we believe that the Earth goes round the Sun, for instance, because it does). This marks a discontinuity between science

[3] All positive atomic claims about witches turn out to be false, but complex claims involving witch-discourse may be true ('*either* she is a witch *or* the earth goes round the Sun'), as do the negations of positive atomic claims about witches ('she is not a witch'). But this is a merely technical issue, and one which I will ignore.

[4] Mackie 1990: 36-38.

and ethics: whereas in science we explain convergence by citing the truthmakers for scientific claims, in ethics a lack of convergence points towards a lack of truthmakers (because, were there any such truthmakers, we would expect convergence).[5]

Now it is not clear that the mere fact of disagreement – the cultural or ethical relativity which Mackie cites – is enough to support a moral error theory. After all, there is widespread, and often persistent, disagreement over straightforward matters of fact, but we do not take this to be evidence that there is no fact of the matter as to, for instance, whether the Earth orbits the Sun, or whether evolution is a gradual or episodic process. More needs to be added in order to derive the required conclusion. There are two (related) key notions which need to be made explicit, namely the *intractability* of moral disagreement, and the *lack of convergence*. How these are related depends on how 'intractability' is understood: if it is taken to describe the persistence of the disagreement, then it is the same as the lack of convergence. However, 'intractable' could also be taken to indicate an explanation for the lack of convergence: there is persistent ethical disagreement, and hence a lack of convergence in people's ethical beliefs, because there is nothing to guide our ethical judgements to agreement. The thought that there is nothing to guide our ethical judgements to agreement contains at least two distinct elements: firstly, there is the claim that the absence of moral properties explains the persistence of disagreement; secondly, there is the claim that there is no mechanism by which we could arrive at an agreement. The first claim explains the second only if we assume that if there were moral properties, there would be some mechanism whereby we could arrive at an agreement. This requires the (reasonable) assumption that were moral properties to exist, they would not be unknowable.[6]

How would moral truths lead to convergence in moral beliefs? There are at least two routes by which this could occur. The first case is one where moral truths are knowable by some generally shared faculty, either rational (where moral truths are readily deducible from some generally shared material), or non-rational (where moral truths are perceived by some faculty of moral intuition, for instance). The second involves morality being amenable to rational discussion in such a way that rational debate, in the absence of confounding factors, tends to lead to convergence. Of course, the two are not mutually exclusive: rational moral debate may

[5] It is unclear whether we explain convergence by citing the truth of the claims, or the truthmaker for the claim. But presumably the thought is that there is some way the world is which explains the convergence; some out-there-in-the-world fact of the matter. Hence the truthmaker interpretation is the natural one.

[6] C.f. Mackie 1990: 38. Mackie suggests that, were moral properties to exist, the faculty whereby we would gain knowledge of these properties would itself be 'queer'.

lead to convergence if the truths under consideration are self-evident, i.e. if merely comprehending those truths is sufficient to justify believing those propositions.

We can, therefore, extract the following argument from Mackie.

(1) There is persistent, widespread ethical disagreement.
(2) Were there moral truths, they would be either readily knowable by some shared faculty, or amenable to rational investigation.
(3) If these truths were either readily knowable by some shared faculty, or amenable to rational investigation, they would be generally known.
(4) Hence there would not be persistent, widespread ethical disagreement.
(5) Hence there are no moral truths.

We could also recast this as an argument to the best explanation, replacing (2) – (4) with '. . . the best explanation of persistent, widespread ethical disagreement requires the absence of moral truths.' But even as an argument to the best explanation, it fails. Each of the premises in the original formulation is contentious. It is far from clear that fundamental ethical disagreement is as widespread as Mackie seems to think - although this is an empirical issue, and one which I will turn to in due course. Worse, even were moral truths amenable to rational investigation, this rational investigation is not obviously an easy process. People are often systematically irrational; even in cases where the truth is determinate and often very much accessible, agreement is not always forthcoming. There exist, for instance, persistent and widespread scientific disagreements even where there is a straightforward – and apparently obvious – right answer. In many cases, people simply lack the cognitive capacities to resolve disputes, at least once those disputes reach a sufficient level of depth and complexity. Worse still, it seems clear that even when procedures for rational resolution of disagreements are available, people do not always choose to take them up.

It should be evident, in any case, that the third premise is likely to be false: the existence of moral truth, even if amenable to rational investigation and discussion, need not guarantee convergence of moral beliefs. However, I think that there is a better way of dealing with the existence of relativity, understood as the persistent lack of ethical convergence over time, which need not commit us to this line of argument across the board. It is – as might be expected – a hybrid approach. That is, disagreement within different areas of ethics requires a different treatment. Rather than argue for it directly here, I am simply going to sketch the account. The core claim is that there is, in some cases, a single correct answer, and that this single correct answer is appreciable along a roughly perceptual model, as well as by rational argument. In other cases, there is a plurality of correct answers, and whereas there are decision procedures available for arriving at these correct answers, one answer need not be decisively more choiceworthy than another. So there is, then, scope for faultless disagreement, and also an expectation of lack of convergence *even where the decision procedure has been widely understood*, because the direction which the procedure will take will be influenced by context. Given variation in context, we can expect variation in decisions. Furthermore, the rational resolution for ethical debates is fairly complex, but people are cognitively limited. We should expect that ethical debates are difficult to resolve for at least three reasons: firstly, the underlying metaethical account is itself complex, and open to rational debate; secondly, the factors relevant for consideration are legion; and thirdly, there *are* cases where there exists a plurality of equally good answers.

Now this might look as if I am claiming that we should expect a lack of convergence because the truth is generally too difficult for people to grasp. There is something rather curious about this claim, however; if the subject matter is too difficult for people to grasp, then that provides mitigating grounds for moral failings. But we generally take people to be fully responsible (and hence blameworthy) for their moral shortcomings. There are two points to be made here: the first is that this is not a matter of attributing to them some epistemic shortcoming, and the second is that culpability implies that they could (realistically) have known better. In the case of children, for instance, we often say that they ‘couldn’t have known any better’. This is not to say that there is no possible world in which the child *does* know better; it is metaphysically possible that they could have known better, but we don’t have any reasonable expectation of them to do so. It may, of course, be that our practices are fundamentally misguided in this respect, but it would at least be counterintuitive to suggest that most people are exculpated by their limited access to the moral facts. As regards the first point, it is evident that when someone acts on erroneous moral beliefs we criticise them for a *moral*, rather than an epistemic, shortcoming.[7] The two may be connected: often, this failing is due to negligence, where, for instance, we criticise people for failing to think adequately about the consequences of their actions. But this, again, is primarily a moral failing; it is incumbent on people to consider their own moral beliefs, partly because the stakes are often high; faulty moral beliefs will lead to bad consequences. Were moral beliefs to be of no consequence, it would be hard to see how failing to form correct moral beliefs would matter.

For these reasons, I think that any acceptable account of lack of convergence relates, in part, to the non-moral factors which influence moral belief. That is, it is evident that there are

[7] Of course the two are not mutually exclusive; but the focus of the criticism is on the *moral* failing, rather than the epistemic one.

numerous ways in which people acquire moral beliefs; by education, imitation, through literature, stories, and so on. People are also influenced by self-interest, such that many of our beliefs are self-serving. For instance, the average person believes themselves to be possessed of an above-average sense of humour, and to be a better-than-average driver. Even where our beliefs track the truth, they do not do so perfectly, and where the stakes are fairly high, factors other than the truth may exert excessive influence. In the moral case, it seems evident that many of our moral beliefs – and sometimes also our metaethical beliefs – are heavily influenced by considerations of self-interest. Common-sense philosophical morality tends to hold that it is permissible for us to prioritise our own interests, life projects, and so on, even in the face of widespread and easily remediable suffering. The corollary of this is the metaethical claim that some forms of consequentialism are simply 'too demanding'. In this case, even were the truth of the matter to be that morality just is as demanding as the straightforward act-consequentialist claims – that is, even were our moral obligations to override almost all of our personal projects and interests – we would reasonably expect large dissent on this issue, even among philosophers, simply because the psychological (and personal) cost of believing this claim would be significant.[8] Likewise, I suggest, with many other moral beliefs. And there are, of course, more factors at work than mere self-interest; for instance, we readily absorb moral beliefs from our peers, regardless of their truth. This partly accounts for widespread homogeneity of moral beliefs within groups, as well as widespread heterogeneity across groups; distinct groups will harmonise amongst themselves, but not with others.[9]

However, even if we can account for the existence of widespread, persistent ethical disagreement, we might worry that the realist still has work to do. Granted, there may be many cases where this kind of disagreement results from human error, biases, paucity of information, and so on. But there also seem to be at least *some* cases where such disagreement cannot be traced back to any of the proposed confounding factors. In such cases, we cannot point to bias, rational shortcoming, etc., on behalf of any of the parties to the dispute. Such a disagreement is said to be *faultless*.

Why would faultless disagreement be a problem for the moral realist? Wright offers the following diagnosis.[10] If two representational systems are focussed on the same object, and are both functioning perfectly, then we expect the same representation. Two perfectly functioning cameras will, in identical contexts, produce identical images. But if these images

[8] Examples of such demanding consequentialist theories are given by Singer 1993, Unger 1996.
[9] See Huemer 2005 for thorough discussion on this topic, esp. pp. 136-144.
[10] Wright 1992: 91-4, 2003: 7-8.

differ, some explanation must be offered, and there are three possibilities: either at least one of the systems is malfunctioning in some way, or they are focussed on different objects, or they are not representational systems. In the case of faultless ethical disagreement, neither system is malfunctioning. And we can assume that the subject matter is the same, as otherwise there would be no disagreement. Hence the existence of faultless ethical disagreement motivates the thought that moral judgements do not function to represent the world. This is simply the denial of one of the central theses of moral realism, i.e. descriptivism.

One possible move, here, is to deny that ethical disagreement is ever faultless in this sense. Now in order to avoid being *ad hoc*, or begging the question, this cannot simply be a reassertion of moral realism. Given that faultless ethical disagreement certainly seems to exist, the burden of proof is on the realist. There are roughly two ways in which the realist can go from here: either give reason to think that moral knowledge is best thought of, at least in some cases, along perceptual lines, so that ethical disagreement in these apparently faultless cases can be traced back to some perceptual malfunctioning; or give an account of some cognitive shortcoming which explains the disagreement.

To recap: we have two related worries concerning ethical disagreement. The first – which is closest to Mackie's argument from relativity – is that the lack of convergence in ethical beliefs suggests that there are no facts to which our ethical beliefs may answer. The second is that ethical disagreement may often be faultless, in that we can point to no cognitive or informational shortcoming which explains this disagreement.[11] Now the existence of widespread, persistent disagreement should be treated with care. It need not imply faultlessness: indeed, given the human tendency towards faulty thinking, we should expect widespread, persistent disagreement even where the subject matter is to be construed along realist lines. Such disagreement indicates *either* faulty representational apparatus (taken to include the existence of external biases), *or* the existence of a plurality of correct answers, *or* the falsity of descriptivism.[12] Given that morality judgement appears to be representational, however, we need to investigate fully the first two options before settling on the third.

[11] There is a quick response to the appearance of faultless disagreement, which is to say that there are right and wrong answers in ethics, and *ipso facto* one party to the disagreement must be in error, and this error must be a distinctively moral error (since we can point to no cognitive shortcoming). But this begs the question.

[12] A fourth possibility is that apparent disagreement results from vagueness at some level – but for the purposes of the present argument, this is covered by the second possibility, viz., the existence of plural correct answers.

Granted, the existence of faultless disagreement would pose a problem for moral realism, unless there existed a plurality of right answers. But there is the following methodological problem here: the notion of 'fault' depends on one's metaethical theory. Now I do not wish to attempt to argue for the details of my metaethical theory within this chapter; those specifics will be addressed in Chapters Three and Four. The point relevant to the current discussion is that the hybrid theory, consisting of realist hedonistic utilitarianism coupled with a construction procedure which leaves room for ethical pluralism, yields the following three predictions. Firstly, we should expect widespread faulty disagreement. Secondly, we should also expect lots of faultless disagreement. Now these two predictions are not particularly interesting, but they *do* indicate that the existence of disagreement of both sorts is compatible with the theory. More interestingly, the theory also offers the following prediction: we should expect *more* convergence within the 'realist' part of ethics than within the constructivist part, all other things being equal. This is a direct consequence of the relation between representationalism and convergence discussed earlier: whereas the construction procedure may yield many different outcomes, we should expect that the (realist) ethical facts can themselves determinately guide inquiry. So just as there is greater convergence on matters of science than of etiquette, we should expect greater convergence within realist domains of discourse than within constructed domains of discourse. Of course, all things are generally not equal: different areas of ethics are subject to different influences, and there is a large number of possible confounding factors. Whether or not these predictions are borne out is beyond the scope of this thesis: furthermore, the extant literature seems inadequate to answer these questions. Although there do seem to be cases of radical ethical disagreement, and also of convergence, it is unclear to what extent the one dominates the other.[13]

2.1.2. The Argument from Queerness

There is a second thread running through error theory. Error theories make at least two key claims, to wit:

1. There is some essential commitment of moral discourse (e.g. to Platonic, non-natural truth-makers, or to categoricity, etc)
2. This essential commitment is never met; there are, for instance, no Platonic, non-natural truth makers; there is no such thing as a categorical imperative; etc.

[13] See e.g. Chagnon 2000, Shweder 2000.

So we have two distinct claims, one semantic, and one ontological. The arguments for the semantic claim may, in fact, turn out to be in the realist's favour, provided that the ontological worries can be met and rebutted. The semantic claim comes in a variety of flavours, some stronger than others. Mackie offers a strong variant, holding that moral discourse is committed to the existence of metaphysically 'queer' non-natural but intrinsically normative properties. But this, one might think, is implausible. Were it to be true that such a commitment existed, we would expect that people who do *not* believe in such properties would be guilty of some deep-seated conceptual incoherence. Making this incoherence clear would provide some pressure, at least, to see moral discourse as incoherent in this sense – or, in Mackie's terms, systematically false. But this raises the possibility of a broad objection to the semantic claim, namely that it is quite possible to reject the existence of spooky, non-natural entities (or properties) whilst retaining moral practice as it stands. Any form of analytic ethical naturalism will meet this criterion. Suppose, for instance, that our best metaethical theory held goodness to be – as a matter of definition - simply a matter of producing the best possible consequences. In that case, we need posit no spooky, non-natural entities or properties. And although analytic ethical naturalism may turn out to be false (there are, after all, many good reasons for doubting it), the point remains: we can dispense with the objectionable commitment without losing the point of our moral practice. In other words, the kinds of metaphysical commitment which Mackie presupposes are accidental to, rather than a necessary part of, moral discourse.

There is a second strand to the argument from queerness. If such metaphysically queer properties were to exist, then our epistemic access to them would have to be mediated by some equally queer faculty of moral intuition. And the suggestion of such queer faculties is, again, implausible. Now it is not clear that Mackie's epistemic claim is true: we might reasonably think that we have knowledge about (for instance) abstract objects such as numbers, in spite of the fact that such objects are very different to the concrete objects which populate the world. Nonetheless, the burden of proof would fall on the realist to provide an account of the relevant epistemology, and this would seem to be a difficult task. I will not attempt to discuss this epistemic issue in any detail, however, as the response for both strands of the 'queerness' argument is the same: namely, to deny that any commitment to queer properties is essential to moral discourse.

2.2. Error Theory: Joyce

Richard Joyce offers a position which he characterises as moral *fictionalism*, which consists of an error theory about moral discourse, coupled with the claim that we nonetheless have good reason to continue engaging in moral discourse, which in turn motivates the thought that we should continue to regard (at least some of) our moral claims as true, at least in some sense[14]. This yields a 'hermeneutic', rather than revisionary, account. His central argument in favour of error theory is, I think, more plausible than those offered by Mackie, and runs as follows.

Even if moral discourse is not committed to the existence of metaphysically queer properties, it *does* seem to be committed to categoricity.[15] Something which provides a moral reason for an agent does so irrespective of his or her interests: we do not escape the demands of morality simply by ceasing to care about it. This is evident in the grammar of moral discourse – on the surface, at least, moral claims are not hypothetical – and is supported by brief reflection on the kinds of consideration which might be held to bear on the issue of whether or not certain moral obligations exist. But this then commits us to thinking that morality provides reasons for action independently of the agent's desires, whims, conative states, etc. In Williams' terminology, these would be reasons which obtain independently of the agent's subjective motivational set – 'external' reasons.[16] And *that*, one might think, is a mistake. Perhaps talk of such reasons is conceptually incoherent, in one way or another. Williams argues that reasons must be able to explain an agent's action, and therefore can only do so if they are able to motivate the agent – that is, to be included in his 'subjective motivational set'. But a reason which bears no relation (or a merely accidental one) to the agent's subjective motivational set cannot (except accidentally) motivate him; hence it cannot (except accidentally) explain his action, and hence cannot (except accidentally) be a genuine reason. Where the agent cannot be brought to include the purported reason in his subjective motivational set by some rational process, talk of that reason is, for Williams, 'mere bluff'.[17]

This is, perhaps, a little too fast. After all, it is simply untrue that all reasons must be motivationally potent. Reasons for belief, for instance, do not depend on the agent's desires – or at least, theoretical reasons do not.[18] *Modus ponens* is compelling whether or not one is interested in thinking rationally.[19] The obvious response, here, is to specify the scope of the

[14] Joyce 2001.

[15] Mackie, as discussed above, shares this view, but holds that the putative categoricity of moral reasons derives from the existence (or otherwise) of intrinsically normative entities – and it is the intrinsically normative entities, rather than categoricity itself, which Mackie finds 'queer'.

[16] Williams 1981: 101-114.

[17] This suggestion has received considerable attention. See e.g. FitzPatrick 2004; Lillehammer 2000; McDowell 1979, 1995; Stratton-Lake 1998.

[18] There might be practical reasons for believing that P, or at least for acquiring the belief that P.

[19] This is actually a rather bad example. *Modus ponens* yields a rational requirement, rather than a reason for belief (see Broome 1999). Accepting modus ponens commits us to *either* accepting the conclusion *or* rejecting one of the premises. But it does perhaps provide the framework within which

argument: Williams is concerned with *practical* reasons – that is, reasons for action – and it is less obvious that we can generate examples of practical reasons which need not be able to motivate agents. But even practical reasons, as Christine Korsgaard has observed, need only motivate agents insofar as they are rational.[20] Theoretical reasons need only be convincing to theoretically rational agents; similarly, practical reasons need only be motivating for practically rational agents.

Similarly, we might worry that Williams fails to allow for a distinction between normative and explanatory reasons. In the case of agency, it is quite plausible to think that explanatory reasons must be able to motivate the agent, and hence must bear some relation to the agent's subjective motivational set. There are no external explanatory reasons. But it would be a mistake to think that all reasons must be explanatory. The fact that an action would cause a great deal of unwarranted suffering is reason for me not to perform that action, regardless of whether or not I share that concern.

Nonetheless, there is a serious issue here. The real force of Williams' argument, I think, lies in his suggestion that talk of external normative reasons can amount to no more than 'mere bluff'. The challenge, therefore, is this: if one thinks that external (normative) reasons exist, one is then obliged to give some account of what kind of things these are. If they are to be such that to fail to act on them is a species of practical irrationality, then an account of this must be given which is not question-begging. But it is hard to see how we can give any content at all to the notion of an external reason.

Korsgaard's response to Williams – which relies on the claim that normative reasons need motivate people only insofar as they are practically rational – fails on this count, because if practical rationality is a matter of responding appropriately to reasons, then the field of normative reasons cannot be substantively constrained by the limits of practical rationality. Either the two concepts are interdependent, in which case Korsgaard's claim is an uninformative truism, or they are not, in which case we need an account either of external reasons, or of practical rationality which does not rely on the agent's 'subjective motivational set'.

reasons for belief get transmitted; that I find the premisses of a logically valid argument convincing provides me with reason to believe the conclusion. This is true regardless of whether I am interested in logical argument.

[20] Korsgaard 1996a: 320-1.

It might seem, then, that what is needed at this point is detailed discussion of various theories of practical rationality (or practical reason). But the general challenge – that of accounting for the categorical nature of moral judgement – can be met more directly. The strategy will be as follows: grant that moral judgements (and, concomitantly, obligations, imperatives, etc.) are categorical. Grant also that this entails that true moral claims entail reasons for action which are *logically* independent of the agent's subjective motivational sets. The task then becomes that of making sense of the notion of an external reason.

2.3. The Answer in Outline

For part of morality, I suggest, we should be moral *realists*. We *can* make sense of the notion of an external reason, insofar as we can make sense of the notion of something which is intrinsically reason-providing. Williams' worry was that reasons need to be able to explain agents' behaviour: in order to do that, they need to be able to motivate agents. There are therefore two ways in which to meet Williams' worry: either show that the reasons under consideration are such as to be appropriately connected to the agents' subjective motivaitonal sets; or give sense to the notion of an external reason in a way which shows talk of external reasons to be more than 'mere bluff'. I will be taking the second approach. If we can make sense of the relevant properties as being intrinsically normative (as having 'to-be-pursuedness' built into them, for instance), then we can begin to give content to external-reasons claims in a way which meets this second criterion.

For the other part, we should be *constructivists*. Here, there are no candidates for intrinsically normative properties. Rather, the normative force of these moral considerations is derived from some other source. Scanlon, for instance, suggests that moral obligation is underwritten by our interest in behaving in ways which are justifiable to others on grounds that no-one could reasonably reject.[21] True moral claims do provide us with reasons for action, because they tell us that a certain action will satisfy this (or a similar) condition. And since it is generally true of persons living in a social context that they have such an interest, true moral claims will provide these persons with reasons for action. However, here the categorical nature of moral judgements is strictly misleading: where people lack entirely the underlying interests, they may turn out to have no reason to keep promises, refrain from telling lies, and so on. In such cases, the metaethical story needs to be further developed. In particular, it needs to be shown that the strict failure of categoricity does not result in a thoroughgoing error theory. I will leave discussion of the precise nature of constructivism to Chapter Four.

[21] Scanlon 1998.

However, there is at least one difficulty with the general strategy of grounding morality on non-moral considerations (that is, taking moral reasons for action to be ultimately dependent on non-moral reasons for action). We initially wanted to say that moral obligations provide reasons for action independently of the agent's subjective motivational set. This is a strong conceptual claim. But it seems as if I am saying that in some cases moral obligations provide reasons for action *in virtue of* the agent's subjective motivational set; thus, for instance, we ought not to lie because we have certain complex interests. In other words, it seems that making moral obligations contingent on the existence of something in the agent's subjective motivational set commits us to an unpalatable revisionism about moral justification: it forces us to offer the wrong kind of moral explanation.[22] Rather than advert to the wrongness of the deed, we turn to non-moral considerations. It is not that we ought not to lie because lying is wrong, but that we ought not to lie if and only lying is appropriately related to our current subjective motivational set. This set may contain moral motivations (desires not to lie, or to do the right thing), but the moral status of an action is not to be explained in terms of the presence (or absence) of these motivations.[23]

A first attempt at a response involves the following move: grant that the reason depends on the existence of some feature of our subjective motivational set, but deny that this feature plays any part in the full specification of the reason. That is, we may have a moral reason – for instance, not to lie – which depends on the existence of certain features of the world in order to be a reason, but where the reason itself does not involve those features. This is similar to the move made in Dancy's *Ethics Without Principles*, where he distinguishes between 'enablers' and 'favourers': 'favourers' (in the practical case) are whichever features of the world count in favour of a certain course of action, and 'enablers' (in the practical case) are those features of the world which allow the favourers to play the role that they do.[24] The fact that an action was one of promising provides a reason – a favourer – for keeping that promise; the fact that the act of promising was uncoerced enables this reason to go through (on the assumption that coerced promises are not binding). Similarly, even if the reason is dependent on some element in an agent's subjective motivational set, this element need not form part of the reason.: there will be a great number of conditions upon which the reason is

[22] The significance of the issue of revisionism about justification was brought to my attention by Hallvard Lillehammer.

[23] Barring special cases – for instance, where the desire to do 'the wrong thing', or to harm another person, contributes to the moral badness of an action because (suppose) it is true that malign intentions make bad actions worse. I am assuming here that, in general, we do not take the moral status of an action to be adequately explained by reference to the agent's subjective motivational set. The relation between the action and the *agent's* moral status is a different matter.

[24] Dancy 2004: 38-45.

dependent, but a full specification of the reason will not include these conditions. These conditions are, in Dancy's terminology, 'enablers', rather than 'favourers'. To take another example: it is (arguably) a necessary condition of my having an obligation (and hence a reason) to perform a certain action that I am able to perform this action (as 'ought' implies 'can').[25] Perhaps I am under an obligation to be under the clock tower at noon because I promised that I would do so. But what the full specification of the reason refers to the obligation – not the obligation coupled with the fact that I *can* be under the clock tower at noon.

However, even where the reason can be specified in a way which does not mention the agent's subjective motivational set – that is, even where we distinguish between enablers and favourers in such a way as to avoid explicitly specifying an element of the subjective motivational set – we are still committed to the denial of categoricity at the conceptual level, since the reason does ultimately depend on the agent's subjective motivational set. Any theory which gives only a non-moral answer to the question 'why be moral?' commits itself to the claim that certain features of the world provide reasons for action partly in virtue of their relation to the agent's motivations. Likewise with any theory which attempts to give a non-moral analysis of moral reasons for action. Constructivism is one such theory. On such an account, moral reasons for action are not (strictly speaking) categorical, since they are dependent on what actual moral agents care about. The task is to make this claim out to be plausible. This, I think, can be done, provided that we draw a distinction within ethics between those moral reasons which are categorical, in some sense or other, and those which are not. It is to this distinction that I now turn.

2.4. Categoricity and Self-Interest

Although it is generally thought that morality is binding categorically, it is worth noticing that the intuition of categoricity varies across cases. For instance, it seems highly plausible that reasons not to inflict wanton suffering are binding on everyone. On the other hand, it is less clear that reasons to refrain from promise-breaking are binding for everyone. In particular, there are certain cases within ethics where we *do* appeal to self-interest in order to explain the force of a consideration. One example is that of wanton lying: here, it *does* seem that the

[25] A possible counterexample is where I enter into an obligation to φ at a certain time, but subsequently do something so as to render myself unable to φ. One might think that the obligation persists despite my later actions – if I borrow money from someone, I ought to repay it regardless of whether I have spent it. But we can, I think, understand the structure of the relevant obligations without violating the '"ought" implies "can"' principle. The issue is in any case tangential to my argument, so I will leave it to one side.

prohibition on wanton lying aligns neatly with enlightened self-interest. ‘Honesty is the best policy’, as schoolchildren everywhere are told. In a similar vein, our self-interest is often best served by being honest, kind, courageous, and so on. Doubtless, this has much to do with the degree to which virtuous behaviour tends to be praised and rewarded, and vicious behaviour blamed and punished. But given the extent to which moral behaviour and self-interest align, it would be surprising if there were no connection between the two. Given this, there is scope for thinking that many moral reasons for action *are* contingent on our self-interest; and hence on our subjective motivational set, given that agents are generally inclined to promote their own self-interest.[26]

It is commonplace, then, that people advert to considerations of self-interest when attempting to justify certain – although not all – moral claims. This does not entail that the truth of these moral claims depends on self-interest, of course. It might be that such appeals to enlightened self-interest are either ‘mere bluff’: perhaps honesty is not, in fact, the best way to maximise self-interest, but is portrayed as such for expedience. Alternatively, these appeals to self-interest might be attempts to persuade people to generate the desired behaviours, even in the absence of the appropriate (moral) motivations. Once the behaviours have been established, the agent may then come to see the point of the practice – but an essential first step is that they form the correct habits.[27]

But the connection between morality and self-interest goes further than mere coincidence. Consider again the case of wanton lying: not only can we convince others that truth-telling is in his or her best interests, but we can do so by rational means. The fact that agents can come to take the prohibition on wanton lying seriously, and to do so by rational processes – in ways driven by considerations of self-interest, or by other extant interests – suggests very strongly that the prohibition derives at least some of its force from these considerations. This is, of course, because honesty generally is (in our present circumstances) the best policy: it is generally prudent to speak honestly. The reason why these appeals to self-interest succeed, therefore, is because certain moral rules genuinely do promote self-interest.[28] And the suggestion on offer, here, is that we can fully explain the normative force of certain moral rules – the prohibition on wanton lying being a case in point – by such considerations.

[26] One might think that there is a necessary connection between self-interest and an agent’s subjective motivational set – or between self-interest and what the agent in question could come to care about. But there is then a problem with agents who obstinately do not care about their own self-interest, perhaps due to depression or simply some mental peculiarity. I do not need to address this problem here, however.

[27] C.f. Aristotle 1925: 28-29.

[28] Or, if not rules, then virtues. The thought that there is such a connection between morality and self-interest has a long history. See Aristotle 1925 esp. 4-7, 17-19, 154-158.

One might worry that this conflates a metaphysical issue ('what makes these claims true?) with an epistemic or psychological issue ('why do we take these claims seriously?'). Perhaps the wrongness of lying is one of the eternal verities of the universe; it is prudent to be honest; what drives us to be honest is prudence; but, nevertheless, the importance of honesty does not depend on its prudential value. But I am not claiming that all moral truths need reduce to self-interest. I am merely claiming that it is plausible to treat the reasons we have for taking (some) moral claims seriously as derivative from self-interest. In that sense, morality is grounded in self-interest. This is not to say that moral claims are true purely because they serve our self-interest; rather, the constructivist holds that we have reason to adopt a moral practice because doing so serves our self-interest, but that the truth-conditions for individual moral claims need not refer to our self-interest. This claim will be unpacked in Chapter Four.

A further worry is that this suggestion fails to acknowledge a distinction between moral considerations, and considerations of self-interest. The two seem to peel apart quite simply; it seems that it can often not be in our self-interest to act morally, and that morality acts as a check on agents' unrestricted pursuit of their own interests, desires, and so on. Again, I am not claiming that all moral behaviour is self-interested; nor, indeed, that the dictates of constructivist morality align perfectly with the dictates of self-interest. Rather, I claim that it is (generally) in our self-interest to act morally; but since the *truth-conditions* for constructivist morality do not specify individual self-interest, the dictates of morality and self-interest can peel apart as required.

However, I do still need to explain the sense in which these imperatives are *moral* considerations. One obvious marker of moral discourse is the appearance of categoricity. Prudential imperatives, in contrast, lack this appearance: they are hypothetical. This appearance of categoricity, however, is problematic, since it seems to commit us to the existence of so-called external reasons – unless we hold, with Kant, that morality is binding on all rational agents as such in virtue of their rationality, which itself is motivationally potent. But this is only problematic insofar as the non-obtaining of external reasons commits us to an error theory. Consider, for the purposes of contrast, the claim that:

L: One ought not to tell lies.

It is then pointed out that there are many circumstances in which one *ought* to tell a lie; when protecting one's children from an axe-murderer, when concealing Jews from pursuant Nazis, or when confronted with a demand for honest sartorial appraisal from one's spouse. Does this falsify the claim? No: rather, it pushes us to read the claim as elliptical for a defeasible generalisation, such as:

L*: In most cases, one ought not to tell lies,

Or perhaps a more opaque claim, such as:

L:** That an action would involve lying counts against engaging in that action.

The *appearance* of categoricity, then, is problematic insofar as it is warranted or unwarranted, rather than strictly true or false (since in cases such as **L**, the appearance is strictly misleading, but we take this to be innocuous, as the appearance is nonetheless warranted). In the case of constructivist morality, I will claim that this appearance of categoricity is entirely warranted, in spite of the fact that there are those for whom constructed moral principles provide no reasons for action.

For most social, interdependent persons (such as ourselves), it is easy to see how a principle which is, on the surface, categorical, might nonetheless be grounded in prudential (and therefore hypothetical) imperatives. The key point, here, is that prudential considerations give us reason to adopt certain moral principles, and to take these principles seriously. In other words, prudential considerations often operate at the level of policy. Acknowledging that we have a reason to adopt a policy of not lying for self-interest, for instance, may lead to adopting this policy. This, in turn, involves reconfiguring the reasons that we take ourselves to have; our desires and self-image; attempting to learn good habits; and so on. Then, when a situation arises when acting against the policy would otherwise be in our self-interest, we still have a motivating reason to act in line with the policy.

Again, I should emphasise that I am not currently trying to argue for any particular constructivist approach. Nor am I suggesting that this part of morality reduces to prudence. What I *am* trying to do, however, is to show that we can reasonably treat some part of morality as deriving force from non-moral considerations. Bear in mind that we are faced with the following triad of apparently incompatible claims:

i. Morality implies (apparently) categorical, and hence universal, reasons for action.
j. Reasons for action must be capable of motivating people.

k. The existence of specifically moral motivation in agents' subjective motivational sets is not universal.[29]

The proposed solution is to claim that, generally speaking, the persons to whom moral principles are addressed have reason to adopt moral principles. There is a large degree of overlap in basic interests and in circumstance, such that individual self-interest can ground motivating reasons to act morally. One problem with this proposal is that these reasons might look like the wrong kind of reason: the motivation behind honest action should be a regard for honesty, rather than prudence. But this does not contradict the claim that *adopting the policy* may be a prudential imperative, and, furthermore, an entirely appropriate justification. Prudential motivation may seem insufficient if it is construed as the sole motive for action: after all, we would not trust someone who would break promises, or tell lies, when it suited her. And this result generalises: there seems to be something inappropriate about taking non-moral reasons to justify moral behaviour. But this is not the case across the board. There *is* something inappropriate about someone who adopts a policy of refraining from torturing children because he thinks it in his best interest to do so. There is, I suggest, no such problem with someone who adopts a policy of keeping his promises because he thinks it in his best interest to do so. Nor is this merely a matter of degree: it is not simply that torturing children is more morally reprehensible than breaking promises. On that analysis, there would still be *something* morally dubious about someone who adopts a policy of honesty out of enlightened self-interest. But this, I take it, is not the case.[30] After all, inculcating dispositions of that sort does often involve adverting to self-interest – in contrast to the process whereby we teach children not to harm others, involving appeals to projective empathy ('how would you feel if someone did that to you?').

The existence of psychopathy might give reason to doubt that the concept of pain brings with it the concept of something that is to be avoided. Psychopaths appear to have a competent grasp of moral and mental terminology, but fail to be motivated to act in accordance with their (apparently sincere) moral judgements, or ascriptions of pain, suffering, etc.[31] In each of these cases we would wish to claim that the agent ought not to inflict gratuitous suffering on her victim, and that she has reason to refrain from doing so. Our pre-theoretic inclination is to maintain both of these claims regardless of whether the psychopath can be brought, by

[29] Amoralists are an extreme example of this. I take it that the lack of universality follows from the fact that the existence of specifically moral motivation is a contingent issue, given that it is implausible to suppose that the presence of specifically moral motivation is partly constitutive of what it is to be an agent. Of course, a thoroughgoing Kantian might disagree. I will leave this issue to one side.

[30] This division provides strong support for the hybrid theory on offer – c.f. the issue of 'revisionism about justification' discussed throughout.

[31] See Roskies 2003; Sinnot-Armstrong 2006: 119-171 & *passim*.

rational means, to be appropriately motivated. Indeed, the psychopath will often verbally accept these claims. So the existence of psychopathy need not give reason to doubt that there is a conceptual connection between pain and disvalue. There are two questions here: the first concerns the connection between moral judgement and motivation (the question of internalism about motivation); the second concerns the question of whether the psychopath *has reason* to refrain from inflicting gratuitous suffering on others (the question of internalism about reasons). I will address the question of internalism about motivation later in this chapter (in Section 4.3). As regards the question of reasons for action, I have claimed that constructivist morality provides reasons (strictly speaking) only for those who are appropriately motivated, but that realist morality provides reasons for everyone. Since at least some of these reasons will be external reasons, however, the challenge is to provide an account of what grounds these reasons. That is a task for Chapter Three.

There is a further problem here, concerning direction of explanation. There seems to be the following dilemma: if moral obligations provide reasons externally, then that retains the correct direction of explanation – the reason is provided by the object, rather than by the subject's interests – but at the cost of an implausible conceptual claim. We are left with the burden of making sense of the notion of something that is intrinsically reason-providing. The parallel, here, with the problem of revisionism about justification should be obvious. And the response, unsurprisingly, runs along similar lines. In some cases – to wit, those which concern pain or pleasure – revisionism about justification is deeply unpalatable, but we can, fortuitously, make ready sense of the notion that pain or pleasure are intrinsically reason-providing. The method is to show that part of the *concept* of pain or pleasure is a concept of a property which is intrinsically reason-providing, or has 'to be avoidedness/promotedness' built into it. I provide a detailed exploration of this claim in Chapter Three. In other cases, revisionism about justification is palatable; we can account for reasons to refrain from lying, for instance, in terms of elements of the agent's subjective motivational set. I explore that claim in Chapter Four.

2.5. Error Theory: Goldman

Goldman offers a further set of arguments against the existence of objective values, and thus insofar as the moral realist is committed to the existence of objective values – a set of arguments against moral realism.[32] He claims, for instance, that objective values would – if they existed – dictate that we attempt to maximise them. But no-one, he points out, actually

[32] Goldman 2007.

behaves in this fashion. Secondly, he suggests, the ‘fully objective’ view (Nagel’s ‘View from Nowhere’), rather than allowing a perspicuous view of values as they are, fails to show any values whatsoever. Thirdly, there is a measurement problem: even if objective values did exist, we would have no way of measuring them. Insofar as values are a function of our valuings, we can measure them; but insofar as they are not, we cannot.

There are three points to be made in relation to Goldman’s claims. The first is that axiology underdetermines what we ought to do. Indeed, axiology is only one part of any ethical theory. We could, for instance, hold that The Good is human pleasure – but that would not determine whether we ought to take whatever means possible to maximise human pleasure.[33] In a similar vein, if the correct axiology is one which specifies plural and incommensurable values, there is then no scale along which to compare values, and hence no obvious way in which these values could uniquely determine action.. Nor is it clear how much of a problem the general lack of compliance with morality’s dictates actually is. Consequentialists, for instance, when faced with the objection that their view of morality is ‘too demanding’, have been known to claim that morality is indeed so demanding that widespread compliance is unlikely. In relation to the second point, it is a mistake to identify ‘objective values’ (or objective anything else, for that matter) as those values (or things) which appear ‘from nowhere’, or ‘from the point of view of the universe’. Objective values, in this context, are those which exist independently of our individual responses to them. In contrast, subjective values vary from person to person, and depend on that person’s responses. But ‘independence’ does not mean that these values cannot relate to distinctively human projects or qualities; nor does it mean that these values should appear once persons or perspectives are entirely eliminated from the picture. The move to imagining things from ‘nowhere’ is better understood as a device for eliminating bias, rather than a characterisation of objective value as such.[34] Lastly, there is an epistemic issue: Goldman claims that ‘. . . we cannot calculate objective value even for [pleasure and cognate states], and so there is no practical relevance to the concept. Yet practical guidance and explanation of the rationality of our choices is the principal reason for positing objective value in the first place.’[35] But this is much too fast. Even if we cannot calculate objective value precisely, we can nonetheless make comparisons. Suppose that it is rational for me to attempt to satisfy as many of my own preferences as possible (one of which is that I remain healthy). I am not able to calculate precisely how many preferences will be satisfied by individual courses of action; perhaps a career in accountancy will satisfy more of my preferences than a career in plumbing, perhaps not. Let

[33] See Broome 2004: 31-33.
[34] See Enoch 2005.
[35] Goldman 2007: 523.

us suppose that there is no way to calculate, in advance, which will be best. This does not entail that there is no fact of the matter about which *is in fact* better; nor does it entail that there is no reason to prefer a career in accountancy over, say, a career in the Foreign Legion.[36] But Goldman is right, I think, to point out that morality had better not be out of the reach of human knowledge. I will explore this claim in the next section. But the objectivist, he thinks, will be hard pressed to account for this requirement. Call this the *epistemic constraint*.

3. The Epistemic Constraint

This stock objection to moral realism turns on the thought that moral truths, whatever they are, must be *knowable*. There is something unpalatable about the notion of an unknowable moral truth, just as there is something unpalatable about the notion of an undetectably funny joke, an unappreciably beautiful piece of art, etc.[37] In the case of the comic, we have a straightforward explanation of why this should be so: 'funny' is a response-dependent property. That is, our best understanding of what it is for something to be funny is dependent on our understanding of how agents would respond to it in certain conditions. A joke which no-one would ever laugh at is simply not funny. So the property (of funniness) and our hypothetical response to it cannot peel apart entirely. In contrast, the laws of physics might well turn out to be unknowable: it is entirely consistent with our limited mental capacities to suppose that full comprehension of the fundamental truths about the universe is beyond us.

So we have an explanation for why there are no undetectably funny jokes (viz., 'funny' is a response-dependent property). We now need an explanation for why there is something unpalatable about the notion of an unknowable moral truth.

One reason why this notion is unpalatable is the connection between blame and knowability. Ignorance often acts as an excuse in cases of wrongdoing. If we could not have reasonably been expected to know that an act was wrong, our wrongdoing is less culpable than it would otherwise be. Nor do we blame each other for unknowable obligations (I take it that there is no salient difference, in this respect, between moral obligations and moral truths in general). And if moral truth were entirely unknowable, there could be no practice of justified blame (because blame requires ascription of wrongdoing, and to justifiable ascribe wrongdoing implies that we can have justified beliefs about wrongdoing, which in turn requires that morality be epistemically accessible). Conversely, when we know that an action is wrong but

[36] Given that joining the Foreign Legion would frustrate a great number of my preferences.
[37] See Lillehammer 2003.

perform it regardless, we are blameworthy. So in order to secure the connection between blame and wrongdoing, we need to suppose that morality is, necessarily, knowable.

A second reason has already been discussed, and depends on the denial of the existence of external reasons. If moral truths must be such that they can be capable of motivating agents, then they must be knowable (since they motivate by way of being believed). And given the platitude that morality is essentially practical, there would be no work for unknowable moral truths to do (although perhaps an ethical theory might entail the existence of at least some unknowable truths, as an unfortunate but unavoidable corollary). As Lillehammer puts it:

'[this condition is] suggestive of the view that ends provide normative reasons by being favourably responded to in circumstances of rational deliberation. Thus, one might reasonably think that if no exercise of sound practical reasoning would make Jack want to torture Jill, then he has no normative reasons to do so.'[38]

But that is too fast. There is a distinction between being *knowable*, on the one hand, and being *knowable by any individual*, on the other. The second condition is much stronger. Consider the following two principles:

J1: If no exercise of sound practical reasoning would make X want to torture Y, then X has no normative reason to do so.
J2: If no exercise of sound practical reasoning would make *anyone* want to torture *anyone*, then *no-one* has normative reason to do so.

Denying J2 is the direct analogue of positing the existence of undetectably funny jokes. Denying J1 is less problematic. We can deny J1, whilst still claiming that morality has to be knowable as a matter of conceptual truth. Certainly, were moral realism just the thesis that moral facts are *sui generis*, out there in the world, waiting to be appreciated, then this epistemic constraint would be inexplicable. That, in turn, would pose problems for the moral realist. However, it should by now be clear that the restricted form of moral realism on offer stands a chance of avoiding this objection, since it deals with pleasure and pain: states which wear their normative quality on their sleeves, as it were. If no exercise of sound practical reasoning would make *anyone* within a community believe that they have reason to refrain from torturing anyone else, then it becomes questionable whether members of that community

[38] Lillehammer 2003: 2.

are competent users of the relevant concepts. I discuss this issue in more detail in Section 5.3 of the next chapter.

4. Expressivism

Expressivists deny *descriptivism*, and claim instead that the purpose of moral discourse and practice is primarily to co-ordinate attitudes, express pro- and con- attitudes towards deeds, convince others to act in accordance with one's wishes, and so on. In doing so they develop the emotivist thesis, promoted by (*inter alia*) Ayer and Stevenson, which claimed that moral statements were no more than mere expressions of emotion (such as Boo! and Hurrah!), but (they claim) in a way which avoids stock problems with simple emotivism.[39] These issues are orthogonal to the current discussion, however. What are relevant are the motivations underlying expressivism, and it is to these that I now turn.

4.1. Ontology

The first motivation is ontological, and depends on a commitment to metaphysical *naturalism*. As a rough gloss, metaphysical naturalism claims that the only entities which we should permit in our ontology are those which can, at some level, be explained in terms of entities posited by the natural sciences.[40] Even if the natural sciences do not use terms such as 'club' (as in 'group'), we might still hold that a club simply is a certain arrangement of human animals, engaged in a certain practice, and that these elements are reducible to terms of the natural sciences. In that sense, a 'club' is a natural entity.[41] Likewise with 'table', 'chair', and other non-fundamental but naturally-constituted entities.

One worry is that non-natural properties would be, as already discussed in relation to Mackie, *sui generis* and metaphysically 'queer'. Worse, even were such properties to exist, then the faculty by which we came to know these properties would have to be a *sui generis*, 'queer' faculty (of 'moral intuition', for instance). So the metaphysical worry entails an epistemic worry.

[39] Esp. Ayer 1946, Stevenson 1945.

[40] See the discussion in Chapter One, Section 2.

[41] This is a very loose gloss. In particular, the formulation is ambiguous between 'currently posited by the natural sciences', in which case metaphysical naturalism turns out to be (almost certainly) false, on the assumption of room for improvement within the natural sciences; and 'posited by idealised natural science', in which case metaphysical naturalism is trivially true. See Brown & Ladyman 2009. Brown & Ladyman propose a core formulation of *physicalism*, viz., that 'the mental supervenes on the physical . . . and that physics in its future development will not posit any mental entities in its theories.' (Brown & Ladyman 2009: 36-37). The loose gloss will, however, suffice for current purposes.

Insofar as moral realism is committed to a view of moral agents detecting spooky, non-natural entities using a spooky, non-natural moral radar, these considerations count against moral realism. Even if we allow that there may be non-natural entities (abstract objects such as numbers, perhaps), I take it that the default position is metaphysical naturalism – particularly given a view of human beings as animals which have evolved, over time, from very simple (non-moral) entities to (relatively) complex moral entities. Unless we are to commit ourselves to a view on which this transition was abrupt, then the account would have to be one of animals which gradually acquired the ability to detect these moral properties. But if the manner of detection is itself non-natural, we need an account of this faculty. None has yet been produced.

Expressivism, in contrast, sits comfortably with a thoroughgoing naturalism. Simon Blackburn, for instance, claims that, 'the state of mind of one who values something . . . is itself a natural, and naturally describable state. Once we find ethics here, we understand the essential phenomenon, which is that of people valuing things.'[42] The focus for metaethics, then, becomes the practice of valuing something, rather than the question of what, if anything, makes ethical propositions true. Gibbard also thinks that realism about practical facts ('to-be-doneness') commits one to the existence of queer properties, and that expressivism provides a way around this.[43] But expressivism is not the only theory according to which morality is perfectly naturalistic. Contractarian theories, such as those of Scanlon and Gauthier also construe morality as a natural system, one which is not committed to non-natural properties.[44] But then, it would be a mistake to suppose that moral realism is committed to the existence of 'spooky' non-natural properties.

There are, however at least two forms of naturalist moral realism. The core claim, here, is that moral properties are at least token-identical to some natural property or properties. A canonical form of naturalist realism, in this sense, might be the utilitarian claim that:

U: An act is right if and only if it maximises happiness.

However, we can disambiguate at least two senses of this claim. They are the analytic reading:

[42] Blackburn 1998: 50.
[43] Gibbard 2003: 5 & *passim*.
[44] Scanlon 1998, Gauthier 1986.

U_A: ‘Right’ just means ‘maximises happiness’,

and the synthetic reading:

U_S: The property of being right is identical to the property of maximising happiness.

Each – as will be discussed in Chapter Three – has difficulties. But more on that in due course.

4.2. The Motivation Motivation

The motivation motivation turns on the thought that there is a necessary connection between moral judgement, on the one hand, and motivation, on the other. I introduced this argument towards the end of the previous chapter. It begins with the thought that morality is essentially practical.[45] It is essentially practical, in that moral judgements have a necessary connection with motivation – or an intrinsic connection, or a conceptual connection. When coupled with the Humean view of motivation, this claim is supposed to provide a straightforward argument against moral realism. The Humean view of motivation claims that:

H1: All motivating reasons are complexes of desires and beliefs
H2: Desires motivate.
H3: Beliefs are motivationally inert.
H4: Desires and beliefs are ‘distinct existences’ (that is, there are no necessary connections between them).[46]

If moral judgements necessarily motivate, then they are necessarily entail both desires and beliefs. But for the moral realist, judgement serves to describe the world. So judgement must be a matter of forming beliefs. In order for judgement to motivate, however, it must also involve forming the relevant desire. Either this is intrinsic to moral judgement (hence judgement is not purely descriptive), or it is extrinsic. If extrinsic, then forming a moral belief necessarily involves forming a desire to do what is right, where that is understood to be either *de re* or *de dicto*. But either option contradicts H4.

Understanding moral judgements as consisting of pro- or con- attitudes, however, provides a natural way of understanding the connection between moral judgement and motivation. Now I

[45] See Sinclair 2007 esp. 1-2.
[46] For a discussion of this view, see Smith 1987.

do not wish to engage in an extended discussion here of whether the Humean view of motivation is correct. There is a problem with making sense of what it means to say that desires motivate, or that beliefs are motivationally inert. More seriously, it is unclear that the partitioning of mental states into beliefs and desires is appropriate.[47] There do seem to be cases of single mental states which serve dual functions – in Altham's word, 'besires'[48]. A thermostat, for instance, serves both to represent the temperature of the environment *and* to regulate it accordingly.[49] It is not clear what would motivate the claim that these two functions require two separate bases. But I will leave this to one side.

We ought, however, to get clear on what the motivation requirement amounts to. Sinclair distinguishes two options: on the first, moral convictions are 'intrinsically connected to the motives of agents'; on the second, they are 'necessarily connected to the motives of agents.'[50] Moral convictions are intrinsically connected to the motives of agents if and only if a moral conviction can suffice, by itself, to provide a motive. Moral convictions are necessarily connected to the motives of agents if and only if the presence of a moral conviction in an agent necessarily entails a motive. The two peel apart: defenders of the 'intrinsicality' model need not suppose that all moral convictions supply motives, but only that when they *do* supply motives, they do so 'by themselves'.

4.3. Internalism

What is the connection between moral judgement and motivation? We can start with the observation that people's motivations do often change to reflect their sincere moral judgements. Someone who comes to judge that it is wrong to torture cats will, hopefully, refrain from torturing cats. If I judge that I ought to give money to charity, then I will be motivated to do so. Conversely, we tend to judge people by their acts, as well as by their words: the fact that someone reliably donates money to charity indicates that they judge it right to do so; the fact that someone reliably refrains from torturing cats indicates that they judge it wrong to torture cats; etc. Call the view that there is some interesting necessary connection between moral judgement and motivation *internalism*.[51] Following Smith, we have:

[47] See e.g. Lewis 1988.
[48] Altham 1986.
[49] Sinclair (2007) adverts to Dawkins 1995: 74 and Papineau 1993, ch 3. §4 on this point.
[50] Sinclair 2007: 9.
[51] Informative discussion of this topic appears in Darwall 1992, FitzPatrick 2004, Svavarsdottir 1999, van Roojen 2000 & 2002.

Strong Internalism (SI): Necessarily, if an agent judges it right for her to φ in circumstances C, then she is motivated to φ in C.[52]

But this is too strong: we want, at least, to allow for agents in whom some defect prevents this motivation from going through. So perhaps we want a *defeasible* connection.

Weak Internalism (WI): Necessarily, if an agent judges it right for her to φ in circumstances C, then either she is motivated to φ in C or she is practically irrational. (Smith calls this 'the practicality requirement').

Strong internalism seems to be false. There are well-documented cases of agents whose capacity for sincere, competent moral *judgement* remains intact, but who entirely lack the relevant motivations.[53] Weak internalism is also problematic: what is meant by 'practically irrational' in this context? Many of those subjects who fail to be motivated by their (apparently sincere) moral judgements appear to be otherwise practically rational, at least insofar as they are capable of forming and executing plans, navigating the world around them, and so on.

Smith argues that both forms of internalism are defensible – partly because the denial of internalism results in a dilemma. Call the denial of internalism, externalism. The externalist is committed to holding that moral motivation is contingent. On what might it be contingent? Plausibly, on the presence of an appropriate desire to do the right thing. But that must be understood as either *de re* or *de dicto*. The *de re* reading, however, fails to explain why our motivations often change in line with our moral judgements. And the *de dicto* desire amounts (or so Smith claims) to a 'moral fetish'. Although Smith may be right about the *de re* reading, it is not clear that *de dicto* desires are 'fetishistic' in the relevant sense. Indeed, a concern to do 'whatever is right' is often wholly appropriate.[54] On this count, at least, externalism should not be seen as implausible.

Let us return to weak internalism. It is not only that there is a problem of explication, here. It also seems that there are people who, although in full possession of judgments about what they ought to do, take a great deal of pleasure in acting contrary to those beliefs. Certainly, it need not be that their behaviour involves inconsistency, *contra* Kant. On a purely *instrumental* understanding of practical rationality (which is, after all, at least pre-

[52] Smith 1994: 61-62.
[53] See Roskies 2003.
[54] See Lillehammer 2003; Dreier 2000; Doyle 2000; Olson 2002.

theoretically appealing), there appear to be actual counterexamples to weak internalism, and many of them: although people do often have moral motivations, and hence ends with moral content, not all people do so.[55] The sensible (and capable) knave, provided that she is suitably competent, will not be failing to take the means to her ends if she fails to have any moral motivations. The challenge, for the weak internalist, would be to specify the nature of the failing of someone who forms a moral belief, but not the appropriate motivation. Perhaps it is true that, if I believe that I ought to φ, then I believe that I have a reason to φ. And if I believe that I have a reason to φ, then either I am motivated to φ or I am practically irrational. But the first claim – linking beliefs about 'oughts' with beliefs about reasons – is not obviously true. Certainly the claim that moral truths entail reasons for action (moral 'rationalism') has been disputed. Can we, then, make sense of the motivating thought behind internalism *without* accepting either of the two above formulations?

Lenman considers the possibility of the *amoralist* (following Brink's terminology).[56] The stock internalist response, he notes, is to claim that amoralists are only making moral judgements in an 'inverted commas' sense. That is, they are not making genuine moral judgements, but rather are reporting on others' usage of moral terms. But reporting on others' usage of moral terms requires that there be others on whose usage to report. In other words, if amoralists are to make 'inverted commas' judgements, then they need a community of genuine moralists surrounding them. Now suppose that there were a world – call it Amorality – whose inhabitants devote a lot of time to detecting and reporting on the moral facts, but who do so in a purely academic fashion. It is 'really *jolly interesting* to know what these moral facts are but it's of no practical significance.'[57] This story is, says Lenman, 'preposterous'. It is preposterous, because 'the inverted commas sense of good is entirely *parasitical* upon the stronger *internal* sense without which it would be altogether empty.'[58]

In order to make sense of moral discourse without motivation, we need to suppose that agents are making 'inverted commas' judgements. But in the absence of *genuine* moral judgement – moral judgement *plus* motivation – there needs to be some way of importing sense into the disquoted (i.e. without inverted commas) judgement. In the case of actual, real-world amoralists, the suggestion is that the 'inverted commas' usage gets its sense from actual, genuine usage. So the 'world of amoralists' story is, if not logically incoherent, certainly preposterous.

[55] Broome 2002, Brink 1989., Svavarsdottir 1999.
[56] Lenman 1999, see Brink 1989, discussed at Chapter 1 §9.2.
[57] Lenman 1999: 446.
[58] Lenman 1999: 448.

This, however, does not get us all the way to *either* strong *or* weak internalism. What it does yield is *Very Weak Internalism:*

Very Weak Internalism (VWI): Necessarily, if a community uses moral discourse, then there is at least one member of that community who is generally motivated to act in accordance with her sincere moral judgements.

This is simply the denial of *hyperexternalism* – the claim that there is no connection between moral judgement and motivation whatsoever. Individual agents may consistently fail to be motivated to act in accordance with their moral judgements, for whatever reason, but nonetheless count as competent users of the discourse in virtue of being appropriately related to those who *are* motivated accordingly.

There are, however, two problems with Very Weak Internalism. The first is that it is simply too weak. This is in part a function of the sheer implausibility of a community where *only one person* is appropriately motivated. A story of a world of amoralists which contained only one moralist would be just as preposterous as the story of the world of amoralists. Not only is it implausible in itself, but it also raises the question of why, faced with such a world, we would think that the *motivated* agents are the ones making the genuine judgements. This question applies to any account which permits the minority usage to be treated as the genuine usage. Here, there is a contrast between the case of morality and the case of science. As far as science is concerned, we have a criterion of expert usage; although the majority may make no distinction between jadeite and nephrite, or be unable to tell the difference between elm trees and beech trees, we nonetheless defer to the minority usage. And our competent usage of the terms 'elm' and 'beech' *is* dependent on being appropriately related to others who are able to differentiate between the two. As far as morality is concerned, however, it is hard to see what grounds we could have for treating the majority practice as a deviant variety of the minority practice. Given this, here is a less weak form of internalism, called *Less Weak Internalism*:

LWI: Necessarily, if a community uses moral discourse, then members of that community are generally motivated to act in accordance with their sincere moral judgements.

The second problem with **VWI**, however, applies to all of these defeasible formulations. We need to give some explanation for why these defeasible formulations obtain, and this takes us straight back to the initial discussion of internalism. The connection can be readily explained by the invocation of an intrinsic connection: moral judgements simply are motivating states.

But this is too strong, for the reasons discussed above. A realist account would have to give some account of moral judgements wherein there is a conceptual tie between making judgements and generally being motivated. At this point, the naturalist might invoke an analytic tie between some natural property, F, and motivation, and then claim a non-analytic tie between that natural property and goodness.[59] Perhaps the property of goodness is identical with the property of our own well-being, and we cannot help but care about our own well-being. But the problem with *that* move, thinks Lenman, is that it leaves the concept of 'good' empty; it need have no *particular* descriptive content (since the property identity is a synthetic one), nor need it have any connection with motivation. This leaves the possibility of the world of amoralists again open. Worse still, the realist has to explain why it is possible for there to be *one* amoralist, but not for *all* members of a community to be amoralists.

I think that Lenman is right to worry about the chances of such an explanation succeeding. But he is wrong to think that attributing *analytic* connections between goodness and motivation, or between goodness and some natural moral property, or – even – between some natural property and motivation, implies *an intrinsic* connection. There is at least one case where an analytic connection does not imply an intrinsic connection, namely the case of Davidson's various interpretative arguments.[60] Similarly, there are relationships between non-moral beliefs and motivation which parallel the defeasible connections outlined by **WI** and **LWI** above. For instance, it is an analytic truth that, within a community, if members of that community believe that an object is dangerous, then generally they will be motivated to avoid that object.[61]

I will assume, then, that **LWI** stands in need of explanation by, and accommodation within, any acceptable metaethical theory. Chapters Three and Four outline how the realist and the constructivist, respectively, can meet this criterion. In the realist case, the strategy will be to claim that goodness is conceptually connected to the notion of something which is intrinsically normative, and that the best candidate for such an entity is also one which is conceptually linked to motivation. In the constructivist case, true moral principles are held to pick out solutions to co-ordination problems. There is therefore a conceptual connection between what we have moral reason to do, and what we have an interest in doing. But I will make this connection explicit in due course.

[59] Following Lenman 1999: 451. An instance of such a proposal is given in Boyd 1988.
[60] See Davidson 1984 esp. 17-43, 125-155, 183-199.
[61] See Simpson 1999.

5. Conclusion

The argument of this chapter has been as follows. I began by discussing three error theorists: Mackie, Joyce, and Goldman.

Mackie offers two core worries, one ontological – the 'argument from queerness' – and one empirical – the 'argument from relativity'. The argument from queerness, I suggested, could be defanged by providing a naturalistic account of ethics. Either moral facts amount to conceptual truths about natural properties such as pleasure, pain, and the like; or moral facts amount to non-conceptual (synthetic) truths about such properties. In either case, they are nothing peculiar. The argument from relativity is more complex: the serious problem is not one of persistent, widespread disagreement *per se*, but rather the appearance of *faultless* disagreement. Here I offered a bipartite response. Firstly, note that the extent of the disagreement is often exaggerated. In particular, the extent of *faultless* disagreement is exaggerated. Where disagreement concerns the realist part of morality, the disagreement is (*ex hypothesi*) faulty.[62] This places an explanatory burden on the realist, namely of making clear where the fault lies. I will address that task in the next chapter. Where disagreement concerns the constructivist part of morality, faultlessness is to be explained by reference to a plurality of correct answers. That is, constructivism permits relativism, and relativism permits widespread faultless disagreement. One significant concern, here, is that relativism permits only the *appearance* of faultless disagreement.[63] I address this concern in Chapter Four.

Secondly, given the prevalence of non-moral factors which influence moral belief, we should expect widespread and persistent disagreement. Hence the presence of such widespread and persistent disagreement fails to strongly motivate the claim that there is widespread *faultless* disagreement. The account being offered allows for both the possibility of faultless ethical disagreement and the possibility of faulty ethical disagreement. Importantly, the account being offered results in a concrete empirical prediction: all other things being equal, we should expect a difference in levels of convergence between realist and constructivist ethical beliefs.

[62] Assuming the falsity of relativism.

[63] That is, relativism threatens to prevent the possibility of genuine disagreement, since, for instance, claims issued from within different frameworks may each turn out to be true relative to that framework, and hence not incompatible.

The discussion then turned to Joyce's error-theoretic argument, according to which moral discourse is committed to categoricity, which in turn implies the existence of external reasons.[64] But such external reasons, Joyce thinks, do not exist. The response is again bipartite: I claim that the realist can account for the existence of reasons which do not depend for their normative force on some element of the individual agent's subjective motivational set. She can do so by identifying a candidate state for intrinsic normativity, such that agents can come to be motivated appropriately by 'coming to see the matter aright'. This enables the theorist to give sense to the notion of an external reason in a way which is not 'mere bluff'. The constructivist, however, cannot do this; in the absence of intrinsically normative states, we have instead to show that the appearance of categoricity is at least warranted. This is done by arguing that it is generally in the interest of agents appropriately situated to adopt some mutually binding set of guiding principles. This is further developed in Chapter Four.

The general strategy, in both cases, is to claim that in some cases there are features of the world which *do* do the theoretical work required of them (realism), and that in those cases where those features do not seem to obtain (e.g. the case of wanton lying) there is a constructivist account available to provide the required connection between moral principles and motivation.

Discussion of Goldman's error theory then lead on to a discussion of the *epistemic constraint* – that any account given should be able to explain the implausibility of the notion of an unknowable moral truth. The realist meets this constraint by giving a restricted account of what moral truths there are (as far as value is concerned, at least); the constructivist incorporates the epistemic constraint by claiming that moral truth is constituted by the outcome of some construction procedure which is itself epistemically accessible.

The final part of the chapter discussed motivations underlying expressivism, with a focus on the motivation motivation: the thought that there is a conceptual connection between moral judgement and motivation which realists have difficulty explaining. I argued that this connection is *loose*, and captured by the formulation of 'less weak internalism' (**LWI**). I claimed that some natural properties are themselves motivationally potent: this permits room for realism. The account on offer here is analogous to one on which external reasons (such as they are) are external to individual agents, but internal to communities. Other properties (the property of 'being a wanton act of lying', for instance) are not themselves motivationally

[64] Joyce self-defines as a *fictionalist*. However, his is an error theory insofar as it holds moral discourse to involve what is strictly a systematic error, viz., the commitment to categoricity.

potent: here we need an antirealist (constructivist) account. I will leave the discussion of constructivism to Chapter Four; the next chapter will address the substance of moral realism.

CHAPTER THREE

Moral Realism II

1. Introduction

The aim of this chapter is to explore the scope and substance of moral realism. I begin with a brief recap of the constraints developed in Chapter Two. Any putative moral facts must be metaphysically non-queer (and compatible with metaphysical naturalism); must be knowable; and must be capable of motivating those on whom they place demands. I then attempt to identify a natural fact (or set of facts) which plausibly provides moral reasons for action; which, in virtue of its nature, has 'to-be-pursuedness' built into it. If anything fulfils this criterion, we should think that facts about conscious experience fulfil this criterion. In particular, facts about pleasant and unpleasant conscious experiences fulfil this criterion.

If there are any intrinsic values, then hedonism is our best theory of intrinsic value - or so I argue. 'Intrinsic' here is meant in the strict sense. There is limited agreement on how to characterise intrinsicality. I will gloss it as follows: a property is intrinsic to an object if and only if that object has that property independently of how the rest of the world is.[1] Consequently, for something to be intrinsically valuable is for it to provide, in virtue of its intrinsic properties, *pro tanto*, agent-neutral reasons for action. I provide five arguments in support of the claim that hedonism is our best theory of intrinsic value. First, I argue that conscious experience is a necessary condition of value, at least for beings capable of conscious experience.[2] At the very least, for beings such as

[1] This is not meant to be a precise definition; see Lewis 2001, Trogdon 2009 for discussion. There is a weaker sense of 'intrinsically valuable', according to which an object is intrinsically valuable just in case it is non-instrumentally valuable. Audi refers to this as 'inherent' value (Audi 2004a; see also Feldman 2000: 320). I will be avoiding this usage.

[2] Kraut (2007: 8-10) points out that sunlight is good for plants; sugar is bad for petrol tanks; hence we have a conception of value ('good for') which is independent of conscious experience. Kraut goes on to develop an account on which good for an entity consists in its flourishing. But Kraut allows that pain is (usually) bad for sentient creatures, and therefore has to make sense of this disvalue in terms of what he calls 'pathologies of the sensory system' (Kraut 2007: 153). Pain's badness is to be understood in terms of our being in a state of 'un-flourishing', not merely in virtue of 'the way that it feels'. This is implausible: even if pain were intrinsic to our flourishing, it would still be disvaluable. Similarly, painful sensations often result from the perfect functioning of our sensory system – despite Kraut's assertion that 'the sensory system we have been given by nature is disordered and not functioning as it should, from the point of view of our well-being . . .' (Kraut 2007: 150).

ourselves, the possibility of conscious experience is a precondition for value.[3] Quite plausibly, value inheres in some forms of conscious experience. Secondly, I argue that hedonism provides our best account of the ends of human action: pleasant and unpleasant experience provides the required degree of finality. Thirdly, I consider an 'evidential' argument: if there are any bearers of intrinsic value, then our moral behaviour points towards them. Fourthly, I consider the related claim that widespread convergence in ethical beliefs points towards bearers of intrinsic value. Fifthly, I give an argument from interpretation: one of the principles which constrain radical interpretation, I suggest, presupposes hedonism. I return to these considerations later in the chapter.

I then consider four objections to hedonism. Firstly, I consider the possibility of agents who are indifferent to their own hedonic states. Secondly, I consider the possibility of agents who are rational, but indifferent (or worse) to the hedonic states of others. The latter are 'rational sadists'. Thirdly, I consider the issue of desert. Fourthly, I consider ethical particularism. Each of these objections can, I claim, be comfortably accommodated by hedonism.

Having identified this candidate bearer of value, I then consider Moore's 'Open Question' argument. This is, broadly speaking, a semantic concern. I offer a diagnosis of the argument itself, in light of the contemporary literature, and suggest that the argument does still pose a concern for ethical naturalists. However, this concern can be addressed. I show how restricting naturalism to a theory of value helps in this respect.

As the value in question is supposed to be moral, rather than merely personal, value, I then make this term more precise. If pain is intrinsically morally disvaluable, then the fact that an action would cause someone pain gives an agent-neutral, *pro tanto* reason to refrain from that action. I then turn to discussion of the notion of an 'agent-neutral reason'. Agent-neutral reasons are reasons for *anyone* to perform an action, maximise an outcome, and so on; our moral discourse is naturally understood as committed to the existence of such reasons. In particular, I consider Nagel's argument in favour of the existence of agent-neutral reasons, and Sturgeon's objections to Nagel.

Lastly, I show how the proposed 'restricted realism' can meet the epistemic and motivational constraints discussed in the previous chapter. Crucially, conscious states such as pleasure and

[3] C.f. Sidgwick 1930: 113-115.

pain have phenomenal content: there is a way that it feels to be in such states. Furthermore, I claim, the phenomenal content of these states is concomitant with their intrinsic normativity: for a feeling to be unpleasant is for it to have a "to-be-avoidedness" about it. And since it is a conceptual truth that these constraints apply, it should also be the case that it is a conceptual truth that restricted realism is true.

2. Constraints

2.1. Naturalism and Non-Queerness

The first constraint is that our moral theory not invoke properties which are metaphysically 'queer'. If the theory is to invoke entities which are somehow intrinsically normative, we need to show that these are not metaphysically queer. One way of doing this would be to identify entities which we already accept in our ontology, which are no more or less metaphysically queer than the suggested moral properties. This could take the form of a 'companions in guilt' argument (we already accept epistemic facts; epistemic facts are intrinsically normative in the same way as moral facts; so we should also accept moral facts), or a 'direct' argument, one that shows how the proposed states themselves can be intrinsically normative. I will take the second approach.

A corollary of the 'non-queerness' constraint is that the proposed moral facts should themselves be compatible with metaphysical naturalism (this is a desideratum of the theory, rather than a strict requirement). There are several motivations behind the denial of non-natural properties. In particular, part of the putative 'queerness' of moral properties, for Mackie, lay in their being utterly unlike anything else in our ontology. We accept the existence of tables, chairs, atoms, electrons, and so on, all of which are taken to be perfectly natural, but non-natural properties would not seem to fit in to this picture; would require non-natural detective faculties, and so on. The strategy on offer aims to show that moral properties are not utterly unlike anything else in our ontology.

2.2. Epistemology

The epistemic constraint requires simply that moral facts be *knowable*. This does not require that they be knowable to all persons (although it would be surprising if they were mostly unknowable). Furthermore, we ought to be able to give an account of *how* we can come to know these moral facts. On the account on offer, both empathy and conceptual analysis can provide routes to moral knowledge. I develop this claim later in the chapter.

2.3. Motivation

The motivational constraint requires that sincere moral judgements be capable of motivating agents. Specifically, it takes the form of 'less weak internalism'.

LWI: Necessarily, if a community uses moral discourse, then members of that community are generally motivated to act in accordance with their sincere moral judgements.

Furthermore, as with the 'epistemology' constraint, this connection ought to be explicable. Whichever form our metaethical theory takes, it should explain how moral judgement is capable of motivating persons, and why, if moral judgement does not always motivate, the *general* connection nonetheless holds. The substance of the realist theory on offer will be illuminating in respect to these issues: I will return to them at the end of the chapter.

3. Hedonism

Hedonism encompasses a range of positions. It is not, however, the purpose of this chapter to discuss the history of, or adjudicate between, these positions. Instead, I will lay out one specific form of hedonism. Nomenclature is, for the time being, unimportant.

Minimally, hedonism is a thesis about value. That is, hedonists think that pleasure is intrinsically good, and that pain is intrinsically bad. More precisely, they think that pleasure is *the* (intrinsic) good (and *vice versa*) .[4] More generally, hedonism is a thesis about what makes a person's life go well. It goes well insofar as it contains pleasure, and badly insofar as it contains pain. Pleasure and pain are both taken to be conscious experiences, and exist in a variety of durations and intensities. Quite obviously, in order to flesh out a hedonist theory of value, we need to add considerable detail regarding our conceptions of pleasure and pain. I will turn to this issue later in the chapter. For the time being, I will stick to talk of pleasure and pain.

[4] C.f. Moore 1922: 62.

Note that, as it stands, hedonism is a claim about what makes a person's life go well. That is, it is a claim about *personal*, rather than moral, value. Personal value provides practical reasons for the potential recipient of that value. If pleasure is what will make my life go well, then I have reason to pursue pleasure - or at least, if possible, to maximise my life's pleasure content. [5] But this by itself does not provide other people with reasons to promote my pleasure, or to maximise, if possible, my life's pleasure content.

However, laying out a coherent theory of intrinsic value does go some way to accounting for moral value. Significantly, once we admit the existence of intrinsically valuable entities, we have committed ourselves to intrinsically normative entities. This should remove worries about the 'queerness' of such entities. This, in turn, allows metaphysical room for entities (or states of affairs) which provide reasons for everyone. Towards the end of this chapter, I explore whether such states do in fact provide reasons for everyone – and, if so, how. I now wish to discuss the thought that intrinsic value attaches to hedonic and anhedonic states, beginning with an argument in favour of 'axiological experientialism' – the claim that the sole bearer of intrinsic value is experience. Hedonism, I will suggest, forms the most plausible extension of this thesis.

3.1. Finality

Robert Audi has claimed that experiences are more 'final' in terms of value and motives than are their objects.[6] They are more 'final' in that we, for instance, go to an art gallery for the sake of the experience of seeing the artworks; we go fishing for the experience of going fishing; we value great symphonies on account of the experience of listening to them, and so on. It is because we can experience these things that we value them. Furthermore, Audi states, experiences can 'figure as the objects of practical attitudes such as desire'; can have '*motivational potential'* [7] (that is, can motive their pursuit); and have 'the internal accessibility appropriate to reasons and their normative contents'.[8]

It would be wrong, however, to suppose that we always consciously desire the experience of going fishing, listening to a symphony, as opposed to the desiring the object of that experience. We often desire to engage in an activity, or to achieve an end, where that desire takes the activity,

[5] There is an exception here, namely where the pursuit of pleasure is self-defeating.

[6] Audi 2004a: 122-130.

[7] Ibid.

[8] Audi 2004a: 124.

rather than the experience, as an object. But in most cases, removing the possibility of experience renders the activity pointless. In such cases, the value of the activity is at least conditional upon the possibility of experience. In other cases – activities aimed at generating morally valuable outcomes, for instance – we tend to think that the possibility of experience does not condition the value of these outcomes. People reliably make provisions to care for their descendants after their death, for instance, whether via wills, life insurance, and so on. But it is an interesting feature of these cases that they require the conscious experience of other persons. Instances where people treat outcomes as morally valuable even though there is no possibility of experience (by themselves or others) are rare.

Similarly, in a world without agents – without some locus of conscious experience – it is implausible to think that anything would be of value in that world. An 'empty' universe consisting entirely of beautiful paintings would be no better or worse than an 'empty' universe consisting of dunghills. Some – Moore, for instance – have thought differently. But there is at least one strong confounding factor here, which is that it is difficult not to conceive of worlds which contain beautiful paintings as presenting *possibilities* of experience. Once there is a possibility of perceiving either world, there is a difference between the two worlds: one contains the possibility of valuable aesthetic experiences, whilst the other does not.

Furthermore, there is a strong argument to be made in favour of the view that values are necessarily anthropocentric. They are anthropocentric, in that they can only be understood in relation to the behaviour and constitution of (as it happens, human) evaluators. Suppose that, overnight, all humans in all places simultaneously underwent some aesthetic inversion: when they awoke, they saw dung-hills as beautiful, and delighted in looking at them, but were left mildly repelled by the contents of the Louvre. It would be implausible to talk of dunghills as repellent and Matisses as beautiful (except in a purely historical, non-evaluative sense). This is because beauty is a 'response-dependent' concept: it cannot be understood except in relation to our responses. If this is correct, then it is hard to make sense of the claim that the empty universe full of beautiful paintings is 'better' than the empty universe full of dunghills. The only viable interpretation is that the empty universe would generate a certain set of responses, in a certain set of conditions.

Now I do think that we should accept that value is a response-dependent term, at least in the weak sense that any genuine moral values must be knowable, and must be capable of motivating appropriately situated agents. In most cases, accepting that the value of an object depends on our responses to it, whether hypothetical or actual, entails that the object cannot be intrinsically valuable, since all conditional properties are relational. But there is one notable exception to this claim: namely, where the object itself is also picked out in response-dependent terms. This, I take it, is close to what Audi means by 'the internal accessibility appropriate to reasons'. Since experiences are responses (broadly construed), they may also be intrinsically valuable.

3.2. Nozick's 'Experience Machine'

There is at least one strong argument against axiological experientialism, namely Nozick's 'Experience Machine' argument. Nozick imagines a machine which can simulate, perfectly, any experience which is wished for (it does so in such a way as to maximise overall satisfaction: the machine won't let you become bored, for instance). The machine yields a life which, from the hedonist's point of view, goes extremely well, being full of pleasure and devoid of pain (except insofar as pain serves to maximise pleasure – by providing contrast, perhaps). However, the machine merely provides a simulation: we are not *actually* surrounded by appreciative friends, for instance. Most people would choose not to be plugged into the machine (so Nozick thinks), and this shows that most people do not take pleasure to be the sole bearer of value for us: '[w]e learn that something matters to us in addition to experience by imagining an experience machine and then realising that we would not use it.'[9] Nor would we choose to use a 'transformation machine', which directly transforms its subject into whatever that subject wishes; nor a 'result machine', which directly produces whichever *result* people desire (although many people do take such shortcuts when offered – fraudsters, for example). So, Nozick concludes, 'something matters in addition to one's experiences *and* what one is like'.[10]

It is possible that Nozick conflates, in this argument, three concepts: what matters to people (understood as a question of what they care about); what matters *for* people (understood as a question of what makes their life go well); and what matters (understood as a question of objective moral value). Certainly, people will often choose the genuine experience over the illusion of experience, even if the genuine experience would be less pleasurable. But this shows something about preference, rather than value, and is readily explicable. Once a desire to Φ has been formed, then that desire has Φ-ing, rather than the experience of Φ-ing, as its object. So we,

[9] Ibid. 44.
[10] Nozick 1974: 44.

for instance, want to go fishing, rather than to have the experience of going fishing. The experience machine, then, does not satisfy any of these desires. But that does not show that what gives these desires their point is the activity, rather than the experience. Conversely, we would not choose to enter a putative 'no-experience' machine, which augments our capacities for fulfilling our projects (although it is still *our* agency that completes these projects) but denies us the experience of their fulfilment.

One might attempt to defend experientialism by claiming that, even if we *do* desire experiences, rather than the objects of experience, the Experience Machine will fail to provide the requisite experience, since the desired experience is an experience *of* its object, whereas the Experience Machine only provides an experience *as of* its object.[11] But this, I think, would be misguided: after all, the experientialist's thesis is that it is the way things are 'from the inside' which matters. That is, value is a function of something to which we have introspective access, or which we are aware of. To allow the value of conscious states to depend on whether they are experiences of friendship, fishing, etc., as opposed to perfect simulacra – experiences as of friendship, fishing, etc. - runs counter to the spirit of hedonism, and experientialism more generally.

Nozick's 'Results Machine' also cuts against this experientialist manoeuvre. The Results Machine – coupled, if appropriate, with the Experience Machine – guarantees that the experiences will be genuine (since the required results actually obtain). But, thinks Nozick, we would still be wary of using the machine. It would still feel like 'cheating'. On this point, I am doubtful. It is certainly the case that there are occasions when it would be analogous to cheating, and where we would be wary of using the machine. But consider the case of games such as solitaire, or computer games played by one person. In these games, the player attempts to achieve a certain result. Achieving that result by simply pressing a button would amount to cheating, and there is good reason not to do so: the game would no longer be fun, or enjoyable. So here the problem with cheating is explained in terms of the impact on the player's experience.

In other cases, however, it seems very obvious that most people *would* choose to use the machine (or machines), and that it would be entirely rational to do so. Suppose, for instance, that one was in severe, continuous pain, or had an 'empty' life, devoid of both pain and pleasure. Or suppose that one had the option of becoming instantly healthy, or improving *just enough* to surmount a

[11] Certainly people *sometimes* desire the experience, rather than its object; the question is whether this is what *matters*.

plateau in one's saxophone-playing ability. It is entirely unclear that people's intuitions do indeed lie as Nozick supposes them to.

There are also methodological concerns here. Even in those cases where people appear reluctant to enter the Experience Machine, their reasons for reluctance may vary. Given that our intuitions are developed in real-world situations, where machines are often faulty, some of this reluctance may be due not to the illusory nature of the experience, but rather due to the possibility that the machine may malfunction, or fail to supply the requisite experiences. Worse still, inverting the experience machine experiment seems to indicate that people's reluctance may be largely due to *status quo* bias – a preference for things to remain as they are - or something similar.[12] In the inverted experience machine case, subjects are asked to suppose that they discover that all of their experience to date has been 'illusory' (i.e. a product of the experience machine); they are then asked whether they would like to unplug from the machine, or remain. Many indicate a preference for remaining. This suggests that people's responses do not depend on the property (genuineness) which Nozick supposes them to.

3.3. Evidence

There is a long history of arguments which take human behaviour as evidence that pleasure (or happiness, or some closely related concept) is the good. Aristotle, for instance, mentions that:

'Eudoxus thought pleasure was the good because he saw all things, *both rational and irrational*, aiming at it, and because in all things that which is the object of choice is what is excellent, and that which is most the object of choice the greatest good; thus the fact that all things moved towards the same object indicated that this was for all things the chief good (for each thing, he argued, finds its own good, as it finds its own nourishment); and that which is good for all things and at which all aim was *the* good.'[13]

We can expand this argument as follows. Some things in the world are choiceworthy, i.e. good, and people's choices are a good indicator of which things are choiceworthy. So if all things pursue pleasure, and avoid pain, this indicates that pleasure is a good to that thing, and the claim generalises: each thing's pleasure is a good to that thing. The thought that choices are an indicator

[12] de Brigard 2008.
[13] Aristotle 1925: 249. Aristotle was, however, critical of this suggestion.

of choiceworthiness is not entirely unreasonable: after all, there is something curious about thinking that choice and choiceworthiness might peel apart altogether. I speculate that this is perhaps just a corollary of the constraint on radical translation which requires that we interpret people as being generally rational, and their choices as generally making sense, i.e. tracking choiceworthiness – but I will discuss this suggestion further in **3.4**. Peter Railton offers a similar suggestion for determining bearers of intrinsic value:

‘To my knowledge, the best-developed method for justifying claims about intrinsic value involves thought-experiments of a familiar sort, in which, for example, we imagine two lives, or two worlds, alike in all but one respect, and then attempt to determine whether rational, well-informed, widely-experienced individuals would (when vividly aware of both alternatives) be indifferent between the two or have a settled preference for one over the other.’[14]

One might think that this presupposes moral realism – or is at least realist in flavour. The realist, after all, claims (*inter alia*) that some things in the world are actually choiceworthy; there are some bearers of intrinsic value. But a sophisticated expressivist can also make sense of the claim that some things in the world are choiceworthy: to say that something is choiceworthy (as with saying that something is valuable) is to express a pro-attitude towards that object, or action, and towards choosing that object, or action, etc. Furthermore, the sophisticated expressivist can make sense of the claim that our beliefs *track* what is choiceworthy. Blackburn, for instance, points out that,

‘[a]ny proposition is given its sense by a set of rules or norms telling us what counts as proper evidence for it . . . these norms also tell us that some situations yield better evidence than others . . . To conduct yourself so that you believe *p* only if *p* is to conform to such norms.’[15]

Gibbard makes a similar claim:

‘. . . we start with opinions, and these tentative opinions are not only matters of meaning. They include rough thoughts on what kinds of moral inquiry are to be trusted. Now, one things we can do is just start with these opinions and proceed, with a special philosophical determination.’[16]

[14] Railton 1984: 149 fn.21.
[15] Blackburn 1998: 293.
[16] Gibbard 1990: 315.

Our beliefs about what is choiceworthy, therefore, track the truth about *p* if and only if they conform to the relevant norms. The relevant norms dictate that reflection, lack of bias, full provision of information, and so on, count as yielding better evidence for our beliefs than their absence. Contrariwise, if our beliefs are guided by unpalatable, alien forces (evil neuroscientists, or the Party of Orwell's *1984*), then they will fail to be appropriately governed.

Now the evidentiary interpretation does, in this context, presuppose a dyadic relation: it holds that there are two distinct elements, being our actual choices, on the one hand, and what is choiceworthy, on the other. If we accept that our moral beliefs are true if and only if they correspond to what is morally the case, then we are committed to thinking that there are at least two elements to the correspondence.[17] This, in turn, commits us to holding that the choiceworthiness of the option can (at least sometimes) *explain* our belief that the option is choiceworthy. The evidentiary interpretation brings with it this explanatory constraint. If the choiceworthiness of the options can explain our belief in their choiceworthiness, then there must be a genuine property – choiceworthiness – capable of explanatory work.

Can this be accommodated without realism? If 'genuine choiceworthiness' is to be understood in terms of responses under ideal conditions, then to say that we believe that Φ is choiceworthy *because* Φ is genuinely choiceworthy is to say that we believe that Φ is choiceworthy because we would believe that Φ is choiceworthy under certain conditions – and, furthermore, these conditions obtain.[18] This second clause is implicit, rather than explicit. When explaining our physical beliefs, the presence of a red object only explains our belief that there is a red object in front of us under normal conditions. Similarly, the beauty of an object (for instance) can explain our belief that that object is beautiful, even if 'beauty' is understood as a response-dependent property. However, this is more difficult for the expressivist. The challenge is to give an expressivist account of the claim,

S: A's belief that P is beautiful is explained by P's beauty

We might follow Blackburn, and gloss **S** as claiming that A's beliefs 'track the truth' about P, where this is a matter of having beliefs regulated by the appropriate norms. But this also requires

[17] Skorupski, personal discussion.

[18] Where the phrase 'ideal conditions' is given an expressivist analysis, as per Blackburn or Gibbard.

that we have a viable notion of “the truth about P” - and the most plausible route here is to invoke superassertibility (as Blackburn does). So to assert **S** is to claim that “P’s beauty” amounts to “the fact that ‘P is beautiful’ is superassertible”, and hence that the superassertibility of “P is beautiful” can be genuinely explanatory of A’s belief that P is beautiful. This, however, is implausible. The suggestion would be that A is constituted as he is *because* he would remain so in relevant respects despite any improvement in his epistemic state, where what counts as an improvement is specified by the norms of the discourse. That seems wrongheaded.

This, however, should be expected. After all, one of the distinguishing features of expressivism is that it, in contrast to descriptivism, understands the function of moral discourse as primarily one of expressing and coordinating attitudes. In contrast, descriptivists hold that our moral beliefs serve to represent some moral state of affairs; that requires that whatever is morally the case can, in principle, explain our moral beliefs.[19] Descriptivism, then, requires that the moral facts be explanatory of our moral beliefs in a way which expressivism does not.

Expressivists may still have a theory of the Good, and it may be one of the norms of moral discourse that human behaviour counts towards determining the Good; both expressivists and descriptivists, therefore, can think that our behaviour counts as evidence for what is morally the case. The two come apart over the issue of explanation. I now turn to a question which forms part of the issue of explanation: that of convergence.

3.4. Convergence

In fact, although there is widespread disagreement over how society ought to be structured, or whether it is acceptable to eat certain kinds of foods, the treatment of homicides, women, etc., there does appear to be widespread convergence on the position that pain (or, more accurately, suffering) is intrinsically disvaluable, and pleasure intrinsically valuable - not only in the sense that individual agents find their own pleasure to be an attractive prospect, but in the sense that agents take the suffering of others to provide agent-neutral reasons for them to relieve that suffering.[20] This is reflected in prohibitions on wanton harming, and praise of benevolent activity, both in commonsense morality and in religious ethics, although is often obscured by the confounding factors discussed in the previous Chapter. There are, of course, many cases where

[19] Otherwise, we have an error theory.
[20] See Westermarck 1906, 1908.

agents treat suffering as valuable: practitioners of certain religions, for instance, treat ritual flagellation as important. More trivially, athletes may come to think of mild muscular soreness as a positive outcome. But all of these cases can, I think, be made sense of as instances of extrinsic value: the monk's flagellation is held to be valuable because mortification of the flesh brings one closer to God; muscular soreness is an indicator of a successful (that is, performance-enhancing) workout; Spartans might seek out painful experiences as a means of demonstrating their fortitude, prowess as warriors, and so on. Certainly, the vast majority of these cases can be accounted for in this way. So the claim about convergence is a highly restricted claim, and relates only to the reasons provided by pain as an end in itself. This supports the claim that pain is intrinsically disvaluable in the realist sense.

3.5. Interpretation

The 'Twin Earth cases' discussed by Lenman, Hare, and others are also relevant to the issue of whether there is a conceptual connection between pain and disvalue.[21] Remember that one worrying case was that of patients who appear indifferent to pain; they do not report aversion to pain, and this makes it look less plausible that pain is intrinsically disvaluable. Now it might be that these patients do not, in fact, experience pain, but rather some related but dissimilar state – call it 'pain*' – which is neither unpleasant, nor disvaluable. But that response is problematic: it is not clear how we could characterise the sensation in a way which retains the connection with pain proper. Imitation watches count as imitation watches in virtue of having a representational relationship to real watches, but is it plausible to think that pain* could feel like pain but not be actually count as pain? Put differently, if it feels like pain, then it is pain. And presumably we should take the sufferers' reports – that it feels like pain – seriously.[22]

However, suppose that there were a Twin Earth whose inhabitants used the terms 'pain' and 'pleasure', but never seemed to be bothered by the occurrence of what they called 'pain', or driven to pursue what they called 'pleasure' (their language is otherwise orthographically identical to ours, as is their environment and constitution, broadly construed). In this respect, they behave just like Earthlings under the influence of certain dissociative anaesthetics (morphine,

[21] See Lenman 1999; Hare 1991; Davidson 1973.

[22] I take it that first-personal reports of pain should be taken as authoritative, but will not argue further for this claim here.

ketamine), or cingulotomy patients.[23] In this case, it would be highly implausible to think that Twin Earthlings share our concept of pain. We would have to give some account which made sense of their usage of the term ‘pain’ without requiring that their experience of ‘pain’ have some ‘to-be-avoidedness’ attached. The parallel with the weaker forms of internalism about moral motivation (c.f. Chapter Two) should be obvious: although it may be coherent to imagine individual cases where pain and unpleasantness (the experience of the sensation as one which is to-be-avoided) come apart, this is only coherent against a background on which pain and to-be-avoidedness are connected. So it is, I think, a conceptual truth that in any community which uses the concept ‘pain’, members of that community represent the experience of pain as generally ‘to-be-avoided’, or providing reasons for its avoidance. It is also a conceptual truth that members of that community are averse to the experience picked out by the concept ‘pain’.

Similarly, as Goldstein points out, one way in which we individuate the sensations of pleasure or pain is by their value or disvalue.[24] This is a straightforward way of rephrasing the claim of the last paragraph: there is a conceptual (if defeasible) connection between ‘pleasant’ and ‘good’. There is further explanation required, however, if we are to make sense of this value as a distinctively *moral* value – in other words, one which provides (pro tanto) agent-neutral reasons for action. Nor should this be treated as providing a direct argument for moral realism: rather, I take it that there is a weak presumption in favour of moral realism, and am suggesting that we can readily make sense of the (supposedly problematic) ‘intrinsic normativity’ of moral properties by pointing out that we are already committed to thinking of at least some mental properties as intrinsically normative.

4. Agent-Neutral Reasons

4.1. Definition

An agent-neutral reason is one that is not bound to any specific person. In contrast, an agent-relative reason will, if fully specified, make reference to some particular person(s). An agent-neutral reason will take the following form:

[23] A cingulotomy is a lesion of the cingulated gyrus, which is a part of the brain which connects the frontal lobes and the limbic system. The frontal lobes are thought to deal with ‘higher order’ thought (including language, planning, and so on), while the limbic system is thought to deal with emotional processing, ‘gut’ reactions, etc.

[24] Goldstein 1989.

ANR: For any agent A, the fact that Φ-ing will reduce X's pain gives A reason to Φ

Whereas an agent-relative reason might be:

ARR: For any agent A, the fact that Φ-ing will reduce the pain of A's children gives A reason to Φ

Thomas Nagel glosses the distinction as follows:

'If the avoidance of pain has only relative value, then people have reason to avoid their own pain, but not to relieve the pain of others (unless other kinds of reasons come into play). If the relief of pain has neutral value as well, then anyone has reason to want any pain to stop, whether or not it is his.'[25]

I now want to consider whether pleasure and pain provide agent-relative or agent-neutral reasons for action.

4.2. Nagel

Nagel originally thought that denying the existence of agent-neutral reasons committed one to a form of practical solipsism, but revised this view in the light of criticisms by Sturgeon, amongst others.[26] In *The View from Nowhere*, however, Nagel presents a revised argument for the claim that pain provides agent-neutral reasons for action.

'pain . . . is just as clearly hateful to the objective self as to the subjective individual. I know what it's like even when I contemplate myself from outside, as one person among countless others . . . The pain can be detached in thought from the fact that it is mine without losing any of its dreadfulness.'[27]

Furthermore:

[25] Nagel 1986: 159. See Korsgaard 1996a: 275-310 for an attack on the distinction.
[26] Nagel 1970; Sturgeon 1974; Nagel 1986: 159 fn. 6.
[27] Nagel 1986: 160.

‘If we assign impersonal value to pleasure and pain, then each person can think about his own suffering not just that he has reason to want it gone, but that it’s bad and should be got rid of.’[28]

And:

‘to say that there is no agent-neutral objection to suffering . . . would again be to overrule the clearest authority present in the situation . . . My objective attitude toward pain is rightly taken over from the immediate attitude of the subject, and naturally takes the form of an evaluation of the pain itself.’[29]

There are at least three distinct lines of argument here. The first argument depends on the claim that when we contemplate pain *per se* it retains its ‘dreadfulness’, or ‘to-be-avoidedness’. The second argument supposes that we naturally think of our own suffering as impersonally bad, and points out that unless we allow the existence of agent-neutral reasons, we must deny ourselves this characterisation, and this denial would be a theoretical demerit. The third argument asserts that from the objective point of view, we must take the sufferer’s own attitude towards the pain seriously – and since the sufferer’s attitude represents the pain as being bad (or ‘dreadful’), we should take this representation seriously.

Although the second argument is relatively weak, the first and third arguments deserve serious consideration. If we look at things objectively, then, thinks Nagel, we are committed to acknowledging the existence of agent-neutral reasons to relieve pain, since it is the sufferer’s viewpoint which is authoritative, and that viewpoint represents the pain as bad. As far as attributions of reasons and values go, we have to take seriously the reasons and values which are actually given.

There is a worry, here, that this view risks showing too much. There are many desires and valuations which are dressed in agent-neutral terms, but clearly do not provide agent-neutral reasons for action. Many of my peculiar desires seem to not to specify any particular agent, and are not obviously self-referential; they are desires that certain outcomes obtain, rather than that these outcomes obtain through my agency. For instance, I may want to become a writer simply

[28] Loc. cit.
[29] Nagel 1986: 161.

because I desire the associated goods; I may want people to have been killed *simpliciter*, rather than wanting myself to have killed people; and so on. Nagel's response to this is that the reason-giving force of these desires turns on their idiosyncrasy: idiosyncratic wishes (those that depend on my peculiar psychology) do not provide reasons for everyone. This is because the reason-giving force of these wishes cannot be detached from the agent: they are distinctively reasons *for her*, rather than reasons simpliciter. That enrolling in a literature degree would make me a better writer, for instance, only rationalises my behaviour when set against the background of my peculiar motivational set. The reason does not detach from the person.

In contrast, the undesirability of pain *does* detach from the peculiar makeup of individual agents. This is Nagel's first argument. The sufferer of the pain wishes to get rid of the pain not because it is his, but simply because it is painful: '[t]he desire to be rid of pain has only the pain as its object.'[30] This is not to say that pains can exist without sufferers, but rather that the concept of pain does not depend on the concept of the self. If we suddenly lost the ability to think of ourselves as embodied, individual agents, we would not therefore lose the ability to think that pain was bad. Put differently, the evaluation attaches to the state, rather than to the state *and* the person. So I am committed to thinking that my being in pain does not just provide *me* with a reason to relieve my pain; I am also committed to thinking that *anyone* has reason to relieve my pain.

There are at least two problems with Nagel's position.[31] Firstly, it is not obvious that the desire to be rid of pain *does* only have 'pain' as its object, rather than 'my being in pain'. Suppose that one were in pain, and had the option of transferring the pain to another person. On Nagel's account, our pain-regarding desire would not translate to a desire to transfer the pain to another person. But it seems highly plausible that people *would* want to transfer the pain to another person (even if this desire would be overridden by moral qualms). So it seems that individuals treat their own pain as supplying agent-relative rather than agent-neutral reasons. This, however, is too fast. It is quite possible – and, indeed, plausible – to maintain that pain yields both agent-neutral and agent-relative reasons for action. Hedonistic utilitarians will, presumably, agree that pain supplies both types of reason: we have prudential reasons to avoid our own pain, and moral reasons to avoid causing pain to others. On a more charitable interpretation of Nagel's argument, the core point is

[30] Nagel 1986: 161.
[31] If not more: see Carlson 1990, Rachels 2002.

simply that our conception of pain (and hence our conscious response to our own pain) is evaluatively loaded.

As regards the third argument, that the correct objective attitude towards the pain is derived from the subjective attitude of the agent closest to it, i.e. the sufferer, the question is whether the sufferer's belief that she has an agent-neutral reason to end her pain gives us any reason to think that she does in fact have such reasons. Stuart Rachels has suggested that since we *do* take introspection seriously as regards questions of *other* normative properties of pain (badness, severity, etc), we should also take introspection seriously as regards the question of agent-neutral reasons.[32] On the one hand, it does seem plausible that the subject of the experience is best placed to tell us about this experience; that people who are suffering tend, almost inevitably, to see their suffering as providing agent-neutral reasons for its cessation; and that our best understanding of suffering is from a first-personal point of view, 'from the inside'. On the other hand, it is easy to find an error theory for sufferers' beliefs about the reasons generated by their suffering. Not only do people often believe what they want to believe, but it is often very useful for us to hold – and express – beliefs that others have reasons to help us. And insofar as introspection is to be accorded weight, it seems that we have to trade off the sufferer's introspected views against the introspected views of other agents. If the majority of other agents, upon introspection, deny that they have any reasons to help the sufferer, then presumably the weight of introspected evidence counts against the sufferer (although the same error theory may apply here). So we should not think that these reasons exist merely because agents think that they do. Nonetheless, since it does seem to most, if not all, sufferers that their suffering generates agent-neutral reasons, and since sufferers *are* in a position of authority compared to (mere) observers, we have reason to take this appearance seriously. The burden of proof, here, is on the theorist who denies the existence of agent-neutral reasons.

4.3. Searle

In a similar vein, John Searle has argued that we are constrained by language to hold that any one person's pain provides reasons for everyone. His argument, in brief, is as follows:

(1) If I say, "I am in pain", I am committed to holding that anyone in such conditions would be in pain.

[32] Rachels 2002, esp. 205-208.

(2) My pain gives rise to a need: I say “I need help because I am in pain”.

(3) I believe that my need for help is a reason for you to help. If I say, “Because I am in pain and need help, you have reason to help me”, I am committed to holding that, were anyone in such conditions, I would have reason to help them.

(4) But because (2) and (3) generalise, I am committed to holding that if *you* are in pain, I have a reason to help you.[33]

Insofar as I say things like “I am in pain”, I am committed to holding that certain things are true of anyone, simply in virtue of the fact that these claims generalise. They generalise, because ‘[w]hen you make an assertion of the form *a* is *F,* rationality requires that you be able to will that everyone in a similar situation should assert that *a* is *F*.’[34] If I claim that my pain provides reasons for you to help me, then I must be able to will that everyone in a similar situation should assert that their pain provides reasons for others to help them.

However, Julian Baggini has objected that (3) is ambiguous: it might be that my belief concerns ‘a reason *for you* for you to help me’, on the one hand, or ‘a reason *for me* for you to help me’[35]. Baggini proceeds to offer a pair of examples. Here is one:

‘I have entered a painting competition and know that if I win, I will appreciate the prize much more than my rivals. That gives a reason *for me* for the judge to choose me as the winner. But it is not a reason *for the judge* for the judge to pick me: the only reasons operative for him are those concerning artistic merit.’[36]

Searle’s universalisation argument will only work if my belief concerns a reason *for you* for you to help me. But I am only constrained to think that there are reasons *for me* for you to help me. At least, so Baggini claims.[37]

[33] Searle 2001: 161-162. This is strikingly similar to Clarke’s claim that “Whatever I judge reasonable or unreasonable that another should do for me: that by the same judgement I declare reasonable or unreasonable that I should *in the like case* do for him” (see Sidgwick 1930: 384-5). Similarly Sidgwick (1930: 209) claims that, ‘[i]f . . . I judge any action to be right for myself, I implicitly judge it to be right for any other person whose nature and circumstances do not differ from my own in some important respects.’

[34] Ibid. 159.

[35] Baggini 2002: 450.

[36] Ibid.

[37] Baggini also claims that Searle’s argument could only yield the existence of *reasons*, and not of moral obligations. I suggest, however, that this is innocuous: realist morality need only yield the existence of reasons, not of obligations. I will omit discussion of Baggini’s second objection.

There is, I think, something peculiar about Baggini's proposed disambiguation. Specifically, Baggini is committed to the following structural claim about reasons:

S: A reason R may exist at time *t,* with strength *z*, for agent *x* for agent *y* to Φ

Which is peculiar, because we tend to think that reasons are either reasons that Φ be done, or for agents to Φ, but not 'for agent *x* for agent *y*' to Φ. There are at least two further unusual features of an analysis such as S. Firstly, it multiplies the number of reasons beyond any necessity. For my neighbours, I have certain reasons; for my tutees, I have other reasons; for my employer, I have yet other reasons; and so on. Note that this is not simply a matter of any reasons I might have to promote others' self-interest. Secondly, it is entirely unclear what these reasons amount to. One might, reasonably, think that a reason for A to Φ is a consideration which counts in favour of A's Φ-ing. But these reasons also have to be considerations which A could *act* on, or at least incorporate into her reasoning.[38] In the example of the painting competition, it would be a good thing *for me* for the judge to choose me as the winner. But in what sense is this a (practical) reason? Well, it might be a consideration which the judge could act on. Alternatively, it might be a consideration which I could act on (it might count in favour of my bribing the judge, for instance). But the phrase 'a consideration for me which the judge could act on' is wrongheaded: if the consideration could be normative for the judge, then it must also be a reason for the judge (not 'for me for the judge' – whatever that might mean).

So I think that Baggini conflates beliefs about reasons for action with beliefs about benefits. It might be a benefit for me for the judge to give me the prize (*that* is a well-formed sentence), but it could not be a *reason* for me for the judge to give me the prize.

A more serious problem, however, is that premisses (2) and (3) are not necessarily true, and, indeed, often false. Even if most people do think as Searle describes, there might nonetheless be those who are sufficiently secure in themselves (or simply sufficiently bloodyminded) such that they do not treat their own pains as generating needs, or reasons for others to help. Sidgwick levels this objection against Kant's arguments for benevolence:

'We can certainly conceive a man in whom the spirit of independence and the distaste for incurring obligations would be so strong that he would choose to endure any privations rather

[38] C.f. the discussion of the epistemic and motivational constraints in Chapters One and Two.

than receive aid from others. But even granting that every one, in the actual moment of distress, must necessarily wish for the assistance of others; still a strong man, after balancing the choices of life, may easily think that he and such as he have more to gain, on the whole, by the general adoption of the egoistic maxim . . .'[39]

The existence of such men is not without precedent. Nor does there seem to be any strict inconsistency in holding that one ought simply to pursue one's own ends, at least given the supposition that the universalisation of this maxim would be to the strong man's benefit. In a similar vein, not everyone believes that needs create reasons for others to provide assistance. One might have a (somewhat peculiar) conception of 'needs' according to which a person has a need if and only if failing to meet that need will be severely detrimental to her welfare. People, for instance, need food, water, social contact, and so on. But it does not follow from this that each person has a reason to help every other person.[40]

Now I do think that these observations cast serious doubt on universalisation arguments of the kind given above. But we should, as suggested earlier, be wary of the thought that universalisation will ever yield distinctively moral imperatives. Even were it the case that the strong man was, by some rational imperative, committed to holding that he had reason to help others *on pain of inconsistency*, a failure to acknowledge these reasons would not be a simple matter of inconsistency (although that inconsistency might offer one way of persuading the strong man to act appropriately). Inconsistency is not a distinctively moral failing; cruelty, or the wanton infliction of gratuitous suffering, is. The failure of universalisation arguments casts doubt on approaches such as Searle's, although the conditional claim – that *if* I claim that my pain provides reasons for you to help me, then I must be able to will that everyone in a similar situation should assert that their pain provides reasons for others to help them – still holds, and shows how agents willing to make some fairly minimal claims about reasons are thereby committed to a substantial position regarding the value or disvalue of certain mental states (for instance). The incompleteness of Searle's argument, however, should not trouble the moral realist.[41]

Bear in mind that the task for the moral realist is to give some account of what it is that the obstinately callous person (or Sidgwick's 'strong man') fails to appreciate. If the strong man can

[39] Sidgwick 1930: 389.

[40] I will return to this line of argument later, in the discussion of the 'rational sadist' (section 5.3).

[41] After all, we hold that there is a right answer to the metaethical issue of whether Sidgwick's 'strong man' has certain reasons for action, even for philosophers who obstinately deny opposing viewpoints.

avoid thinking that his own pains provide reasons for others to help, then he can avoid thinking that he has reason to relieve other people's pains. But this is beside the point, on two counts. Firstly, the key issue is not whether we can avoid thinking that we have certain reasons, but whether we do in fact have these reasons for action. Secondly, the reason that we have for relieving other people's pains is not dependent on the reasons that other people have. It is related: their reasons for relieving our pains are, on some level, identical to our reasons for relieving their pains; they have a common source, namely the character of the sensation itself.

The strong man, that is, ought to think that he has certain reasons for action. What is it that the strong man fails to appreciate when he denies that he has such reasons? This question is best answered by considering what it is to respond to these reasons appropriately. At least one way in which we come to an awareness of moral value and disvalue is by exercising our empathic capacities; by seeing how it feels from someone else's point of view. This, at any rate, gives a plausible account of how we come to believe that the pain of others is disvaluable. This in turn generates a candidate account of where the strong man goes wrong: he exhibits an empathic failure. I will discuss this in more detail in Section 7.1. The key point, here, is that we can show what kind of failing Sidgwick's 'strong man' is guilty of, and can do so in a way which does not amount to a merely stipulative attribution of failure. Empathy is, I take it, a capacity rather than a form of preference. In this respect, the requirement that adequately functioning moral agents possess full empathic capacities does not amount to the mere requirement that adequately functioning moral agents possess moral motivations; *that* would fail to give an informative account of how Sidgwick's 'strong man' falls short. Of course, it is not obvious that empathic capacities themselves necessitate any particular direction of motivation. I consider the possibility of 'rational sadism' – that is, the rational infliction of pain by someone with intact empathic abilities – later, in Section 5.3.

It is worth noting that this model aligns neatly with McDowell's 'secondary qualities' model of values. A secondary quality is 'a property the ascription of which to an object is not adequately understood except as true, if it is true, in virtue of the object's disposition to present a certain sort of perceptual appearance . . .'.[42] The attribution of disvalue to a state of affairs is only to be understood in terms of our response to that state of affairs (under certain ideal conditions). So the attribution of disvalue to pain is to be understood in terms of our judging it to be disvaluable; this involves, *inter alia*, generally acquiring a motivation to minimise instances of pain. But there is a

[42] McDowell 1998: 133.

perfectly natural faculty – that of empathy – which both provides an understanding of others' mental states, and is able to supply the relevant motivation. McDowell also points out that coming to appreciate the presence of an external reason need not involved a transition via correct deliberation, but rather a transition *to* correct deliberation. Given this, we can give sense to talk of the required reasons in the case of Sidgwick's 'strong man', and hence to the charge of a rational failing insofar as he fails to appreciate these reasons: '[t]here would now be no question of a bluff, any more than one need to be bluffing if one says, to someone who cannot find anything to appreciate in . . . twelve-tone music, "You are missing the reasons that there are for seeking out opportunities to hear this music."'[43] Again, I will return to the claim that empathy is conceptually linked to the relevant motivations in Section 5.3

5. Clarifications and Objections

5.1. Symmetry

I have suggested that value attaches to pleasure, and disvalue to pain. On this simple reading, pleasure and pain are to be understood as opposite: if pain is simply whichever sensation feels distinctively bad, then pleasure is whichever sensation feels distinctively good. Let's call this the Symmetry Thesis:

ST: Pleasure has the same formal properties as pain, but at least one of its qualitative properties is reversed (e.g. both are non-intentional mental states, individuated by phenomenal quality, but one rationally justifies its pursuit, while the other rationally justifies its avoidance).

It is not clear, however, that the Symmetry Thesis is true. For instance, pleasure often takes an intentional object, whereas pain need not do so: I may be pleased that I have successfully poached an egg, but the anvil on my foot causes me pain. Pain has a location – in my foot, for instance – but not so with pleasure. Perhaps, then, pain is a sensation whereas pleasure is some kind of pro-attitude towards sensations, events, or states of affairs.

As far as the hybrid theory is concerned, it is important that the bearers of realist value are symmetric. It is important for two reasons. Firstly, the intrinsic nature of the bearers is supposed to explain why it is that value attaches to them. The disvalue of pain is to be explained partly in

[43] McDowell 1995: 78.

terms of *how pain feels*. Secondly, the bearers are to be identified partly in virtue of their extrinsic features – the extent to which they can rationally justify action, come to motivate agents, and so on. Now it might be that pain and pleasure, whilst different, still occupy opposite positions when it comes to justifying actions, or bearing value: that is, the Symmetry Thesis might turn out false, but pain would nonetheless be intrinsically disvaluable, and pleasure intrinsically valuable; that an action would lead to pain would rationally justify the avoidance of that action, etc. The problem with this view is that it renders the higher-level symmetry inexplicable.

Furthermore, remember that part of the specification for the value-bearers was given in terms of 'intrinsic to-be-pursuedness' and 'intrinsic to-be-avoidedness'. The states identified as the bearers of value, then, ought to be symmetric in this respect. It had better not be, then, that pain only contingently possesses this feature – as it would do, for instance, if pain were only extrinsically unpleasant. This is a live possibility: it has been reported that certain lesions to the prefrontal lobes leave subjects able to feel pain, but entirely indifferent to its presence – although this claim has been disputed.[44] Stronger cases are those of 'pain asymbolia', where patients not only display complete indifference to reportedly painful stimuli, they also often appear to find these stimuli entertaining, responding with laughter.[45]

5.2. Indifference

It does seem natural to think that pleasure (in and of itself) is a good thing, and pain a bad thing. But it would be hasty to conclude that these connections are necessary, rather than merely observations of general regularities. The experience of being in pain, in particular, need not always be unpleasant. Even normal subjects may be indifferent to very mild pains – such as when muscles become stiff after strenuous exercise. Furthermore, the terms 'pleasure' and 'pain' both extend to cover a wide range of states: we experience pleasant sensations (or 'sensory pleasure'), are pleased at finishing a piece of work, undergo painful operations, are pained by our failures, etc. Not only do the subjects of the experiences vary (good food, success, music, may all be sources of pleasure), but the subjective nature of the mental states also varies: pleasure may dominate our consciousness, be acute, mild, or – on some accounts – entirely devoid of phenomenal feel.[46]

[44] See e.g. Dennett 1978: 220-222; Elithorn, Glithero & Slater 1958.

[45] Grahek 2007, esp. 30-35.

[46] See Feldman 1997: 104 & *passim.*

The natural conclusion to draw from the lobotomy and pain asymbolia cases, I think, is that pain need not always be disvaluable – since if we are genuinely indifferent to its presence, it doesn't matter whether we feel it or not. What is missing is an aversion to the sensation: were the patients averse to the sensation, then there would be reason for them (and us) to diminish that sensation. So we may distinguish between 'mere' sensory pain, on the one hand, and suffering, on the other, where suffering is to be understood as a combination of a sensation and a certain kind of aversion to that sensation. Following Feldman, let us clarify that this aversion needs to be 'intrinsic' – i.e. an aversion to the sensation *for what it is*, rather than for some other reason. This rules out, for instance, being averse to the pleasant taste of chocolate because we know that it will lead to obesity. I leave the phenomenal nature of this aversion open for the time being.

There is an immediate worry here, that if the key property is 'intrinsic aversion', rather than the sensation itself, then *any kind of sensation whatsoever* may give rise to this aversion – even sensations which are generally considered pleasant, such as the sensation of a cool breeze on a hot day. But this, I think, is an advantage, rather than a problem, for the theory. There are those who – although rare – have peculiar tastes, or pathological aversions to certain stimuli. For these people, their lives would go much better if they managed reliably to avoid whichever normally-pleasant sensations they are averse to.

Now it does seem slightly curious to allow that the 'brute' sensation may detach from our response to it. There is a great deal of evidence – both from introspection and from more rigorous sources – for the existence of top-down as well as bottom-up processing for 'brute' sensations. That is, the quality of a 'brute' sensation may vary according to how averse we are to the sensation, what we expect of the sensation, and so on. I am not, therefore, claiming that sensations, on the one hand, and our responses to them, on the other, are entirely independent. What I am claiming – and this seems to me to be fairly plausible – is that in certain abnormal cases the relation between the sensation and the response may be non-standard, but that the sensation and response may nonetheless be identifiably similar to standard sensations and responses. Pain asymbolia patients, for instance, do indeed feel pain, but are simply indifferent to it (as opposed to feeling pain*, rather than pain). Yet there is a strong intuitive resistance to the thought that, in the case of such patients, there is no particular reason to refrain from torturing them (provided that it doesn't cause them physical harm, that is). It isn't acceptable to inflict pain on others simply because they don't mind it (although this might be less obvious in the case of the masochist who actively welcomes certain kinds of pain). I suspect, however, that this intuition is mistaken, although warranted. The entirely justifiable reluctance to inflict pain on others is unlikely to be sensitive to such exceptional cases as those of patients with pain asymbolia.[47] Equally, we should expect it to be difficult to empathise with such exceptional cases. In order to see the world from the point of view of someone with, for instance, a severe phobia of certain colours, we need to work very hard - and even then, it is unlikely that we will be particularly successful.

Similarly, one might raise the following concern. I have suggested that empathy provides a route to moral knowledge: by putting ourselves in the others' shoes, we can see that there is (agent-neutral) reason to relieve their pain. But if the brute sensation detaches from its unpleasantness, such that there can be pleasant (or at least non-aversive) instances of pain, then empathy may not give us an accurate indication of the reasons that there are. The relevant question, here, is whether empathy is a matter of understanding the other's brute sensations, or whether it is a matter of putting ourselves in their shoes. After all, what we need to know is how to treat other people. Understanding that, for instance, lobotomy patients are not averse to pain may give us less reason to refrain from harming them. The key question is not what brute sensations they are experiencing - after all, patients with pathological aversions to certain stimuli may be experiencing perfectly standard 'brute sensations' - but whether they are averse or amenable to these sensations. Empathy may allow us to understand that a stimulus which would be amenable to us is, for them, deeply unpleasant (or *vice versa*).

The upshot of the present discussion is that we should, for the purposes of the hybrid theory, talk of 'pleasure_P and suffering', where pleasure_p is to be understood as a mental state involving some kind of pro-attitude, and suffering is to be understood as a mental state involving some con-attitude. This proposal mirrors Feldman's – unsurprisingly, given the shared theoretical motivation (viz., hedonism). If it is possible for there to be pains to which we are not averse – and which are not disvaluable – then pain *cannot* be intrinsically disvaluable (at least, not on the strict understanding of intrinsicality at use in this thesis, according to which an object or property X is intrinsically F if and only if at every possible world W at which X exists X is F), since to be intrinsically disvaluable requires being disvaluable in each and every situation in which pain

[47] Clinical researchers may constitute an exception in this instance.

arises.[48] It also locates value in the complex experience of enjoying whichever sensations, facts, or possibilities we are experiencing, contemplating, or imagining, rather than in those sensations, facts, or possibilities themselves.

This, of course, leaves open the question of what pleasure$_P$ and suffering are. There is a possible problem here: one might think, with Feldman, that some cases of pleasure$_p$ are phenomenologically inert. There is simply nothing that it is like to experience some cases of propositional pleasure. Feldman, for instance, claims that 'propositional pleasure is not a feeling . . . from the fact that someone is taking propositional pleasure in some fact, it does not follow that he is experiencing any pleasant feelings.'[49]I may be pleased that the war in Bosnia has ended, but this does not mean that I am experiencing any pleasant sensations. Although there is a sense in which this is certainly true – I can be pleased that some fact obtains without experiencing any pleasant sensations – there is also a sense in which it is false. Provided that we assume that 'being pleased at . . .' is a conscious experience, we must assume that there is something that it is like to be pleased at. Pleasure, in the sense that the hedonists are interested in, should be understood to be a conscious experience – otherwise, we don't have much reason to think that it is valuable. Even supposing that Feldman were right, and that 'being pleased at x' is a state which, like 'hoping for x', may be unconscious, that kind of pleasure would not interest the hedonist. The key intuition which is being traded on here is that undetectable goods are of no value to the recipient, and hence of no value simpliciter. So the relevant notion of 'being pleased at' is one of 'being consciously pleased at', or 'being pleased at in a way which impinges on our consciousness' (I take it that being pleased at something might have a noticeable impact on our conscious experience, even if we are not aware of being pleased as such). And in *that* case, although being pleased at something is not a matter of having a pleasant sensation, there is something that it is like to be in this state – something distinctively pleasant.

Even if we allow that there may be instances of pains to which we are indifferent – or perhaps even pleasurable pains – there does seem to be some conceptual connection between pain and unpleasantness (or suffering). This connection can be brought out by some now-familiar Twin-Earth considerations. On Earth, it is generally true that individuals who feel pain perceive it as unpleasant, wish to avoid it, and so on. The inhabitants of Sensation Twin-Earth, on the other

[48] It is possible to understand the notion differently: as Feldman suggests (Feldman 1997: 140 fn. 29), height is an intrinsic property of persons, but any given individual might have had a different height. But I will leave this issue to one side.
[49] Feldman 1997: 143.

hand, claim to feel pain but display indifference to it, whilst being alike Earthlings in every other important respect. For the abnormal cases on Earth – lobotomy patients who feel pain but are indifferent to it, instances of pain asymbolia, etc. – we can maintain that they are competent users of the term ‘pain’ in virtue of their relation to their pre-lobotomised selves, or to others. On Twin Earth, however, this is not possible: we should not, therefore, be willing to treat Twin Earthlings as being competent users of the term ‘pain’. One might think, then, that there is a conceptual connection between pain and unpleasantness such that, necessarily, for any given community of language-users, it is generally true that pain-sensations are unpleasant. This, in turn, would go some way towards explaining why pain and pleasure are naturally – if erroneously – thought of as opposites.

5.3. Rational Sadism

The case of rational sadism poses a serious concern for the account on offer. A rational sadist is an agent who causes pain precisely because it is unpleasant to the sufferer, but appears to be entirely rational in all other respects, and appears to display competent mastery of the concept ‘pain’. When the victim complains that the pain is too awful to bear, the sadist replies that this is precisely the reason for inflicting the pain in the first place. The problem is that the sadist appears to take the qualitative nature of the sensation to count in favour of, rather than against, causing that sensation. On the account being developed, however, there is a conceptual connection between pain and disvalue such that the fact that someone is in pain suffices to provide reason to alleviate that pain – and we become aware of this reason by coming to understand that they are in pain. An increased capacity for empathy should make us more, rather than less, vividly aware of these reasons. But the rational sadist uses *her* capacity for empathy to make the pain more, rather than less, severe.[50]

The term ‘rational’ is, here, problematic. Let us understand ‘rational’ in this context to mean ‘devoid of procedural irrationality’. Perhaps the sadist’s preferences are, in some sense of the word, irrational – but that is beside the point. The worry is that, on the account being given, the qualitative nature of unpleasant conscious states provides reason to avoid causing them, and these reasons are inaccessible to the rational sadist; but it was supposed to be our *empathic* and *rational* capacities which show us that the nature of these states provides reason to avoid causing them. The rational sadist has intact empathic faculties *and* intact rational faculties. And if our concept of

[50] C.f. McNaughton 1988: 140-144.

pain (or unpleasantness) is intrinsically normative, in that it comes with attached ‘to-be-avoidedness’, then the rational sadist turns out not to be using the same concept as us.[51] That seems implausible.

In relation to this topic, I wish to make three points. Firstly, it is question-begging to suppose that the sadist is rational in the sense of being fully responsive to all reasons present, although equally question-begging to suppose that she is not: whether or not she is fully responsive to all reasons present is precisely what is at issue here. Secondly, the question of whether or not the sadist is genuinely rational may be something of a side issue: being fully rational need not entail being responsive to all of the (agent-neutral) reasons present. It will help, here, to distinguish between two senses of rationality, one which is merely a matter of internal consistency (call this the ‘purely procedural’ conception) and one which requires responsiveness to certain reasons (the ‘substantive’ conception).[52] The sadist, then, is fully rational in the ‘purely procedural’ sense; this is uncontroversial. But in the substantive sense, she fails to be fully rational, since she fails to respond to certain reasons for action. The fact that the sadist is acting consistently does not, by itself, show that there are no reasons for her to refrain from acting as she does. Thirdly, even if there is a conceptual connection between pain and disvalue, this does not entail that all agents need perceive pain as disvaluable: as with the motivational constraint discussed in Chapter Two, the connection here may be governed by a ‘necessarily, generally’ operator, such that it need only be the case that agents *generally* perceive pain as disvaluable.

Now it does seem that, in actual cases, empathy is dependent on performing some kind of internal modelling and re-enactment.[53] If empathy requires simulation of others’ mental states, therefore, rational sadism seems less plausible. On the other hand, if attribution of mental states to others is a matter of theorising, rather than simulation, there is clear scope for rational sadism. I do not have the space here to adjudicate over the debate between simulation theory and theory theory here, however.[54] I discuss the issue of empathy in more detail in Section 7.2.

As regards the issue of rationality, consider Searle’s ‘universalisation’ argument outlined above. If Searle is correct, then the sadist may open herself up to rational criticism. For if she, when in pain, believes that others have a reason to help her, then she is committed to thinking that she has

[51] This chapter, Section 3.5.

[52] See Hooker & Streumer 2004. I take it that both senses of ‘rational’ are at work in common parlance.

[53] See e.g. Batson et al. 1987; Bavelas et al. 1987.

[54] But see Arkway 2000; Nichols & Stich 1993; Goldman 1992.

a reason to help others who are relevantly situated. Of course, a particularly hard-nosed sadist may deny that others have reasons to help her; perhaps she has conditioned herself such that she does not resent being in pain, does not believe that it creates a need on her part, and so on. People like this are few and far between, if, indeed, any exist. Indeed, I suspect that a world full of such agents would be so unlike our own world that their concept of 'pain' would be distinct from ours. In particular, they would have to think that the unpleasant character of pain gives reason to cause it in others. This returns us to the now-familiar 'Twin Earth' cases which have so far been adduced in relation to internalism: if the reason for causing these states in others really is their intrinsic character, as opposed to the delight which their infliction causes to the sadist, this fact would be inexplicable.[55] Perhaps some extrinsic reason – a desire to harm others, for instance – could explain the sadist's behaviour, but not the qualitative nature of their victim's suffering as such.

Irrespective of whether sadists can be shown to be internally inconsistent, however, I suspect that these cases are something of a side issue. This is partly because not all normative reasons need provide motivating reasons, or even be responded to by all agents. Rather, as the discussion of Chapters One and Two indicated, the epistemic and motivational constraints apply to groups of agents, rather than individuals. But even if rational sadists do exist (in the strict sense, taking there to be no intrinsic reasons to avoid promoting others' suffering), their existence need only be troubling insofar as it is widespread. There are parallels here between practical and theoretical rationality: it is troubling for any theory of normative reasons if these reasons turn out to be epistemically inaccessible, but surely a merit for the theory if it allows that we may, on occasion, be sufficiently obtuse that we are blind to these reasons.

Suppose, further, that we could show sadists to be irrational. What would follow from this? Insofar as the sadist cares about being rational, she can thereby be brought to see that she has a reason to refrain from harming others. But she might equally delight in being inconsistent (some Zen philosophy seems to do precisely this), or simply refuse to admit that there is a reason not to be inconsistent. So again, even if there are cases where the subject is simply blind to the relevant reasons, this does not show that such reasons do not exist. Nor should we think it a constraint on our metaethical theory that we must be able to convince all persons in all contexts that they have reasons to, for instance, refrain from causing harm.

[55] Since sadists commonly inflict pain as a means to sexual gratification, etc.

5.4. Desert

A stock objection to naive hedonism (according to which all pleasurable experiences are *pro tanto* valuable) is that pleasurable experiences are only valuable when they are *deserved*. One might even think that pleasurable experiences are only good *for their subject* when they are deserved. Kant, for instance, held this view, arguing that we fail to treat people as ends in themselves when we fail to punish them, because in doing so we fail to give them their due. There are at least two distinct concerns here. The first is that the valence of pleasure (or whichever states the hedonist invokes) is not intrinsic, as it is conditional on an extrinsic factor (desert). The second is that desert itself may be inexplicable on any account which takes consequences to be morally fundamental. Utilitarians, for instance, have traditionally had problems accounting for the phenomenon of moral desert: for instance, if what we ought to do is purely a matter of what would maximise overall utility, then there is no room for other considerations (i.e. desert) to impact on what we ought to do.[56] I will assume, for the sake of argument, that a moderately sophisticated rule-utilitarian approach – on which desert-involving practices are justified by reference to utility-maximisation, even where individual acts mandated by the practice need not themselves maximise overall utility – can deal with the second problem[57]. However, the first objection needs to be addressed: even if utilitarianism leaves room for the thought that we should treat people as they deserve, we still need to account for the worry that pleasure is only valuable when it is deserved.

There is one way in which pleasure might be intrinsically valuable, yet only conditionally yield reasons for action, and that is if the following analysis holds:

A: P is intrinsically morally valuable if and only if the intrinsic features of P provide *prima facie* agent-neutral reasons for promoting P

A reason is *prima facie* if it ceases to exist when overridden; otherwise, it is *pro tanto*. Although there is not a consensus on whether moral reasons are *pro tanto* or *prima facie*, it seems highly plausible that values are to be understood as providing (or consisting of) *pro tanto* reasons.[58]

[56] See Feinberg 1970 esp. 75-94.

[57] See, e.g. Hooker 2000; Hare 1981

[58] In particular, see Kearns & Star 2009 for the claim that *all* practical reasons are evidence that one ought to φ, and hence *prima facie*.

After all, intrinsic values are supposed to possess their normative status irrespective of how the rest of the world is. And construing the resulting reasons as *pro tanto* does indeed give a natural way to explain 'rational regret' – the fact that we may justifiably regret the fact that, although we caused the best outcome, we failed to bring about some other outcome.[59]

In any case, *prima facie* reasons are best suited to modelling epistemic conflicts, rather than moral conflicts. Moral values, whatever else they may be, are to be maximised; this requires that they can be agglomerated, and compared. The *pro tanto* model captures the thought that the (moral) reasons make it the case that one ought to φ, rather than simply serving as evidence for the claim that one ought to φ.[60] But to decide that the balance of reasons stands in favour of one course of action does not entail that there is no longer anything to be said in favour of another course of action. If we are faced with two patients, one of whom is in severe pain, and one of whom is in mild pain, but we only have one anaesthetic available, it is clear that we should give it to the patient in severe pain – but this does not mean that the reason for giving it to the patient in mild pain (that it would relieve this pain) ceases to operate.

Furthermore, as Susan Hurley points out, conceiving of values as supplying *pro tanto* reasons yields a way of making sense of *akrasia* – cases where an agent knowingly acts contrary to what she considers herself to have best reason to do.

'The capacity of *pro tanto* reasons to influence what an agent does is no more exhausted by their contribution to his deliberated all-things-considered evaluations than the capacity of interest groups to influence what a democratic society does is exhausted by their contribution to its government and laws: each may go on trying to get its way in the face of legitimate authority.'[61]

Akrasia, on this account, is not a matter of simple insanity, but can be understood in terms of 'familiar though conflicting subsystems'.[62] When I judge (correctly) that what I have all-things-considered reason to do is to generate prose, but instead choose to attend a garden party, it is not that I am irrationally choosing to act on a nonexistent reason. Rather, my rational appreciation of a *pro tanto* reason for attending the garden party manages to derail my rational appreciation of a

[59] See Williams 1966. There are, however, other ways of making sense of rational regret: we could do so on a rule-utilitarian account, for instance.
[60] Broome 2008 esp. 100-103.
[61] Hurley 1989: 137. See also Hurley 1989: 170-173.
[62] Hurley 1989: 162.

pro tanto reason for generating prose. Accounting for judgements about reasons as judgements of reasons which persist even when overridden, here, allows us to make sense of these akratic failings.

5.5. Particularism

Ethical particularism seems to pose a threat to the particular form of moral realism on offer, although not to realism *per se*. I take the canonical formulation of ethical particularism to be that provided by Jonathan Dancy:

Particularism: 'The possibility of moral thought and judgement does not depend on the provision of a suitable supply of moral principles.'[63]

Dancy begins by dispensing with the 'subsumptive' model of morality, according to which specific moral truths derive their status from being subsumed under general moral principles. So whereas we might commonly think that an individual lie gets to be wrong by falling under the general principle 'lying is wrong', Dancy denies this claim. In this respect, my formulation of moral realism agrees with Dancy's particularism: the claim that suffering is bad is true *because* individual instances of suffering are bad. Here, the general ethical principle serves to describe the individual cases. But the motivation behind Dancy's particularism is hostile to my formulation. Dancy is a holist about reasons:

Holism: 'a [non-moral] feature that is a reason in one case may be no reason at all, or an opposite reason, in another.'[64]

So, for instance, the fact that an action would cause another agent pain may be a reason to refrain from that action in one case (the standard case), but a reason to engage in that action in another case (where we wish, justifiably, to punish someone). The fact that an action would raise the temperature of the fridge is a reason to engage in that action in some cases, and not in others. Dancy claims that '. . . nobody has ever really debated the question whether ordinary practical reasons are holistic or not . . . the examples, which are legion, should be allowed to carry the day

[63] Dancy 2004: 5. Of course, we might think that all moral truths derive from some set of principles (the Decalogue, for instance) but that a particularly saint-like agent could fine-tune her moral judgement to such an extent that she no longer needed the principles. This would still be a form of generalism.
[64] Dancy 2004: 75.

without resistance.'[65] And there is, he thinks, good reason to think that all reasons behave alike; it would be surprising if one subdomain of rationality turned out to have its own particular logic.

I do not, here, have time or space to do justice to Dancy's arguments. However, I do want to indicate how the form of moral realism on offer here (specifically, one which takes hedonic or anhedonic states to provide agent-neutral reasons for action) is compatible with the thought that holism does accurately characterise practical reasons, at least at some level. To begin with, note that it is supposed to be a distinctive feature of the mental states relevant to hedonistic utilitarianism that they provide agent-neutral *pro tanto* reasons for action. Intrinsic value, strictly construed, entails the denial of holism, since the intrinsic properties of whatever it is that bears the value will not change across cases. Given this, it would be surprising if this subdomain of reasons did *not* turn out to have its own particular logic. Secondly, note that invariantism about moral value says nothing about whether our all-things-considered reasons are themselves invariant. Even if pain is invariantly disvaluable, we may still have all-things-considered reason to inflict pain on a deserving subject. When we claim that the pain-causing aspect of an action is reason to engage in that action (as in the case of punishment), we are claiming that this aspect of the action can explain why we ought to proceed. But that does not, by itself, entail that that aspect of the action does not also yield *pro tanto* reasons to refrain. That an action would cause me immense pleasure, for instance, may sometimes be a reason not to engage in that action (where I need to concentrate on a delicate task), but surely there is still something to be said in favour of the action.

In fact, the existence of *some* invariant reasons is something to which Dancy is quite amenable – 'so long as the invariance is not a matter of the logic of such reasons, but more the rather peculiar fact that some reasons happen to contribute in ways that are not affected by other features.'[66] Were such invariance to obtain as a matter of the logic of such reasons (that is, as a constitutive rather than an accidental property of these reasons), then we would, he thinks, be committed to a 'hybrid conception of rationality'.[67] This is not entirely accurate: even if rationality is to be understood as a matter of sensitivity to reasons, it is entirely possible for a single faculty to respond to commensurable but metaphysically distinct elements (some in principle variant, some in principle invariant). But what *is* worrisome is the thought that the distinction between variant

[65] Dancy 2004: 74.
[66] Dancy 2004: 78.
[67] Dancy 2004: 77.

and invariant reasons is a ‘brute fact’, hence inexplicable and peculiar. But when it comes to intrinsic value, this kind of contribution is to be expected, rather than being peculiar. The invariance of these reasons is a matter of the metaphysical status of these reasons. Whether or not this amounts to a matter of ‘the logic of such reasons’ is uncertain. In any case, limiting the scope of intrinsic value renders the theory as a whole much more plausible: the painfulness of a punishment may count towards engaging in that punishment *as far as justice is concerned*, but is disvaluable in and of itself.[68]

6. The Open Question Argument

6.1. The Problem

I now want to consider a stock objection to non-moral analyses of moral terms: Moore’s ‘Open Question’ argument (henceforth OQA). Moore thought that ‘good’ was an unanalysable term, since, given any proposed analysis of the form ‘F is the good’, the question ‘given that x is F, is x good?’ remains open:

‘The hypothesis that disagreement about the meaning of good is disagreement with regard to the correct analysis of a given whole, may be most plainly seen to be incorrect by consideration of the fact that, whatever definition be offered, it may be always asked, with significance, of the complex so defined, whether it is itself good.’[69]

So whereas the question

Given that x is good, is x good?

is closed – the answer is available to anyone who understands the terms being used – the question

Given that x is (insert proposed definition), is x good?

is open: it still makes sense to pose the question. Therefore the proposed definition, whatever it happens to be, fails. And this is a perfectly general point: whereas it makes sense to wonder

[68] Which, as I have suggested, gives an explanation of why the punishment is appropriate.
[69] Moore 1922: 15.

whether an outcome which meets the proposed definition is good, it does not make sense to wonder whether an outcome which is good is good. Given this, Moore concluded that ‘good’ denotes a non-natural, undefinable property.

Moore is widely accepted to have been wrong about the conclusion of his argument. The OQA applies equally to non-natural as to natural properties. Secondly, the OQA operates at the level of concepts, rather than properties. Thirdly, analytic truths need not be obviously analytic: complex mathematical proofs, arguably, are, if true, analytically so, but wondering whether their conclusion holds may still amount to a sensible question. Conceptual analysis may yield truths which are analytic, but where the argument is complex, the conclusion may not be obviously true. In any case, for the proposed theory to be a substantive, informative theory, we should expect these conceptual gaps: informative analyses lead to questions with ‘open’ feels.[70] Nonetheless, it is equally widely accepted that Moore was on to something.[71] Whichever property or concept we pick, it does seem that the claim ‘x is good’ is genuinely contentful. The question then becomes, what explains this appearance?

6.2. Diagnosis

Darwall, Gibbard and Railton, in their panoptic ‘Towards *Fin de Siècle Ethics*’, suggest that the openness in question ‘may stem from our seeming ability to imagine, for any naturalistic property R, clear-headed beings who would fail to find appropriate reason or motive to action in the mere fact that R obtains (or is seen to be in the offing).’[72] But moral terms, such as ‘good’, are action-guiding in a way that non-moral terms are not.[73] If I realise that an action would generate a well-balanced painting, that does not by itself provide me with a reason to engage in that action. But if I judge that such an action would be good, that does seem to provide me with a reason. So ‘good’ is normatively loaded in a way in which the proposed definiens will not be. The opennness of Moorean open questions, then, just is the space for wondering whether we really do have reason to promote the suggested property. In other words, the openness of such questions stems from a normative gap between ‘good’ and whatever is being offered as a definiens.

[70] At least at first: as we become convinced of the truth of the analysis, the question will begin to seem less open.

[71] See Altman 2004, Campbell & Woodrow 2003, Rosati 1995, Strandberg 2004.

[72] Darwall, Gibbard and Railton 1997b: 4.

[73] See also Bloomfield 2006.

This, however, is a slight mischaracterisation of the OQA. After all, it also seems that we can imagine clear-headed beings who would fail to find 'appropriate reason or motive to action' in the putative fact that X is good. This is, of course, a combination of a worry about rationalism in ethics (the question of whether if I judge that *x* is good I therefore judge that I have reason to bring about *x*), and internalism about motivation (the question of whether moral judgements bring with them motivation to act accordingly). But Darwall, Gibbard & Railton's analysis does not seem to provide room for clear-headed beings who could sensibly ask, 'given that x is good, do I have reason to promote x?', since they hold that the claims 'x is good' and 'I have reason to promote x' are similarly action-guiding. Nonetheless, it does seem plausible that wondering whether x is good may often amount to wondering whether there is reason to promote x.

This does not, of course, rule out the possibility of finding meaningful connections between whichever property is proposed, on the one hand, and goodness, on the other. After all, the OQA seems to equivocate between a *de dicto* and a *de re* reading (Moore certainly conflated the two).[74] Disambiguating these two readings gives, for any proposed definition (F) of goodness:

De dicto: It is an open question whether everything that instantiates F-ness is good

And

De re: F-ness is such that it is an open question whether everything that instantiates it is good

In other words, the Open Question argument tells us about relations between the concepts at hand, rather than the properties. In order to derive a metaphysical conclusion about the nature of the property denoted by 'good', we need the second (*de re*) reading. But the most that the OQA will deliver is the first (*de dicto*) reading.[75] So one possible response to the argument is to claim that, although the question remains open at the conceptual level, the property of F-ness can still be identified with the property of goodness. This strategy has been put into practice by the so-

[74] Kalderon 2004.

[75] Bloomfield (2006) claims that even this reading may be innocuous, as the open nature of the question derives (or so he argues) from a very general aspect of rule-guided behaviour. Following Wittgenstein, Bloomfield claims that the rule laid out in 'F-ness is good' is open-ended in much the same way as any other rule is; hence it is always open to us to ask, of any instantiation of F-ness, whether it is good. I find this thought plausible, but it is – as will become apparent – orthogonal to my response to the OQA.

called 'New Wave' moral realists, who claim that the relevant ethical definitions will be synthetic definitions – statements of property identity.[76]

This indicates one possible solution to the OQA. But we need to be careful about the nature of the analysis on offer. Bear in mind that the proposed realism takes the following form:

R: if an action will cause suffering, there are intrinsic, *pro tanto*, agent-neutral reasons to refrain from engaging in that action.

Which may also be phrased as,

R*: causing suffering is intrinsically bad,

or

R**: suffering is intrinsically disvaluable,

Where **R, R*,** and **R**** are taken to describe the same fact (since intrinsic disvalue, on this account, amounts to the provision of agent-neutral *pro tanto* reasons to refrain from its promotion). To adopt Scanlon's terminology, I hold a 'buck-passing account' of (moral, intrinsic) goodness, whereon the property of (intrinsic, moral) goodness is just the higher-order, formal property of possessing features which provide reasons in the specified sense. So claims about the intrinsic disvalue of suffering are at some level claims about the reasons for action which suffering provides, although not synonymous with those claims.

Hooker & Stratton-Lake claim that adopting this position 'rules out the possibility of simultaneously accepting that A is good and responding "so what?" . . .'.[77] But this conflates a claim about the meaning of 'is good' with a claim about the property of goodness. So whereas I maintain that for something to be good *just is* for it to have certain non-moral (reason-providing) properties, it is quite possible for someone else to deny this connection (at least, it is intelligible for them so to do). And given this, it is intelligible for someone to accept that A is good and respond "so what?" – this is a consequence of 'less weak internalism', **LWI**, as discussed earlier.

[76] See e.g. Boyd 1988.
[77] Hooker & Stratton-Lake 2006: 163.

What *would* be curious is for someone to accept that A is good, *and* to consciously accept a buck-passing account of reasons in relation to goodness, *and also* to respond “so what?”. The buck-passing account is compatible with the possibility of amoralism, since the normative buck is passed at the level of properties, rather than of concepts.

But what is the status of the claims **R**, **R***, and **R****? They are certainly not intended as an *analysis* of the term ‘good’; rather, the claim is that these hedonic or anhedonic states are the bearer of goodness, where this is to be understood as a matter of providing certain kinds of reasons. The OQA cuts against reductive forms of analytic naturalism, but the account on offer is not such an account. Nor is it intended to be an informative account of the *property* of moral goodness, beyond the claim that intrinsic moral goodness consists in the provision of *pro tanto*, agent-neutral reasons for action in virtue of the intrinsic properties of the obtained states. This is why **R** is left as a one-way conditional, rather than a biconditional (although it is consistent with the claim that suffering is the *only* source of such reasons).

However, we might still ask, in virtue of what (if anything) are **R, R***, and **R**** true? At the fundamental level, I have suggested, **R** is true in virtue of the phenomenology of the relevant states; these states provide reasons for action in virtue of the way that they feel. Insofar as this is true, **R** is a claim about the relation between two properties – the property of suffering, and the property of providing reasons for action. However, we should also expect that concepts with correlated real-world properties are informative about those properties. Hence the arguments in favour of **R** are primarily conceptual, and hence *a priori* rather than *a posteriori*. Given this, a modified ‘open question’ worry presents itself; that is, it seems that the question,

Q: Given that φ-ing would cause suffering, is there some *pro tanto*, agent-neutral reason to refrain from engaging in this action?

and the associated question,

Q*: Given that *x* is an instance of suffering, is *x* (intrinsically) bad?

are open – although it is not clear that these questions are *in fact* open. This is potentially worrisome, as it casts doubt on the claim that there is a conceptual connection between suffering and (moral) badness. If the two were conceptually linked, then questions such as **Q** and **Q*** should be closed, since mastery of the relevant concepts should allow clear-headed individuals to get clear on the relevant connections. But it is possible for **Q** and **Q*** to be genuinely open questions even if there is a conceptual connection between suffering and moral reasons, provided that this connection is 'loose'; that is, provided that the connection is such that necessarily, generally, agents who are competent users of the relevant concepts may come to treat such questions as closed.[78] Whilst Darwall, Gibbard and Railton claim that the open feel of such questions is a matter of our being able to imagine a *world* of clear-headed beings who see that an action causes suffering, but do not see any reason to refrain from this action, all that is needed to account for the open feel of these questions is the possibility of *an individual* who meets these conditions.[79]

Granted, it is logically possible to accept that φ-ing would cause suffering, but deny that there is reason to refrain from φ-ing. It is also conceptually possible that there exist clear-headed agents who understand that φ-ing would cause suffering, but deny that they have reason to refrain from φ-ing. But it is not clear what follows from this. Certainly, the possibility of a world of clear-headed, rational sadists, the inhabitants of which are competent users of the relevant concepts (concerning hedonic and anhedonic states, and reasons for action) but do not share our view of the connections between them, might be explained by the fact that the proposed analysis is incorrect. However, as discussed above, mastery of the relevant concepts does not itself entail that the relevant connections are clear. There are many cases where we may be clear-headed, competent users of concepts but fail to grasp the connections between them. Even given a correct analysis, there is still room to wonder whether that analysis is correct. For instance, one paradigmatically analytic truth is 'bachelor' means 'unmarried man'. We might ask whether a Bachelor of Arts must be an unmarried man. But even we treat the occurrence of 'bachelor' in this phrase as a homonym for 'bachelor' in the context of 'unmarried man', we might wonder about, for instance, young men below marriageable age. Understanding 'unmarried man' as elliptical for 'unmarried but marriageable man', however, does not end the inquiry; we might further wonder about gay men in committed long-term relationships (perhaps in a legally binding civil union). Presumably the word 'bachelor' would, in this context, be a misnomer. And in the individual case, I may coherently wonder whether particular aspects of my use of a term are

[78] C.f. Rosati 1995 esp. 49-51.

[79] Darwall, Gibbard and Railton (op. cit.), amongst others, hold that this diagnosis motivates noncognitivism. But it would be more accurate to suggest that it motivates moral judgement internalism – the claim that there is an internal connection between moral judgement and motivation. Whether noncognitivism is best placed to account for this connection is a further issue.

idiolectal (indeed, this often turns out to be the case, despite competent and clear-headed use of the relevant terms).

To summarise, my response to the Open Question Argument runs as follows. Firstly, the OQA itself does not pose a direct threat to the account on offer, as the OQA cuts against proposed analytic reductions, whereas the account addresses relations between properties. Secondly, the relevant questions – **Q** and **Q*** – are not as clearly open as the question, 'given that *x* maximises happiness, is *x* good?'.[80] Thirdly, and most importantly, even if **Q** and **Q*** are indeed open questions (in Moore's sense), we can make sense of this open feel without losing the conceptual link between suffering and disvalue. The response runs in part along the lines proposed by Darwall, Gibbard and Railton, and suggests that the open feel of these questions is a function of our being able to imagine clear-headed beings of a certain disposition.[81] The possibility of such beings is secured by the possibility of a weak form of internalism, according to which it is a constraint on possession of the relevant concepts *by a community* that its members, generally, take suffering to be reason-providing in the sense specified (and *vice versa*). And given that even clear-headed, competent use of the relevant terms need not make obvious the relevant conceptual connections, the Open Question Argument is inconclusive as regards the conceptual issue.

I now turn to how positing a weak form of internalism, as introduced earlier, allows the moral realist to meet the epistemic and motivational constraints outlined in the previous two chapters.

7. The Constraints Revisited

7.1. Epistemology

The epistemic constraint amounts to the denial of the possibility of unknowable moral truths. This posed a worry for naive moral realism, for at least two reasons. Firstly, there is no principled reason why the realist moral facts could not transcend our ability to appreciate them; secondly, the non-naturalist realist has a problem of accounting for the mechanism whereby we could detect

[80] There is a difference between the usage of 'good' as it occurs in this question, and the usage of 'intrinsically good' as it appears in my account: whereas 'good' in the first instance may entail 'what we ought, all things considered, to pursue', 'good' in the second instance entails nothing more than 'what we have some reason to pursue'. In the first case, the answer is not settled by the concepts involved; the second case is less clear-cut.

[81] Note that being able to *imagine* such beings does not entail that these beings are conceptually possible. I will avoid discussion of this issue.

these moral facts.[82] But note that the form of moral realism on offer here is quite different: it is a modest naturalistic form of realism. There is, therefore, no commitment to any (problematic) non-natural faculties. And the realism here is *modest*, amounting to a combination of modest mind-independence (independent of individual minds, rather than all minds), cognitivism, and descriptivism. Our talk of intrinsic value aims to locate states of affairs which, in virtue of their intrinsic properties, provide us with reasons for action. I have already argued, following Foot, that moral truths cannot peel apart from human interests (that is, there are anthropocentric constraints on value). The hedonist theory outlined up to this point is, I take it, a plausible sketch of a morality which is tied to human interests in the requisite way.

The phrase 'moral truths', here, encompasses two distinct elements. There is the claim (for which I have been arguing in this chapter) that pleasure is intrinsically valuable, and suffering intrinsically disvaluable, where 'intrinsically valuable' is to be interpreted as meaning 'provides, in virtue of its intrinsic properties, *pro tanto*, agent-neutral reasons for action.' This claim should be knowable, and a plausible epistemology for it should be supplied. But this extends to individual cases: agents should be able to detect the putative agent-neutral, *pro tanto* reasons provided by the hedonic (or anhedonic) states of others.

The epistemology which attaches to this account has, therefore, at least two elements; two routes via which we may come to knowledge concerning intrinsic value or disvalue. The first route is immediate: our capacity for empathy allows us to put ourselves in others' shoes, and to take the qualitative nature of their suffering to provide us with reasons to relieve that suffering. But is it true that the reason-providing nature of suffering might not be undetectable? In order for this to be the case, we would have to be unable to put ourselves in other peoples' shoes; we would be unable to empathise with them, or to construct a theory of mind which adequately represented what it is like to be another person. Snow suggests that S empathises with O's experience of emotion E if and only if: (a) O feels E; (b) S feels E because O feels E; and (c) S knows or understands that O feels E.[83] But there is a weaker sense of 'empathy' which elides (b), and where the core element is that S knows what it is like for E. Here, a well-developed imagination will suffice; simulation is not needed. I will assume the weaker reading, at least for the time being. Now although we should allow that individuals may be unable to empathise, but

[82] Quite how problematic a faculty for detecting non-natural facts would have to be depends, of course, on how we are to interpret 'non-natural'. But since the theory on hand is a naturalistic theory, I will leave this issue to one side.

[83] Snow 2000: 68.

nonetheless count as competent users of the relevant terms (in the parasitic sense discussed in relation to 'loose' conceptual connections, as with internalism about motivation), we should be, I think, unwilling to allow that *everyone* might have a concept of suffering without any understanding of what it is like for other people to suffer. There are two reasons for this. The first is straightforward: we can understand ourselves as creatures which respond in certain (mental) ways when confronted with certain stimuli, and just as we can induce that the application of hot irons will cause suffering to ourselves in the future (since our future selves are relevantly similar to our present selves), we can induce that other creatures relevantly similar to us will feel similarly when confronted with relevantly similar stimuli. This is a general consequence of a concept of suffering as a resultant property.[84] Note that these considerations do not require that moral knowledge will, as it were, force itself upon us, but only that there must be a route to moral knowledge (here, about the realist part of ethics) available.

There is a second, and more interesting, argument, which turns on considerations of the possibility of a private language. Following Wittgenstein, I take it that we should deny the possibility of a private language, at least for sensations. This is not to say that the *use* of language requires, in all cases, more than one person; we may, for instance, admonish ourselves, engage in soliloquy, etc.[85] But the meanings of terms are only fixed against a background of a shared linguistic practice; in that sense, the language is essentially communal.[86] Since what fixes the use of the term is communal, there must therefore be a shared usage of the term – and *that* entails that individuals must be able to understand what other people mean by their uses of the term. Consequently, each individual must be able to understand what it is like for others to experience the sensation picked out by the term; so insofar as I myself understand the term 'suffering', I must be able to understand what it is like for other people to suffer.

There is a second route to moral knowledge, at least as regards intrinsic value, and that is the direct conceptual route. We can tell, just by thinking about the matter aright, that there is something to be said (*per se*) in favour of actions which cause pleasure, and against those that cause suffering. So I hold that we can come to be justified in believing that suffering is

[84] That is, a mental term which picks out a property resulting from other, non-idiosyncratic properties; a property of sentient entities in general, rather than one particular sentient being.

[85] See Canfield 1996: 478 & *passim.*

[86] Else, in the case of naming our own sensations, we would have no way of understanding what it is to be mistaken in naming our sensations, and hence there would be no sense in which we were following a rule.

intrinsically disvaluable just by understanding the proposition.[87] But it is also the case – as the argument of this chapter has attempted to show – that we can come to be justified in believing that suffering is intrinsically disvaluable though inference. Since such inferences can be grasped (if not accepted) by all who possess the relevant concepts, they are appropriately accessible, and hence ensure that the claim 'suffering is intrinsically disvaluable' remains knowable.

There is a third route available, which I have not so far discussed, and that is the route of 'reflective equilibrium'. The broad picture, here, is that we take our considered judgements (whether of particular cases, general principles, or both), and attempt, by careful deliberation, to bring them into balance, discarding those that seem out of place or implausible, leaving us with a collection of (hopefully) stable, systematically-interrelated beliefs. This is, for instance, a method commonly used to justify utilitarianism.[88] Note that this does not amount to a coherentist epistemology; it is not that moral claims are true (or false) in virtue of their coherence with the network of other moral claims, but that the method of reflective equilibrium is our best way of engaging in first-order moral theorising. So this method *may* lead to moral knowledge, but there is no guarantee that it need do so.

7.2. Motivation

There are two questions regarding motivation: how can moral properties come to motivate agents, and what explains the connection between moral judgement and motivation? Evidently, one route by which moral properties could come to motivate agents is through forming veridical moral beliefs (coupled with some account of how beliefs can motivate agents, either on their own or coupled with some external desire to act accordingly). So the epistemology given above, insofar as it gives an account of how we can come to propositional moral knowledge, coupled with an account of how moral judgements can motivate, can explain how moral properties come to motivate agents.

However, the account of moral properties on offer here provides room for these properties motivating agents directly. The suggestion is that our capacity for empathy provides us both with knowledge of the value (or otherwise) of the hedonic states of others, and also supplies a route

[87] Audi 2004a provides an account of moral beliefs which are justifiable both inferentially and non-inferentially (i.e. by intuition).
[88] See Mulgan 2007: 57.

whereby this value can come to motivate us.[89] This contrasts with the putative badness of, for instance, lying – where if lying is bad (in the realist sense) it is hard to see how this could come to motivate us to avoid lying without going via a belief about the badness of lying. Bear in mind that for a state to be intrinsically disvaluable is just for it to provide *pro tanto*, agent-neutral reasons for minimising the occurrence of that state. I am suggesting that empathy provides a route to appreciating these reasons, and that appreciating these reasons involves acquiring motivations to act accordingly.

There is a deep problem here, which is that the exercise of our empathic capacities need not in itself entail the appropriate motivation. When engaging with fiction, for instance, persons often seem to empathise with the characters, but are not always motivated accordingly (although young children may shout advice to imperilled film characters).[90] In a more extreme case, we may imagine a torturer whose empathetic capacities make him particularly effective – he knows precisely what his victims like the least, and uses this insight to tighten the thumbscrews appropriately. The 'imaginative torturer' has full empathetic capacities, but sees no reason to refrain from making his victims suffer. One possible response, here, would be to suggest that whilst the case of the individual imaginative torturer is conceivable, there would be some difficulty in conceiving of a *world* of imaginative torturers. Perhaps whatever is picked out by their usage of the term 'suffering' is sufficiently deviant that it is distinct from whatever is picked out by our usage of the term 'suffering'. But suppose that a normal Earthling paid a visit to the World of Imaginative Torturers; in this case, we would still suppose that Earthlings and Torturers would be genuinely disagreeing with each other in their discussions of suffering and reasons for action. Given this, it seems that Earthlings and Imaginative Torturers share the same concepts; Imaginative Torturers do indeed share our concept of suffering. Now it is possible that the disagreement here operates at the level of properties, rather than concepts: if one person holds that the Evening Star has a mass of 5 x 10^{24}kg, and another holds that the Morning Star has a mass of 8 x 10^{24}kg, then at some level they are involved in a genuine disagreement (they ascribe different masses to the same object), despite using different concepts to locate the same entity. In the case of the Earthlings and the Torturers, the thought would be that their uses of the term 'pain' serve to express distinct concepts, but have the same reference. It is far from clear,

[89] See Hoffman 1987 for remarks on the connections between empathy and moral judgement.
[90] Neill 1996. It is not obvious that this counts as genuine, as opposed to 'as if', empathy.

however, that the Earthlings' use of the term 'pain' and the Imaginative Torturer's use of the term 'pain' differ in this manner.[91]

A second response would be to point out that although the imaginative torturer may have full empathetic capacities in the *weak* sense of empathy (see above), he lacks full empathetic capacities in the *strong* sense (on which empathy is a matter of sharing the subject's feelings).[92] Sharing the subject's feelings would, for the torturer, be unpleasant, and provide some motivation to desist.[93] Furthermore, for the torturer to empathise with the victim requires that the torturer take the victim's pain to provide reason for the victim's desire that the pain cease; sharing the victim's emotions requires that we engage with how the landscape of reasons lies for the victim.[94] In this sense, empathy involves sensitivity to reasons. On the strong conception, therefore, the imaginative torturers *do* display a straightforward psychological flaw: a lack of empathy.[95]

However, this response is unsatisfactory. Although lack of empathy in the strong sense may suffice to explain the torturers' lack of (appropriate) motivation, there are two problems with such an explanation. Firstly, it is relatively unhelpful, being tantamount to claiming that the torturers behave as they do because they are cold-hearted. Secondly, this looks like a purely motivational, rather than cognitive, failing; it is not that the torturers are not aware of what it is like for the victims, but rather that they do not care. But the original claim was that suffering provides reasons for action in virtue of the way that it feels; hence a full appreciation of the way that it feels, for rational agents, ought to suffice to provide motivation. The possibility of a world of

[91] There is also a thorny issue regarding reference. In the case of the Morning / Evening Star, there is a distinct entity which we can rigidly designate by the terms 'the Morning Star' and 'the Evening Star'; we can point to the object involved, such that distinct concepts can pick out one and the same entity. The case of mental properties is less clear cut: it is not obvious (indeed, highly doubtful) that 'pain' refers to mental states in the same way that 'the Morning Star' refers to a physical object. I will leave this issue here.

[92] In the experimental-psychological literature, empathy is characteristically defined as involving vicarious affect, or similar – and hence, understood along the strong conception. See Eisenberg & Strayer 1987.

[93] Darwall 1998: 271 points out that, '[s]ince Ezra Stotland's first experiments in 1969 . . . studies have consistently shown that subjects who projectively empathize report actual emotions and show physical symptoms that parallel the likely reactions of their targets.'

[94] Darwall 1998: 270.

[95] Kennett (2002) shares this view of empathy, although maintains (with Snow, op. cit.) that empathy cannot be *necessary* for moral agency, since high-functioning autistic persons, in spite of extreme empathetic impairment, nonetheless display some (albeit sometimes idiosyncratic) moral behaviour. On the hybrid theory developed by this thesis, however, there are plural sources of moral reasons; I therefore agree that empathy is not necessary for moral agency, but suggest that it *is* a route to detecting moral reasons in the realist sphere.

imaginative torturers is, therefore, deeply problematic.[96] Nonetheless, people are, in general, disposed to think of the relevant elements (the phenomenal character of hedonic and anhedonic states) as providing reasons for action. And there does seem to be something peculiar about the thought of the world of imaginative torturers. Now although the world of imaginative torturers seems to be conceptually coherent (as indicated by the Twin Earth considerations discussed above), it is still possible that its inhabitants share some cognitive shortcoming.

The world of the imaginative torturers is certainly peculiar. One diagnosis of this peculiarity lies in the impact of their beliefs about reasons; such a society would be a highly dysfunctional one. And it is hard to see how such a society could come about, regardless of what the correct metaethical account is. Each individual's interest in the cessation of his or her own suffering would quite naturally push that individual to ascribe some agent-neutral reason in favour of that cessation, regardless of whether or not such reasons exist. And given the close connection between suffering and evolutionary success, it is hard to see how dispositions which would, overall, reduce conflict, promote group welfare, and so on, would nonetheless fail to win out over dispositions which do precisely the opposite.[97] It is difficult to imagine precisely what such a world would turn out like, in part because such creatures would have to be very different to ourselves.

The imaginative torturers cannot be brought to see that the nature of their victims' experiences counts against promoting those experiences. Whereas Earthlings may demand assistance from other Earthlings, and invoke talk of reasons to support their demands, the torturers do not do this. The imaginative torturers hold that each has reason to pursue his or her own happiness (for instance), but that the welfare of others is unimportant. But the central issue, here, is not that the torturers must be taken to have peculiar empathic capacities, or peculiar moral conceptions (which they certainly do); nor is it that they have no use for reasons-discourse (perhaps they retain talk of agent-neutral reasons for pursuing beauty). Rather, the peculiarity is that they represent their own hedonic or anhedonic states as normative *only for them.* Doing otherwise – representing their own hedonic states as normative *per se* – would commit them to holding that the pain of other persons (with which they are imaginatively acquainted) provides reasons not

[96] Note that the relevant possibility is that of *global*, rather than individual, failure to perceive the putative reasons – since individual conceptual competence may be parasitic on the functioning of other members of the community.

[97] There does seem to be a clear practical advantage to possessing empathic capacities in the strong sense – namely insofar as simulation facilitates understanding of the mental states of others – but this understanding might, of course, be supplied by other means.

only for the sufferer, but also for the torturer. That is, I suggest that what is genuinely puzzling about this thought experiment is not that the torturers fail to see that they have reasons to help other persons, but rather that they deny that other persons have agent-neutral reasons to help them, in that there is something which the other person could count as a reason regardless of their particular desires.[98] The psychological pressure against this, in the case of normal persons, is significant, and we should, at least, be sceptical of the possibility of global behaviour of this kind.

We can bring further pressure to bear on the possibility of fully empathic, imaginative torturers. Specifically, there do seem to be cases where a full understanding of what things are like, for individual agents, requires an appreciation of certain reasons for action. Suppose that I am trying to decide whether to engage in a certain course of action – delivering a paper before an audience of my peers, for instance. Having done so in the past, I am aware that I found the experience traumatic; at the time, I was vividly aware that this amounted to a strong reason to avoid giving papers in future. But my present determination to advance whichever views I hold prevents me from fully appreciating this reason, and I happily commit to giving the paper. While there is a sense in which I know precisely what it is like to undergo such experiences, there is also a sense in which I am failing – at the time of deliberation – to fully appreciate what it will be like to repeat the experience. At the time of the decision, I fail to take one of the available reasons for action into account, despite displaying a full (weak) empathic understanding of my future self. There is, here, a direct analogy between empathising with one's future self, and empathising with other selves. If I am capable of displaying an imaginative awareness of my experiences (either prospectively or retrospectively) without appreciating the attendant reasons for action, then it is at least possible that this holds for the inter- as well as intra- personal case. That is, appreciating the attendant reasons for action may be a necessary condition of fully appreciating the nature of the experience. Hence, in spite of the imaginative torturers' apparent grasp of their victims' mental states, there may still be reasons which they fail to appreciate, where this amounts to a cognitive failing. Given these considerations, we should take the burden of proof as resting with the theorist who denies the existence of such reasons.

This chapter has been concerned to explore the claim that moral discourse aims (in part) to describe the way that things are. But our own dispositions and capacities constrain this: we are

[98] From their point of view, there may be something that would count in favour of being helped by others – namely that it would increase their welfare. Hence, there is some reason for others to help them. But this reason is merely agent-relative; a reason *for them* for others to help them.

interested in establishing how things are for creatures relevantly similar to ourselves.[99] And for creatures like us, empathy in the strong sense does provide a route to being appropriately motivated. But even empathy in the weak sense will also provide such a route, provided that our characterisations of our own hedonic or anhedonic states are evaluatively laden. As discussed in sections 3 and 5, I hold that this is indeed the case. We characterise suffering as an experience which *feels bad*, and pleasure as an experience which *feels good,* where this represents some element of the experience as objectively desirable – a good thing, in and of itself. Whereas there is a sense in which the torturers display an awareness of what it is like for their victims, there is also a sense in which they display a deviant understanding of their victims' experience: they fail to appreciate the reasons which attend the awfulness of their victims' suffering. There is, I have argued, moderate theoretical pressure to adopt the construal on offer, and consequently a sense in which the torturers are guilty of some cognitive shortcoming.

7.3. Naturalism

This account is a naturalistic account, if not a reductive naturalistic account; the reason-providing aspect of the relevant mental states is understood in terms of the phenomenal character of these mental states, and our capacity for empathic response and judgement. There is therefore no commitment to any metaphysically queer faculty of moral intuition; nor to any metaphysically queer properties. The intrinsic 'to-be-pursuedness' of hedonic states is, I have suggested, innocuous – it is no more or less queer than the notion of a pleasant or unpleasant mental state - and we can make good sense of the categorical reason-providing nature of such states, since we have a grasp on the (non-moral) defects operating in someone who fails to appreciate these reasons.

8. Conclusion

I have attempted to develop a limited form of moral realism, specifically realism about intrinsic value. A state is intrinsically valuable, on my account, if and only if its intrinsic properties provide agent-neutral, *pro tanto* reasons for that state's promotion. The chapter began with a recap of the constraints on an acceptable realist theory, and then moved to attempting to identify candidate bearers of value. The most plausible option, I suggested, is felt states – conscious states with a positive or negative phenomenal feel. I then turned to the question of agent-neutrality:

[99] C.f. Hurley 1989: 98-101.

whether such states could be plausibly understood as providing agent-neutral reasons, and whether we have good reason to do so. We can, and should, conceive of these states as providing agent-neutral reasons. A key element to doing so is providing an account of what goes wrong in an agent who appreciates the existence of such states, but fails to construe them as providing agent-neutral reasons for action – and to provide such an account in a way which is not question-begging. This can, I suggested, be done: the core element, unsurprisingly, is our capacity for empathy. That, in turn, secures the motivational connections to the relevant moral properties. Although these arguments are not intended to provide a knock-down argument for the existence of intrinsic moral value in the realist sense, they do show that a coherent account of intrinsic moral value (in the realist sense) can be given for the limited domain in question.[100] This gives a substantive theory answering to the presumption of moral realism established in Chapter One.

[100] That is, for pleasant and unpleasant mental states.

CHAPTER FOUR

Constructivism

1. Introduction

The aim of this chapter is to motivate, and then develop, a substantive constructivist theory. I begin by detailing two ways in which constructivist accounts are appealing: firstly, they offer an account which is close to moral realism but lacks objectionable ontological commitment; secondly, they sit well with a view of morality as a social, co-ordinating practice. The Prisoner's Dilemma is used as a test case. I then turn to a discussion of three constructivist theorists: David Gauthier, Onora O'Neill and Christine Korsgaard. From this discussion, two key themes emerge: one concerning the public use of reason, and another concerning moral motivation and its connection to our practical identities. I show how both of these themes relate directly to the role of morality as a social, co-ordinating practice, and highlight the role of universalisability in a plausible constructivist approach. I then turn to a critical discussion of two constructivist theories which are restricted in scope, those of John Rawls and Thomas Scanlon, and derive a concrete proposal on which acts are morally wrong if they would be ruled out by principles the following of which by the group is acceptable to each member of the group as a proposal for mutually coordinating action. Principles are acceptable for a group if (in outline), a) it is in the self-interest of each member of that group that the group follow those principles, b) their joint implementation is practicable, and c) members of the group are able to see the principles as justified. I conclude with a discussion of potential problems, with a focus on the issues of scope, categoricity, and the 'Moral Twin Earth' arguments developed by Horgan and Timmons.

2. Constructivism

Constructivism, as I will be using the term, claims that ethical propositions are truth-apt; that moral judgements express beliefs; that the truth of at least some of these propositions is knowable; that some of our moral beliefs are true; and that moral error is possible. Unlike the realist, the constructivist sees moral truth as a product of some procedure, rather than as simply 'out there, in the world, waiting to be discovered.' Sharon Street offers the following characterisation of constructivism:

‘*Constructivist views in ethics* understand the correctness or incorrectness of some (specified) set of normative judgments as a question of whether those judgments withstand some (specified) procedure of scrutiny from the standpoint of some (specified) set of further normative judgments.’[1]

However, allowing for the possibility of moral error entails that we can understand the truth of moral propositions as being (in some sense) objective. This does not commit the constructivist to maintaining that moral facts are thoroughly mind-independent; rather, it amounts to the denial of simple subjectivism. An analogy with colour might be useful here: even if (plausibly) our colour concepts are to be understood as response-dependent, we may nonetheless be mistaken about individual judgements of colour. Similarly, even if moral truths are determined by our responses, the relevant responses need not be our immediate, individual responses, but may instead be the responses (broadly construed) of groups, which may in turn be idealised to some extent. There are, of course, many different forms of constructivism; precisely how constructivism is to be understood should become clearer in the discussion of O’Neill and Korsgaard, later in this chapter.

The key point to be noted, here, is that constructivism aims to secure metaethical territory which shares as many of the key features of realism as possible whilst avoiding the stock problems which face moral realism. This, therefore, is the first motivation underlying constructivism. If it can be made to work, constructivism saves the appearances of moral discourse, and offers a way to resolve moral disputes: since moral truth, for the constructivist, is constituted by the outcome of some construction procedure, following the procedure provides an accessible moral epistemology. But more on that in due course. There is a second reason for treating constructivist accounts as desirable, and that is that they are well suited to account for the distinctively social or political role of morality. This is the role that morality plays in mediating between agents with potentially conflicting, or unco-ordinated, interests. Let us take the Prisoner’s Dilemma as a test case.

The Prisoner’s Dilemma (henceforth PD) is a familiar example from game theory. It involves two (independent) parties, each of whom is faced with a choice: to either testify against their compatriot (i.e. ‘defect’), or remain silent (i.e. ‘co-operate’). If both testify, they each receive a moderate sentence; if both co-operate, they each receive a short sentence. But if only one testifies, then his compatriot receives a hefty sentence whilst the defector goes free. The payoff matrix, then, looks like this:

[1] Street 2006.

	A Co-operates	**A Defects**
B Co-operates	A & B serve 3 months	A goes free B serves 10 years
B Defects	B goes free A serves 10 years	A & B serve 5 years

The rational choice for A and for B is to defect, because the possible outcomes are then <freedom, 5 years> as opposed to <3 months, 10 years>. Note that this is independent of what each expects the other to do: from A's point of view, A will be better off if she defects, regardless of what B does, and *vice versa*. Defection is said to be the 'dominant' strategy. However, if both A & B choose rationally, then they will each serve 5 years – which is a significantly worse outcome than if both had co-operated, and received minimal sentences as a result. The PD is paradoxical because individually rational, self-interested behaviour produces collectively suboptimal results.

It is clear that this feature of the PD is widespread. For instance, my choosing to download a film from the internet, rather than purchasing it, is in my interest (provided that I am undetected). But if everyone chose to download, rather than purchase, their films, the industry which provides the films would collapse; each person's desire to watch films would be frustrated. More generally, there are many systems whose existence is valuable (to people, in general), and which tolerate a degree of freeloading, but which would be destroyed by universal (or widespread) freeloading. These include physical systems (particularly natural resources, such as fish stocks), as well as industries and establishments, together with more abstract systems such as promising, truth-telling, and so on.

Coupled with the observation that such freeloading is in many cases thought of as immoral, it is tempting to think that morality is (partly) to be explained with reference to its role in mediating Prisoner's Dilemma-type situations.

2.1. Should we be worried about the Prisoner's Dilemma?

The thought under consideration is that part of morality is concerned with Prisoner's-Dilemma type cases. This thought is controversial. Ken Binmore, for instance, claims that game theorists 'think it just plain wrong to claim that the Prisoner's Dilemma embodies the essence of the problem of human co-operation.'[2] According to Binmore, the Prisoner's Dilemma is a neat representation of a situation in which defection is always the rational strategy, *regardless* of how altruistic we are feeling. Kantian (or other constructivist) accounts which attempt to show how it can be rational to co-operate (whether simply as a constraint on practical reason, or in an attempt to maximise joint goods) are entirely misguided. Binmore presents a 'knock-down refutation of this nonsense', which runs as follows.[3] First, note that elements of the pay-off matrix which indicate the possible outcomes of the Prisoner's Dilemma are supposed to represent gains – or, as Binmore has it, preferences. It makes more sense to talk of preference-ordering, rather than absolute gains, since the prisoners may, for instance, be entirely indifferent to money or have peculiar, or very different, attitudes towards captivity.[4] And the way in which we detect people's preferences is by examining the choices that they make. So if 'co-operation' were rational, then it would be preferred (since both agents are assumed to be rational), and hence represented by a higher pay-off in each case (regardless of whether the opponent 'co-operates' or 'defects'). But the essence of the Prisoner's Dilemma is that it is in A's interest to defect when B defects, and to defect when B co-operates. Hence, showing that it is rational to co-operate merely shows that the Prisoner's Dilemma is no longer a Dilemma – it is a 'Delight'. Put more simply: if each prisoner would choose to co-operate even if the other were to defect, then the pay-offs change, and there is no dilemma.

Does this argument conclusively demonstrate that it can never be rational to co-operate in PD situations? In the narrow sense in which Binmore understands the contents of the pay-off matrix, it certainly seems to. But it is far from clear that we should run together rational

[2] Binmore 2005: 63.

[3] Ibid.

[4] There is, of course, a problem of 'adaptive preferences': the downtrodden, unfortunate, or uninformed may appear content with their lot, even though they are obviously much worse off than their equally-content neighbour (see Sen 1988: 45). Alternatively, I may convince myself that I never really wanted to be a professional rugby-player, although this preference arises (let us suppose) as a defensive response to being woefully unsuited to playing rugby. As Nietzsche puts it, '[w]hen stepped on, a worm doubles up. That is clever. In that way he lessens the probability of being stepped on again. In the language of morality: *humility*.' (Nietzsche 1990, §31).

choice, preference, and gain in quite this way. Clearly, if rational choice and preference are identified, then behaviour indicates preferences. And it is often taken as axiomatic that rational agents will choose whatever has the highest payoff – where 'payoff' doesn't mean 'money / happiness / etc. for the agent', but rather just 'whatever the agent values'. Now it is not obvious that what the agent values just is whatever the agent would rationally choose, nor that either of these is identical to what the agent consistently prefers (and, as Binmore points out, what does the work in game theory is consistency of choice).[5] But I think that Binmore's argument runs too fast, for at least two reasons.

The first of these begins by pointing out that, despite what the argument shows, maybe morality *did* arise as a response to PD cases. The kinds of Prisoner's Dilemmas with which philosophers tend to be concerned are well-described: we are indeed concerned with cases where, were it not for some altruistic or moral motivation, it would be rational to reliably defect. But given altruism, it can be rational to co-operate. Morality thus has an important role in dealing with *what would otherwise be* PD cases, for although it may be unable to show non-moral agents that it would be rational *for them* to co-operate, it can provide a way in which groups of agents can transform Dilemmas into Delights.[6] Suppose that the two Prisoners are equally averse to time spent in prison, and hence value mutual co-operation over joint defection to the same extent. If each side prefers to co-operate in any case, then mutual co-operation becomes even more strongly preferred.

The second is that, regardless of agents' preferences, we may still talk of them being harmed or benefited by the choices that they make. Perhaps – quite plausibly – I reliably choose to smoke cigarettes. If preferences reflect value from the agent's perspective, then smoking cigarettes is valuable to me. As it happens, I smoke cigarettes because I am weak-willed, rather than because I actually prefer to smoke. Our actual choices, however consistent or predictable, need not always reflect what is in our best interests, what is of most value to us, nor our deep-seated preferences (although they may, at least sometimes, provide overwhelming evidence in that direction). Similarly with so-called 'adaptive preferences': it is clear that people's choices may on occasion reflect a lack of hope, or be constrained by a malign way of viewing the world, and that, consequently, the world is full of downtrodden

[5] See Broome 2004: 31 – 39 & *passim*.

[6] See Blackburn 1998: 184, 'So-called countertheoretical actions do not reveal the irrationality of the players, but the impropriety of this application of the theory.' That is, co-operation in apparent Prisoner's Dilemmas should raise a question about what is being modelled, rather than a facile dismissal of the actors as irrational.

housewives who are ‘happy with their lot’, would-be rugby players who do nothing to fulfil their ambition, etc.[7] This is in contrast to modern utility theory, which

‘ . . . doesn’t say that Eve chooses *a* rather than *b* because the utility of *a* exceeds that of *b*. On the contrary, the utility of *a* is chosen to be greater than the utility of *b* because it has been observed that Eve always chooses *a* rather than *b*.’[8]

To reiterate: even were the equation of utility with choice warranted, we could still retain the notion that the Prisoner’s Dilemma is a relevant test case. Showing that it would be rational to co-operate amounts to turning the Dilemma into a Delight: that consequence is innocuous. But the equation of utility with choice is unwarranted. Regardless of whether the one is a reliable indicator of the other, we ought not to identify the two.

2.2. To what extent is Game Theory relevant?

It should be entirely unsurprising that game theory is well-placed to shed light on our moral behaviour and intuitions. After all, many of the test cases which game theorists consider are designed to bring out the fact that we are driven by considerations of fairness, desert, and so on, and not merely by neutral evaluations of cash rewards. What is less clear is the lesson that we should take from game theory. Clearly, in certain circumstances it becomes rational to adopt a stable strategy of co-operation (together with a disposition to punish those who fail to co-operate, renege on agreements, etc.), such as when the PD (or some variant) is indefinitely reiterated. And, indeed, indefinitely reiterated PD cases are common: arguably, even one-off Prisoner’s Dilemmas, as played out in the laboratory, ought to be treated as instances of a broader scheme of such cases (where the Prisoners have relatively stable dispositions, rationally acquired, to co-operate, and an interest in displaying their co-operativeness to anyone who happens to be watching).

Now one might be concerned that in iterated PD cases, it is simply rational to be mostly co-operative (that is, to play forgiving tit-for-tat), at least given that, for most people, defection is likely to be detected and punished. Similarly in the case of the Ultimatum game, where a sum of money is to be divided between two parties: one party offers the other a take-it-or-leave-it deal, but if the other party refuses, neither party receives any money. Here it is rational to refuse ‘unfair’ offers, provided that the conditions are right (the chances of reiteration are

[7] See Sen 1988, esp. 29-56.
[8] Binmore 2005: 98.

reasonable, there are onlookers, etc).[9] This, again, is innocuous: as I argue later in this chapter, we should think that it is generally rational to *behave* morally, even if agents are understood to be narrowly self-interested. The more interesting case is, of course, that of the agent whose interests turn out to be best served by behaving immorally. But more on that in due course.

Game theory is, I suggest, relevant for at least three reasons. Firstly, it can illuminate the way in which apparently irrational behaviour (such as refusing unfair deals even when doing so results in an opportunity cost) can be in our long-term interests. There may, therefore, be good reason (given certain other conditions) to behave morally. Secondly, it can shed light on the kinds of motivations that agents actually possess, and consequently on the relevant moral concepts: fairness, reciprocity, promise-keeping, and so on. Thirdly, it offers a way to understand the underlying structure of (what I take to be) the constructivist domain: constructed morality is functional, because it maps onto sets of stable strategies which can be commonly adopted within societies.[10] If this third point is correct, then one might reasonably worry that the best understanding of this part of morality is some form of straightforward realism: what is right is whatever forms a stable set of strategies which can be commonly adopted for societies, for instance. But this would be too fast.

Bear in mind that one central question for constructivists is the question of what (if anything) makes moral claims true. Now as far as the game-theoretic story goes, moral codes are mechanisms which push agents towards ideal sets of behavioural strategies. The ideal set of behavioural strategies for any group, or society, is that which forms an optimally efficient Nash equilibrium – that is, forms a set of strategies where each player's strategy is a best reply to the strategies of the other player, and where this set is at least as efficient as any competing set. Moral codes are (*inter alia*) neat ways of specifying such sets; efficiency then provides a means of comparing codes. The truth of moral claims does not then boil down to the question of locating the most efficient equilibrium. Rather, the moral code forms a kind of intermediary between the equilibria, on the one hand, and agents deliberating under conditions of uncertainty, lack of information, etc. Put differently, constructivist morality is not concerned with *optimal* solutions, but with *satisfactory* solutions. Hence constructivists can happily take on board a great deal of the game-theoretic (or evolutionary) story, without committing themselves wholesale to a reductive naturalist realism – or, for that matter, to a

[9] 'Rational', here, is understood as meaning 'maximising the agent's own expected utility'.

[10] I take it that game theory is useful for more reasons than this – but these three are, at least, sufficient.

‘New Wave’ moral realism, on which moral properties are identical to natural properties.[11] In any case, constructivism is neutral on the issue of whether or not the proposed analysis is reducible to (i.e. type-identical to) natural kind terms, or even irreducibly natural (i.e. token-identical to natural entities). In the process of discovering what makes constructed moral claims true, we also gain an understanding of how we can come to know moral truths – and this, of course, is useful.[12] But if we are to take naturalism seriously, and to attempt to locate moral claims in relation to the claims made by the sciences – to solve the ‘location problem’ for ethics – then we ought, I think, to take the notion of finding the relevant equilibria very seriously indeed.[13] Understanding the function of moral discourse and practice as related to the solution of co-ordination problems addresses this problem.

Earlier in this thesis I noted that one desideratum for any metaethical theory is that it capture the point of the practice. Indeed, this is one of the primary means of capturing the distinction between, for instance, realist and expressivist theories of ethics: realists hold that moral discourse and practice aims at representing (or describing) the world, whereas expressivists hold that the point of moral discourse and practice is to enable us to express and co-ordinate (in a broad sense) our attitudes. Part of the point of moral discourse, on my account, is to report on various phenomenological considerations. A further part – supported by the considerations from game theory – aims at solving co-ordination problems, broadly construed. But the truth conditions for moral claims in this domain do not reduce to this level: I do not suggest that acts are morally permissible if and only if they would provide solutions to co-ordination problems. Rather, I suggest that the permissibility (or otherwise) of certain acts is dependent upon their falling under, or being excluded by, certain principles, where those principles in turn are determined by the construction procedure. There is a distinction to be had between the justifications for (and truth-makers for) the principles, and the justification of the practice. Constructivism, then, is an account of the grounding of certain moral reasons. The subject matter of the construction procedure may be moral principles; may involve the conception of certain moral concepts, such as justice; values, such as fairness, and so on. There is no reason why the constructivist account need revolve exclusively around deontic principles, rather than values, but the subject matter of the construction procedure *does* need to be (as I discuss later in this chapter) apt for public deliberation, hence relatively concise, and general in form. If constructivist morality is to fulfil the role which I have

[11] In terms of the present discussion: where the property of moral rightness is identified with (for instance) an act’s compliance with a moral code which, if instantiated, would yield an efficient Nash equilibrium.

[12] C.f. Binmore 1994: 335, ‘The game of morals . . . serves as an *equilibrium selection* mechanism for the game of life.’ If we are to take this claim seriously, then Binmore is not a reductive naturalist, despite his claims to the contrary.

[13] See Jackson 2000: 113-139.

outlined, it will not be able to do so on a case-by-case basis. As regards ethical particularism, the constructivist account will turn out to be anti-particularist.

I have suggested that the construction procedure may address values, as well as deontic principles. It may turn out that, for instance, the construction procedure dictates that we treat social equality as valuable. And I take it that for an outcome to be valuable is (at least) for it to provide reasons for its promotion. But axiology, by itself, underdetermines action. Since a key role of constructivist morality is to provide satisfactorily determinate solutions to co-ordination problems, the construction procedure will need to address deontic principles, as well as (if appropriate) axiology.

With this in mind, I now turn to one direct attempt to address PD cases – that of David Gauthier.

2.3. Gauthier

Gauthier maintains that, for rational, self-interested agents, there is reason to adopt a disposition of co-operating in PD cases, and that rational agents embodied and situated in roughly the ways and environments that we find ourselves will therefore consistently choose to co-operate. [14] It is in our interest to adopt these positions because cheats are readily detected: we are 'translucent'.[15] It is therefore rational to behave morally (which includes not only acquiring a stable disposition to co-operate, but also to complain against, and attempt to punish, defectors, and so on).[16]

There are problems with Gauthier's theory. It is unclear, for instance, whether Gauthier can provide an accurate account of moral theorising. If moral judgements are merely attempts to find solutions to co-ordination problems, then the role of mediating concepts, such as respect for persons, stands in need of explanation. Moreover, Gauthier's account does not deal with the observation that some people are highly opaque: there are multiple instances of successful liars, crooks, and suchlike. For the Gauthier-style theorist, all that results is an injunction to

[14] See Gauthier 1986.

[15] Gauthier 1986: 169 'Constrained maximizers can . . . obtain co-operative benefits that are unavailable to straightforward maximizers . . .'. A constrained maximiser will choose to co-operate if, given her assessment of her compatriot, 'her own expected utility is greater than the utility she would expect from the non-co-operative outcome' (ibid.)

[16] Thus Scanlon 1998: 190, in a discussion of Hare's universal prescriptivism, claims that '[r]ationally defensible moral principles will thus be those that lead to maximum satisfaction of the rational preferences of all affected parties.'

behave morally if you can’t get away with behaving otherwise. This is, I take it, unsatisfactory.[17]

Nonetheless, this approach carries with it much appeal. Although, as I argued in the previous chapter, we should think that part of the point of our ethical practice is to allow us to describe a salient feature of conscious experience, it also highly plausible to think that part of the point of our ethical practice is to be understood in terms of co-ordinating actions, mediating between conflicting interests, and so on. Perhaps other systems can go some way towards solving these problems: the legal system, for instance, provides much structure for our interpersonal dealings, and also provides a system whereby conflict is moderated. But these systems will not, by themselves, be sufficient. We might think, as David Copp has suggested, that unless people tend to view themselves as under an extra-legal duty to comply with the law, there will be insufficient intra-legal compliance.[18] In any case, the salient feature of morality, as opposed to law, is its pervasiveness. Moral approbation and condemnation attach to more types of action, and do so more rapidly, than legal sanction or punishment. This is not merely an observation about current practice: a law mandating general social co-operativeness would be impractical (barring special cases), but moral condemnation of individual social unco-operativeness is both practical and effective.

Constructivism is well suited to addressing this observation because it is – as should shortly become clear – positioned to track solutions to these various co-ordination problems without losing normative force. Here is one formulation of constructivism, from the early Rawls:

‘. . . moral objectivity is to be understood in terms of a suitably constructed social point of view that all can accept. Apart from the procedure of constructing the principles of justice, there are no moral facts.’[19]

Constructed moral objectivity, therefore, may provide a way for parties involved in PD cases to deliberate about their situation from outside of their individual perspectives – or, at least, from a shareable perspective. Similarly, accounts which treat morality as a function of what we could, or would, jointly agree to; or as a function of the content of some hypothetical

[17] A challenge of the same variety can be levelled at Scanlon’s theory. Scanlon claims that we have an interest in behaving in ways which are justifiable to others on grounds that they could not reasonably reject. But even if this is true, all that follows is that we have an interest in behaving in ways for which we could give a (not reasonably rejectable) cover story. And given that an individual action might be underwritten by a variety of principles, it is not clear what is stopping us from acting on a morally dubious principle whilst proclaiming the opposite.

[18] Copp 1995: 106-107

[19] Darwall, Gibbard & Railton 1997a: 248

contract; and so on. Constructivist accounts vary in their construction materials and procedures. The materials may be more or less metaphysically rich – and, of course, the plausibility of the theory varies with the richness of the suppositions. There is, here, something of a dilemma for constructivists: on the one hand, the more minimal the starting materials, the harder it is to derive substantive ethical constraints; on the other, the richer the starting materials, the more the starting materials are open to debate.[20] It will therefore be illuminating to look at two prominent constructivisms – that of Onora O'Neill, representing a metaphysically innocent approach, and that of Christine Korsgaard, embodying more substantial commitments. I will examine each position in turn. Each of the two constructivisms discussed here, I will argue, has elements which should be retained.

2.4. O'Neill

O'Neill attempts to generate a constructivist account of ethics from a minimal base, viz. that 'anything that is to count as reasoning must be followable by all relevant others'.[21] This 'followability' condition entails two constraints: firstly, the proposed principles (or reasons, plans, etc.) must be intelligible; secondly, they must be 'real proposals for action'.[22] A consequence of the second condition is that for any proposed principle 'its universal adoption in the relevant domain would not be incoherent', where 'incoherent' is understood as implying some kind of practical contradiction.[23] The principle could not successfully mandate acquiring a monopoly over resources, for instance, since not everyone could adopt such a principle (although there might be room for some kind of competition towards such an end). Importantly, this minimal base involves only abstraction, and not idealisation: in contrast to Rawls, O'Neill's constructivism does not rely on any (or, at least, any particularly substantial) metaphysical claims.

One might worry that some (morally permissible) principles will not turn out to be 'real proposals for action' in the required sense. Henry Richardson, for instance, worries that whilst I understand perfectly a principle which mandates eating extra vitamins whilst pregnant, this will never be a principle which I could adopt.[24] Depending on one's views on personal identity, we could consider this principle as a straightforward counterfactual (were my

[20] See Enoch 2005.
[21] O'Neill 1996: 3. Also, '[Practical reason] should . . . *at least aim to be followable by others for whom it is to count as reasoning*' (ibid. 51, my emphasis); 'anything that is to count as practical reasoning must be followable by others within the relevant scope' (ibid. 55).
[22] Ibid. 56
[23] Ibid. 58
[24] Richardson 1999: 599

circumstances radically different, then . . .), or as a counterpossible.[25] Alternatively – and this, I suspect, is the move most congruent with O'Neill's overall strategy – we could abstract further from our situation, and consider more general principles of action. Specifically, the operative principle here would be one of caring for dependent others (or something similar): for O'Neill, it is left to our judgement to implement such general principles in an efficacious way. And the principle in question is still a real proposal for action, albeit not for me: if the account needs to be amended, this can be done without losing the central theme.

The minimal base can, it is claimed, be added to. We are, O'Neill maintains, required to apply whatever assumptions are needed for purposes of quotidian activity to ethical theory. This amounts to the three conditions of *plurality* (the assumption that there are others, distinct from the agent); *connection* (that those others are connected to, or interact with, the agent) and *finitude* (those others have limited but determinate powers). This approach avoids the pitfalls of attempting to fix the scope of ethical theory by appealing to some prior criteria (rationality, the ability to suffer, etc.), and instead holds that 'a reasoned way of resolving practical questions about others' ethical standing can be constructed on the basis of the corrigible assumptions agents make about connected others . . .'.[26] The picture so far, then, runs close to Kant's own ethical theory: moral requirements are constraints of practical reason. There will be principles which fail to be universalisable (i.e. jointly instantiable) in virtue of their intrinsic properties (e.g. a principle which mandates acquiring a monopoly over some good); principles which fail to be universalisable in virtue of some contingent fact about the world (e.g. a principle which mandates acquiring sufficient resources for survival, in certain cases); and principles which do not fail to be universalisable. Those principles which fall into the first category are immoral, and provide a grounding for the rest of our (constructed) ethical theory. So, for instance, an principle of causing wanton injury to others cannot be universalised, and this 'provides the material both for constructing an account of other more determinate principles of justice . . . and with them of just special obligations, rights and relationships [etc.]'[27]

This yields three interconnected grounds of moral obligation. Since individual morality involves concern for others, principles which permit direct injury are ruled out; similarly, since those others are dependent on limited resources (specifically, our shared environment), principles which permit overuse of these resources are ruled out; and since we are dealing

[25] Specifically, depending on whether or not one thinks that sex is an intrinsic or accidental characteristic of one's personal identity.
[26] O'Neill 1996b: 121
[27] Ibid. 166

with '*connected others* . . . those who reject injury must reject activities, institutions and practices that gratuitously or systematically deceive . . . indirectly injuring the connections between lives.'[28]

Perhaps the main worry, here, is that which was raised previously in discussion of theories which attempt to found ethics in rationality. Even if ethical constraints are, at base, rational constraints, why should we care about *those* constraints? We might seek to answer this question by pointing out that the question itself presupposes a commitment to taking rational constraints seriously. Any possible answer to the question would involve a normative reason, a reason to care about certain constraints. Posing the question sincerely involves being prepared to take answers seriously, and hence requires that we already care about rational constraints. But there is a problem with this manoeuvre, which is that there is conceptual space between 'fully rational' and 'fully irrational'. This, in turn, makes room for the question 'why be *fully* rational?'. Perhaps I need to be generally in the business of taking normative reasons seriously in order to sincerely pose such a question.[29] But I do not need to be thoroughly committed to taking all such reasons seriously. I might think (quite plausibly) that what I most want is to generally, but not always, conform to the requirements of theoretical or practical reason. For instance, we may sometimes want to be able to get out of the business of theoretical and practical reason altogether. Euthanasia for an agent whose life prospects are unbearable is one instance of this.[30] Furthermore, although very few people (if any) are fully rational, such rational failings as they exhibit do not result in an overall collapse of rationality. Quite possibly, the most reasonable attitude to have towards one's own rational capacities is a satisficing attitude: we care about meeting most of the constraints of practical reason, most of the time, but not all of the constraints all of the time – and representing oneself as practically irrational is, in any case, a small price to pay.

O'Neill's 'minimalist' position is at a slight advantage over her competitors on this issue: the less substantial the construction materials, the less work that needs to be done to justify them. Unlike Rawlsian constructivism, which takes an idealized conception of the agent (as mutually independent, the head of a household, etc.), O'Neill's constructivism employs only abstraction – which we are forced to use in identifying actions, since we have to pick out some description under which the action falls, and this is a paradigmatic case of abstraction.

[28] Ibid. 179

[29] Railton 1997 discusses the viability of 'constitutive' arguments in favour of non-hypothetical requirements on practical reasoning – where it is argued that taking these requirements seriously is partially constitutive (and a necessary condition) of agency – and concludes that these arguments fail. Even if taking these requirements seriously were necessary for agency, one might not be concerned about this.

[30] C.f. Railton 1997: 74

As to the question of why we should care about the procedure, it is clear that O'Neill thinks that universalisability is a constitutive element of rationality (c.f. the discussion of Korsgaard's 'publicity' condition later in this chapter). Nonetheless, a worry about the justificatory force of the construction procedure remains. Fitzpatrick raises the following dilemma: if the principles of practical reason have normative force, it is because they result from some construction procedure; but if the construction procedure itself has normative force, this must either be derived (in which case a regress threatens) or intrinsic (which is both implausible, and entails a form of normative realism).[31] O'Neill responds by choosing the 'regress' horn of the dilemma, although she argues that the regress is neither infinite nor vicious: justification is a function of 'agreement based on principles that meet their own criticism.'[32]

This claim is, of course, open to the stock objections which face any coherentist account of justification. In particular, there may be any number of bizarre, and contradictory, but internally coherent justificatory frameworks: this is problematic partly because it then becomes indeterminate whether any given statement is justified (relative to one framework) or not. Equally, we might think that there is more to the matter of justification than the question of whether or not any one system is vindicated by its own principles. Worse, we lack a positive argument for the claim that agreement of this sort can generate normative force. The denial of realism closes off one horn of the dilemma, but that does not make the second horn true by default.

There is also a worry about direction of explanation. O'Neill seems to be claiming that the content of reason is a function of its universal followability, rather than *vice versa*. This is a curious claim: after all, we think that what makes *modus ponens* universally followable is that it is a valid form of inference. It is not clear why we should think that other rational constraints are any different.

Furthermore, there is a minor worry about vacuity: if it is constitutive of reason that it deals in principles which are followable by all, then we have the problem of how to pick out the relevant 'all'. This cannot simply be 'human beings', since there are human beings who – whether due to accident, injury, constitution, immaturity or senility – lack the ability to follow the various principles. But nor can it be 'rational beings', since, absent a prior notion of what is to count as rational, we have no way of picking out this set of entities. Perhaps the only viable option is to talk about all and only those beings who are capable of engaging in shared

[31] Watkins & Fitzpatrick 2002: 353.

[32] O'Neill 1989: 38 (cited in Watkins & Fitzpatrick 2002: 345).

deliberation; this, of course, raises the question of what is to count as shared deliberation of the required sort. That, however, is less problematic than the question of what is to count as rational: we do have some idea of what is, and is not, to count as shared deliberation, even if there may be borderline, or vague, cases.

Relatedly, one might think that O'Neill's move from the followability constraint to the universalisation constraint is illegitimate. After all, the claim that moral principles must be 'followable by all' might mean '*jointly* followable by all persons' (or 'followable by all persons simultaneously'), or simply 'followable by each person independently'. A principle of acquiring a monopoly over some good, for instance, is for each person a followable principle, albeit not for all. However, it would be uncharitable to think that the argument turns on such an obvious equivocation. Bearing in mind that reason is a public entity (in the sense that its dictates are followable by all in the relevant domain), and that moral principles are addressed to persons in general, rather than to particular individuals, we can think of these principles as being constrained to provide 'real proposals for action' for groups of persons, and not just for individuals. But this, in turn, raises the question of why the principles which guide individual action should obey similar constraints to the principles which guide group action. One answer might be that whether something counts as a rational constraint must be decided by the character of the constraint itself, rather than by contingent features of the world. That is, the acceptability of the principle cannot depend on the majority of individuals exercising restraint. This being so, constructivist ethical principles will be general in their form – that is to say, they will not be 'hedged' principles which enjoin us to 'do this, provided that most others are not doing so'. The principles which guide individual and group action should therefore be similarly constrained. I discuss this issue further in Sections 3.1 and 4.2.

As discussed in the previous chapters, there is also a worry here of revisionism about justification: wanton injuriousness is not wrong because it instantiates a non-universalisable principle, but rather because it harms others (although the story for other wrongdoings may differ). Similarly, there is a worry about redundancy. Susan James, in her critical notice of O'Neill's *Towards Justice and Virtue,* observes that:

'It is arguable that the principle of rejection of injury derives its persuasiveness not from its universalizable character, but from its content . . . members of different societies will interpret the principle in different ways: they will have varying conceptions of injury, of when an injury is gratuitous [etc] . . . But if they ever articulate O'Neill's most general principle, its persuasive force is likely to derive more from its affinity with comparatively specific principles to which they are already committed than from its universalizability.'[33]

I think that this is quite plausible, but perhaps not all that problematic for O'Neill. After all, if we are to take our intuitions about wrongdoing seriously, then we ought to think that they are non-accidentally connected to the truth in morality, however that is to be understood. Then, of course, we had better have some account of how this could be so, in a way which fits with O'Neill's universalisation tests. Fortunately, it is quite easy to see how such an account might turn out: if moral principles are those whose universalisation is possible, then it would be reasonable to suppose that widespread uptake is a good indicator of universal followability, and hence of moral truth.

2.5. Korsgaard

Christine Korsgaard, primarily in her *The Sources of Normativity*, develops a substantially different constructivist theory. She begins by identifying what she calls 'the normative question': for her, this is the question of what justifies the demands that morality makes on us. Since moral concepts have various practical implications (such as the connection with motivation), we have a criterion of explanatory adequacy; but we also have a criterion of 'justificatory adequacy'. Whatever our metaethical theory, it must be one which vindicates the putative moral demands. And the answer must be a response to someone who asks for a justification of these moral demands. Taken together, these yield what Korsgaard calls the 'transparency condition':

Transparency: Any acceptable answer to the question of what justifies the demands of morality must be such that, once we understand it fully, we will still take those moral demands to be justified.[34]

More controversially, Korsgaard also claims that the answer must appeal to our sense of identity, because there must be occasions where to do the wrong thing would be worse than death, and the only thing which is as bad as death is the loss of our identity (which, Korsgaard thinks, amounts to the same thing).[35] Lastly, if the answer is to meet the justificatory

[33] O'Neill 1996; James 1998: 262-3.
[34] See Korsgaard 1996b: 17.
[35] c.f. Korsgaard 1996b: 102: 'It is the conceptions of ourselves that are most important to us that give rise to unconditional obligations. For to violate them is to lose your integrity and so your identity, and to be no longer who you are . . . it is to be for all practical purposes dead or worse than dead.' See also Korsgaard 1996b: 16.

requirement, it must be such that it renders the normative question redundant; once we understand the answer, there should be no further need to ask 'why should I care?'. This is one reason why Korsgaard thinks that moral realism is unsatisfactory: even if we posit brute moral facts, that does not answer answer the question of why we should care about these facts. We will only be happy to admit intrinsically normative entities (she claims) to the extent that we are already convinced of the claims of morality.

The notion of 'personal identity' which Korsgaard has in mind (and on which, for her, ethics rests) is a strong one: she claims that '[t]he conception of one's identity in question here is . . . better understood as a description under which you value yourself, a description under which you find your life to be worth living and your actions worth undertaking.' And the connection between personal identity and action is very strong indeed: it is not just that we act morally because we self-identify as moral beings, but rather that '[i]t is necessary to have *some* conception of your practical identity, for without it you cannot have reasons to act.'[36] The thought underlying this claim, it seems, is the following: unlike other animals, people have the ability to reflect on, and adjudicate between, their impulses; but this requires that we have a sense of the self which is doing the adjudication, and consequently generating reasons for action.

As Baehr points out, Korsgaard's claim about the relation between valuing one's humanity and reasons for action seems to be simply false: sometimes we come to have practical identities for no particular (normative) reason, but – even if Korsgaard is right about the connection between practical identities and reasons for action – these identities may still provide us with reasons.[37] For most people, their practical identity in fact seems to derive, at least in part, from such arbitrary sources. For instance, a salient part of many persons' practical identity is their religious identity – but they acquire this identity from the culture in which they grew up, rather than selecting it. More generally, our personal identity often depends on contingent and unchosen facts about our environment, which are simply present, rather than chosen. Furthermore, Korsgaard appears to be claiming that the requirement that we value our humanity depends on 'see[ing] that your need to have a normative conception of yourself comes from your human identity.' We might – as, presumably, most people actually do – fail entirely to see that this connection obtains. Worse, some people may have deeply immoral self-conceptions: perhaps I think of myself as a serial killer, and value myself under this description. Suppose further that my serial-killing image is not a merely accidental

[36] Korsgaard 1996b: 120.
[37] Baehr 2003: 485.

feature of my identity, but rather a deep, essential component.[38] Korsgaard offers the following two-stage argument to defang this objection, beginning with a (transcendental) argument for the claim that each individual must value their humanity:

‘You must value your humanity if you are to value anything at all . . . because now that you see that your need to have a normative conception of yourself comes from your human identity, you can query the importance of that identity . . . since you cannot act without reasons and your humanity is the source of your reasons, you must value your humanity if you are to act at all.’[39]

How should we get from the claim that each must value their humanity to the claim that each must value each others’ humanity? Korsgaard’s argument is notoriously unclear on this point: although it seems that her argument turns on the thought that reasons are essentially public, it has also been argued that the key issue is not the publicity of reasons, but our shared identity as reason-givers.[40] She does, after all, claim that ‘[a]n animal can obligate you in exactly the same way another person can. It is a way of being *someone* that you share.’ One consequence of the opacity of Korsgaard’s argument, unfortunately, is that different commentators identify different ‘key’ arguments – the more sympathetic commentators identify more plausible arguments, and *vice versa*, with accuracy being, perhaps, sacrificed on occasion.

Nevertheless, the overall direction of Korsgaard’s argument seems to be as follows. Reasons have to be public, because they are linguistic entities, and there is no such thing as a private language. This publicity amounts to shareability. But the shareability in question is a strong kind of shareability: it is as if language provides a space in which our consciousnesses may impinge on each other. ‘Why shouldn’t language force us to reason practically together, *in just the same way that it forces us to think together*?’, she asks.[41] Using various words (such as ‘how would you like it if I did that to you?’), I can force you to put yourself in my shoes, and hence to think that, were you in my shoes, there would be reason for me to treat you in a certain way, and hence, since this reason is grounded in your humanity, so my humanity must also ground reasons for *you* to treat me in a certain way. Since I can’t hear your words as ‘mere noise’, I am committed to understanding you as being a person, and to sharing in a joint process of practical deliberation, one in which your reasons count for me, as well as for you -

[38] Korsgaard allows that there may be conflict between the specific demands of morality and those of some more contingent form of identity: see Korsgaard 1996b: 126.

[39] Korsgaard 1996b: 123.

[40] Ibid. 135 – & c.f. the following (slightly curious) claim: ‘I take this [viz., the publicity-as-shareability thesis] to be equivalent to another thesis . . .that what both enables us and forces us to share our reasons is, in a deep sense, our social nature.’ See also van Willigenburg 2002.

[41] Korsgaard 1996b: 142, my emphasis.

unless, of course, you speak an entirely different language. Even then, there is a difference between hearing it as a language and hearing it as 'mere noise': if the former, then I am still committed to treating you as 'someone'.

This is unsatisfactorily vague (and not just as a characterisation: consider Korsgaard's claim that '[t]o act on a reason is already, essentially, to act on a consideration whose normative force may be shared with others.').[42] Problematically, it does not seem that the connection between the 'essentially public' nature of reasons, on the one hand, and the requirement that I take your reasons to be normative for me, on the other, goes through. The difficulty, I think, lies in the way in which 'shareability' is being used.

Suppose, for the sake of argument, that Korsgaard is right to adduce Wittgenstein's objections to private languages, and that these objections go through. For something to count as a reason, it has to be public; for it to be public, there have to be standards for correct and incorrect usage of the term. But this, by itself, does not render internalism about reasons false. If it did, then mere talk of internal reasons would be meaningless – and it isn't. Publicity depends on shareability in the sense that you must be able to understand that my reasons can justify, or explain, my action. Reasons must be shareable, in this sense, in order to be public. That, however, is not the required sense of 'shareability': to make the argument go through, 'shareable' must mean 'normative *for* me as well as for you'. If the normative force of my reason can be shared with you, that can mean *either* that you understand that this reason has a certain relation to me, *or* that you take this reason seriously (incorporate it into your practical deliberation, or similar). The publicity claim entails the first interpretation: Korsgaard needs the second.

Korsgaard also seems to suppose that a particular (albeit deep-rooted) fact about our psychology – that we tend to resent it when others harm us gratuitously, and to think that they have reason to stop – grounds an obligation to refrain from harming others gratuitously. Insofar as this is plausible, it contains an element of truth: we do tend to resent being harmed, and asking 'how would you like it if someone did that to you?' is often an excellent way of preventing others from behaving in this way. But suppose that I am a particularly hardened internalist: when pressed, I agree that I would resent being harmed in that context, but interpret that resentment as a disposition to engage in 'mere browbeating'. I don't *actually* think that you would have reason to desist, but will nevertheless engage in external-reasons talk in an attempt to make you desist. It certainly doesn't seem that such thinking is, on the

[42] Korsgaard 1996b: 136.

surface of things, incoherent: it doesn't involve my hearing your protestations as 'mere noise'.[43]

In a further essay – her 'The Reasons We Can Share'[44] – Korsgaard makes the stronger claim that our response to being unjustifiably harmed is not just a (contingent) psychological fact, but rather some form of rational constraint; one which is grounded in being treated, and treating others, as an end, and not as a mere means. If I am inflicting gratuitous suffering on another, and he asks me how I would feel were the roles reversed, I see that I would resent being treated as a means in that way because '[i]t would be impossible for me to consent to be so treated and so I would have to rebel . . . my victim demands that I either cease using him as a means, or give up my own claim not to be so used by others. But the latter is impossible . . .'[45] In a sense, of course, I cannot consent to be treated as a means rather than an end: consent implies choice (or autonomy), and being treated as a means implies the opposite. But there are at least two things wrong with Korsgaard's claim. Firstly, I can consent to put myself in situations where I will be treated as a means, rather than an end: perhaps I engage in a conflict where one possible outcome is that I be captured and enslaved. I may waive any claim to certain future protections, in order that I gain immediate benefits; equally, I may compromise my immediate autonomy in return for future gains (by committing myself to a period of long service, for instance).

Consenting to being treated as a 'mere means' is, therefore, only impossible where the object of my consent is one and the same as my capacity to consent. This requires that the two be contemporaneous: I can consent to give away my *future* capacity for consent, but not my immediately present capacity. It also requires that the object of my consent be related to 'my capacity to consent', under that description. I can, of course, consent to situations which, unbeknownst to me, involve me being treated as a means to an end, provided that the object of my consent is the situation (under some suitable description), rather than 'being treated as a means'. That is, consent will be problematic if and only if I am consenting to being treated as a 'mere means' where this is read *de dicto*, rather than *de re* (since the *de re* reading allows that I unwittingly consent to diminishing or disrupting my own agency).

Secondly, and more importantly, there is a logical gap between consenting to being treated as a mere means, and denying that there would be reason for people to refrain from treating me in this way. Consent is a positive matter; denial of the existence of reason to refrain is not. I

[43] Korsgaard 1996b: 143.
[44] Korsgaard 1996a: 275-310.
[45] Korsgaard 1996a: 299.

might, for instance, think that there is no reason for certain corporations not to send me junk mail – but this does not entail that I have consented to receive it. The denial of the existence of reasons to refrain only implies consent if one assumes, to begin with, that the absence of consent entails (agent-neutral) reasons to refrain from treating me in a certain way. The argument, therefore, is question-begging. Similarly, there is more than one sense in which we can withdraw consent. On the one hand, we might maintain that the other has (agent-neutral) reason to refrain from treating us in whichever way is objectionable: the withdrawal of consent, on this view, is a matter of taking a stance on what reasons there are.[46] On the other, we might simply kick and scream: withdrawal of consent, on this reading, is a matter of resistance, rather than an ethical issue. And again, *even if* 'rebellion', as Korsgaard has it, is a necessary response to being treated as a means to an end (because, let us suppose, we cannot consent to being treated in that way), that implies no more than the second reading: refusal to consent may entail resistance, but it does not imply the existence of agent-neutral reasons.

The last difficulty with Korsgaard's argument (although this is, strictly, a problem which arises at an earlier stage) is her claim that normativity depends on taking our practical identities (and hence our humanity) to be valuable. Korsgaard appears to think that the issue of normativity raises the danger of a regress: if there are to be reasons for valuing something, then we must have reason to care about those reasons, and so on. This regress can only be halted by positing something which renders the question 'why should we care?' redundant or meaningless. Here she thinks she has an advantage over the naive moral realist. There are two obvious worries about this aspect of her position: firstly, it is not clear that such a regress threatens; secondly, it is not clear that practical identity halts the regress in the required way. Our identities, on the Korsgaardian view, ground reasons for action, because they provide descriptions under which certain actions become choiceworthy; we must therefore value these identities, and hence our humanity as such. Normativity, she thinks, has to enter the picture somewhere, and it does so at this fundamental level. But from 'X is ϕ' and 'a necessary condition for X being ϕ is Y', we cannot derive 'Y is ϕ' – at least, not without some further argument.[47] We must have a description under which to rationally prefer certain actions over others, but that doesn't entail that we must rationally prefer that description (over the absence of such a ground). However, one might quite reasonably think that the case of reasons for

[46] There is a middle ground on which claims about agent-neutral reasons are attempts to browbeat our interlocutors into refraining from treating us in a certain way (where those reasons are external, in Williams' sense, to the interlocutor): this is the expressivist strategy. If expressivism is true, then 'taking a stance on what reasons there are' might be given a deflationary reading, where 'taking a stance' just means 'adopting a disposition to behave in such-and-such a way, and make such-and-such noises'. This is not a strategy which I wish to pursue here.

[47] As Gaut notes, '. . . it isn't true that if something has the power to confer some property, then the thing must possess that property.' (Cullity & Gaut 1997: 174).

action is unusual in this respect. Since we can ask for choices of actions to be justified, the justification may – at least, if Korsgaard's picture is correct – advert to our 'normative conception of our identities'. This justification is then either sufficient in its own right, or subject to further justification; in this case, there is a further justification to be given, namely that, as a fact about our humanity, we need *some* conception of our identities in order to get the process of practical justification off the ground. Our humanity, that is, needs these identities (in order to function); but 'needs' is an instrumental notion, such that if we don't care about our humanity, we don't have reason to care about our identities either. So the move from 'X being ϕ' to 'A condition of (X being ϕ) being ϕ' is misdescribed: if the argument works at all, it works because *unless* humanity is valuable, the possibility of ceasing to care about our identities remains open. As mentioned earlier, however, I suspect that this is an inconclusive argument: there are many things in the world which I might rationally cease to care about, but which I nonetheless have reason to pursue.

Furthermore, it isn't clear that the first step of the putative regress goes through. Cohen, in his response to Korsgaard, thinks that her argument turns on the following premise:

(P) 'If we did not have a normative conception of our identities, we could have no reasons for action . . .'

which is, he thinks, false.[48] The thought underlying P seems to be something along the following lines: insofar as we have reasons for action, we take one course of action to be more valuable than another (because, since we have reflective distance from our impulses, thinking how to act is not simply a matter of summing the vectors of our various impulses, but rather reflectively endorsing one, or some, of them). Or, in other words,

(R) *nihil appetimus sed sub specie boni; nihil aversamur, sed sub specie mali.*

In fact, Korsgaard phrases this differently: *nihil appetimus sed sub* ratione *boni*. So, for Korsgaard, it isn't just that choice depends on the appearance of value, but that it depends on the reasons which value provides – that is, desirability. But even if reflection (i.e. my reflective distance from my impulses) is the source of obligation for me, the most that follows from this is that my conscious awareness of you can generate an obligation for you – and not,

[48] Cohen 1996.

as Korsgaard supposes, that my conscious awareness of you can generate an obligation for me.[49]

And whilst it is certainly true that we sometimes make choices according to what seems valuable, or good, it seems equally true that we sometimes simply make choices on a whim – choices which are, nonetheless, our own (autonomous) choices. I am not here thinking of occasions when we simply give in to our strongest impulses (a desire for cake, or a cigarette), although I suspect that many of these might also properly be characterised as autonomous. Rather, the interesting cases are those where there is no good reason (at least, no obvious or accessible reason) to choose one course of action over the other, and we let ourselves be guided by our instincts. Now we might think that there *is* some good reason to choose one action over the other, namely that in certain situations where a range of acceptable options present themselves we have a reason to act on our whims. But that there is a good reason to make a choice does not entail that the choice is dependent on the desirability (or otherwise) of the option chosen. The choice, for instance, may be forced: the fact that there is some external threat compelling us to make *some choice or other* provides a strong reason to make the choice, but does not require that we treat one option as more desirable than the other. That is, the reason need not be intrinsically related to the objects of choice. Even if, therefore, we allow that there is reason to allow our whims to select between acceptable available options, this does not entail that whim-driven choosing is directed by beliefs about value. If R is true, then in such cases one choice appears to us as better than the other. But, *ex hypothesi*, there is nothing for us to ground our decision on; so if R is true, such cases cannot ever obtain, unless R is understood as a thesis about desiring *as such*. That is, unless R is understood as:

(R*) If I desire X, then I believe that X is *pro tanto* good (or *prima facie* good).[50]

However, this is both incongruent with Korsgaard's position on the relation between reflective endorsement and R, and controversial. It is controversial for at least three reasons: firstly, I may often have pathological desires (suicidal desires, for instance) which are unmotivated, or are simply brute facts about my psychology (and hence lack the corresponding belief); secondly, it violates the Humean prohibition on necessary connections between metaphysically distinct entities (although this prohibition is not uncontentious); thirdly, we are presumably much more comfortable with attributing desires to animals than

[49] C.f. Wiland 2000a: 104, '. . if we want to preserve the analogy [with intrapersonal obligation] when we turn to the topic of interpersonal obligation, at best we may conclude only that the person who is conscious obligates the person who is the object of consciousness.'

[50] C.f. Railton 1997: 62-65. Railton describes this account as a 'High Brow' view of agency.

we are attributing value concepts. Dogs may desire snacks, but that does not entail that they have beliefs about the desirability of leftovers. Maintaining the truth of R* for humans, but not for other animals, commits us to thinking that human desire and animal desire is qualitatively different, and this counts as a theoretical demerit. A full discussion of the reasons in favour of and against R* is beyond the scope of this thesis, but I will assume, for current purposes, that R* is false. In any case, there are sufficient independent reasons to doubt the success of Korsgaard's project.

2.6. Conclusions

What follows from this discussion? I suggested, at the beginning, that although there are reasons to doubt the success of these constructivist projects, there are elements of truth to them. In particular, I believe that there are two distinct but related themes which can be rescued from this discussion. The first concerns the public use of reason, wherein formal constraints, such as universal acceptability (and hence universalisability) may generate concrete proposals for action; the second concerns moral motivation and its connection to our practical identities.

The core problem with O'Neill's proposal is, I suggest, that the process of generating principles which are 'real proposals for action' and 'followable by all' needs some additional motivation. It may well be that, once we have begun to engage in the process of generating principles and adducing justifications, we will end up with the kinds of conclusions which O'Neill suggests. But such a coherentist picture of justification provides for justifications which are internal to that framework: we may, therefore, fail to care about the framework entirely, and have no reason to do so. A similar problem arises with Korsgaard's position: even if reasons are essentially shareable, and in that sense public, this entails no more than that they be comprehensible to others. We are not rationally constrained to obey the categorical imperative, nor to care about the reasons which others have for action. As a corollary of this, neither Korsgaard nor O'Neill has managed to provide a convincing argument for the existence of agent-neutral reasons.

However, both of these constructivist accounts *do* manage to capture the thought that morality deals, in part, with solutions to co-ordination problems, PD cases, etc.; and both correctly require that ethical principles be followable by all in the relevant domain, and hence universalisable (because jointly instantiable). O'Neill's account, in particular, shows how constraints on what may count as publicly acceptable reasons might serve to mediate between individual deliberation, on the one hand, and ethical principles, on the other. It is therefore worth attempting to build on this account.

There is one way in which we can amend both of the accounts under consideration, and that is to abandon the claim to agent-neutral reasons. This move would be entirely destructive as the accounts stand, since both O'Neill and Korsgaard are attempting to construct an account of why we should care about other people (refrain from injuring them, respect their humanity, and so on). So the theory needs an additional modification: namely, to restrict the constructivist account to that part of morality which, I have argued, is not accounted for by considerations of intrinsic value in the realist sense – the reasons provided by pleasure, suffering, and cognate states. That makes room for an account according to which our reasons for behaving morally may depend on our (contingent, but deep-rooted) reasons for behaving, and reasoning, in broadly social ways.

3. Restricted Constructivism

Restricted constructivism, therefore, is a more promising option than thoroughgoing constructivism. It is not a theory about morality in its entirety. Rather, it is a theory about a certain class of reasons for action: one which depends for its existence on some level of prosocial motivation, and which governs peculiarly social or political issues – co-ordination, bargaining, and so on. The theory to be developed will be similar to Rawlsian constructivism, at least insofar as it takes a notion of an idealised public deliberator as central; and it is similar to Scanlonian constructivism, insofar as it allows people's evaluations to be partly determinative of moral truth.

I should make it clear, at the outset, that I do not think that the role of morality in mediating these issues is best characterised in terms of a contract, hypothetical or otherwise. It is true that participants in the Prisoner's Dilemma, were they able to enter into a contract to secure mutual co-operation, would be well advised to do so. As an explanatory story for the existence of certain legal institutions, I take it that this is a plausible account; however, talk of contracts is implausible when it comes to discussion of justice, fairness, and cognate issues. The difficulty is that the contract must either be actual, or hypothetical, but each interpretation is problematic. We can rule out the existence of an actual contract: after all, contracts are generally supposed to be agreements between people, but the binding force of ethical propositions need not depend on the establishment of such an agreement. In the case of parties which have never met before, one still supposes that there are standards of appropriate and inappropriate behaviour, even before there is a chance to establish an agreement. Hence, if there is to be an actual contract, it must be such that it can be established automatically and unconsciously, prior to any other interaction; and this, I take it, is a curious usage of the term, even if we permit the tacit formation of contracts.

On the other hand, it does not seem that hypothetical agreements (or contracts) are binding, even when restricted to 'reasonable' or 'rational' hypothetical agreements. For instance, suppose (plausibly) that you would rationally have agreed to a wager which is heavily weighted in your favour: a coin toss, where if the coin lands 'heads' you give me £10, but if it lands 'tails' I give you £100. I am not, however, entitled to toss a coin and then claim money from you if it lands 'heads'.[51] This does not, of course, entail that considerations of how you would agree to be treated fail to bear on how you ought to be treated. Minimally, if I know that you would not agree to a certain plan of action which involves you, that provides me with some reason not to implement that plan. But I doubt that the notion of *agreement* is critical here: one might equally recharacterise these considerations in terms of what you would be happy to have done to you, or would like to have done to you, or would complain about.

3.1 Universalisability

Nonetheless, it does seem that moral discourse and practice often serve to secure outcomes which are mutually beneficial, or to help avoid outcomes which would be mutually disastrous.[52] This thought is partly captured by rule-consequentialism, but not fully: the rule-consequentialist is forced to treat the justification of moral principles in terms of *overall* consequences, rather than mutual benefit. What we are after is a system which generates rules which are acceptable to all interacting parties.

There are at least two distinct notions of acceptability available here. According to the first, a set of principles is acceptable to all parties just in case all parties have good reason to believe the set of principles to be true. According to the second, a set of rules is acceptable to all parties just in case all parties have good reason to believe these principles, *and* these beliefs are not collectively self-defeating.[53] Note that these rules need not appeal to the existence of agent-neutral reasons. I can readily hold that, for instance, pristine lawns are good for me, without thinking that there is any reason for people in general to avoid trampling on them. However, it doesn't take much effort to conclude that if everyone trampled the lawn, the lawn

[51] Rakowski 2001: 207.

[52] Part of the plausibility of contractualist theories, I suspect, derives from this fact; we would, if we were reasonable, agree to avoid outcomes which would be mutually disastrous.

[53] A related suggestion is that morality seeks to locate rules which could not be reasonably rejected by any relevant party. This is Scanlon's proposal, and I discuss it later, in Section 3.3.

would no longer be pristine. From that thought, I can derive a general objection to individuals walking on the lawn: were all of them to walk on the lawn, each of them would be equally responsible for the muddy state of the lawn. But if I have an objection against any of them in this case, then I have an objection against each of them. And if I have an objection against each of them, then that holds independently of what the other is doing. In other words, my objection doesn't depend on the person; it is impartial. And this is naturally expressed using the phrase 'one ought not to walk on the lawn'.

This line of thought turns on the assumption that a consequence of people believing lawn-walking to be morally permissible is the muddy state of the lawn. Widespread uptake of this principle leads to the loss of a common good; it is therefore in each person's interest that every person adopts a principle protecting this common good. However, notice that this is not purely a function of the consequences of universalising the principle; it is not a formal test (as in Kantian theories of ethics). As an example, consider a principle which permits leaving for work early in order to avoid heavy traffic. If everyone left for work early in order to avoid heavy traffic, the net result would be earlier heavy traffic; in this sense, the principle cannot be universalised. But this outcome does not obtain, even given that people treat this behaviour as morally acceptable. More generally, there are many behaviours which are parasitic on some widespread practice, but which are entirely acceptable. So whether or not a principle can be universally acted upon – whether or not it is possible for everyone, rather than merely each person, to get up early and beat the traffic – does not by itself determine the moral status of that principle. Rather, the salient feature here is what would happen were that principle to be universally accepted, where acceptance is crucially different from implementation. Freeloading on practices is permissible where others will restrain themselves independently of moral considerations; freeloading on practices is impermissible where we are depending on widespread moral restraint. This is particularly evident in the case of common goods: if the continued existence of these common goods depends on moral restraint, then moral restraint becomes important; if not, then not.

This being the case, it becomes clear why universalisability – in terms of whether or not the proposed course of action can be universally implemented – seems to be, and, indeed, sometimes is, morally relevant. It is morally relevant, because where people would naturally be inclined to act in such-and-such a way, the consequences of *everyone* acting in such-and-such a way become relevant. If these consequences are sufficiently bad, then it matters that there is some means by which people may co-ordinate their self-restraint. As discussed in Section 2.4, this is not a matter of universalisability acting as a constraint on rational choice:

rather, universalisability is a constraint on acceptable moral principles precisely because moral principles are proposals for group, rather than individual, action.

There is a very natural thought – which is commonly offered to the young as justification for certain moral principles or systems of justice – which asks ‘what if everyone did that?’. I am claiming, here, that we should take this thought quite seriously. We should take it seriously because it accurately characterises one of the core elements of the justification of moral principles. The case of significant common goods is one instance where this justificatory element becomes obvious. I also suspect that, whilst Korsgaard’s project as a whole seems implausible, Korsgaard does touch on something which is directly relevant to this: our sense of personal identity does underscore our interest in behaving in certain ways, particularly where the social sphere is concerned. A further good reason for taking this question seriously is that we often conceive of ourselves as social animals, members of a certain society, etc.; and even if we fail to do so, it is nonetheless true that we are animals situated in communities on which we depend for various goods. And our personal identity itself matters, insofar it is a concern which most people, quite reasonably, share. This, in turn, gives reason to be concerned about the shape of the society which we inhabit; its common goods, practices, and so on.[54]

3.2 Rawls

The later Rawls offers a picture of constructivism as providing a framework to address a very specific problem – of how, given the existence of competing conceptions of the good, to provide a framework which can accommodate these conceptions whilst remaining neutral between them. The method offered in *Political Liberalism* is to take the familiar ‘original position’ as a device for modelling ‘both freedom and equality and restrictions on reasons in such a way that it becomes perfectly evident which agreement [regarding conceptions of justice] would be made by the parties as citizens’ representatives.’[55] The original position is a thought experiment, in which we are asked to consider which principles of justice would be acceptable to agents who have knowledge of what primary goods there are, but no knowledge of the situation they will end up in (thus no knowledge of social position, gender, etc.), and no knowledge of their full-blown conception of the Good. Unlike the positions offered by other

[54] Sen notes that, in spite of thoroughgoing defection in finitely iterated PD cases being the dominant strategy (because we can induce backwards from the rationality of defection on the last iteration to the rationality of defection on the last-but-one iteration, etc.), co-operation *does* emerge. He suggests that one reason this happens is because people are thinking about what they *jointly* could do – i.e. are adopting a ‘social’ point of view. See Sen 1988.

[55] Rawls 2005: 26.

constructivists, however, the resultant understanding of justice is not made true by the outcome of the decision procedure on offer; rather, the decision procedure given by the original position serves to model the various factors which we might wish to bring to bear on the problem. The thickness of the 'veil of ignorance' (which denies agents in the original position knowledge of their various accidental attributes) serves to model the thought that it would be unjust to privilege one race over another, for instance.[56]

In one sense, the project outlined by Rawls in *Political Liberalism* is orthogonal to the account being developed here. *Political Liberalism* is interested only in the political issue of which principles of justice are reasonable, rather than true.[57] In contrast, I am interested in the truth conditions for certain moral principles, and in the grounds of certain moral reasons more generally. But I think that Rawls is correct in thinking that the original position provides a good way to model those features which need to be modelled in a discussion of justice, fairness, and so on. Two points need to be highlighted here: firstly, these concepts are concerned with how we ought to co-ordinate action, broadly construed, rather than with the question of the Good (that is, the Rawlsian project is supposed to accommodate plural conceptions of the Good, but nonetheless to enable deliberation about how society is to be structured). Secondly, there is, as Rawls argues, a close connection between the reasonableness of a principle and the extent to which it can be justified to all relevant parties – where it is not merely that the principle can be justified to all relevant parties because it is reasonable, but that the principle is reasonable because it is justifiable to all relevant parties who share an interest in finding some basis for a general, unforced agreement.

3.3 Scanlon

Scanlon, in his *What We Owe to Each Other*, gives what is perhaps the most important contemporary development of this thought.[58] He begins by taking the notion of a normative reason as primitive, and noting that we all have an interest in behaving in ways which are justifiable to other people. 'Justifiable', here, means not only that others *could* see our behaviour as justified, but that they, in some sense, cannot help but do so, insofar as they are reasonable. This is then developed into an account of moral obligation – or 'what we owe to each other'. The central argument offered by Scanlon to connect justifiability and moral obligation concerns motivation. Firstly, he claims that it seems 'phenomenologically accurate' to claim that, when I think that an action is wrong and therefore refrain from doing

[56] O'Neill (2003) argues that Rawlsian 'constructivism' is better termed 'contractualism'.
[57] Rawls 2005: 394-395.
[58] Scanlon 1998.

it, my thoughts are best characterised as thoughts about what I could justify to others.[59] Secondly, he claims that this account gives 'an ideal of relations with others which is clearly connected with the content of morality and . . . has strong appeal when viewed apart from moral requirements.'[60] I take it that both of these are plausible claims; I will forgo a detailed discussion of Scanlon's arguments for his 'contractualism' here. However, I do want to look more closely at the nature of the account on offer. Here is Scanlon's own formulation:

(C_S) An act φ is wrong if 'its performance under the circumstances would be disallowed by any set of principles for the general regulation of behaviour that no one could reasonably reject as a basis for informed, unforced general agreement.'[61]

There are several immediate worries with this formulation. Firstly, the term 'reasonably' is doing a lot of work here. The emphasis on 'reasonable' rejection is intended *primarily* to exclude 'rejections that would be unreasonable *given* the aim of finding principles which could be the basis of informed, unforced general agreement.'[62] But 'reasonable rejection' also incorporates quite general reasons: for instance, the fact that a principle would arbitrarily favour one gender over the other counts as a good reason to reject that principle. Not only can we reasonably reject a principle on the grounds that it would compromise our well-being, but we could also reject it on the grounds that it conflicts with certain important values, such as that of friendship. Furthermore, the fact that a principle would conflict with the recognition of some *impersonal* value can, in certain circumstances, provide sufficient reason to reject that principle – although this relation is indirect, and dependent on the significance of being able to live in ways which involve the recognition of these impersonal values.[63]

There are, therefore, two concerns which one might have about the use of the term 'reasonableness'. The first is that the term is not sufficiently well-specified. In light of the extensive discussion provided in *What We Owe to Each Other*, I think that this concern is misplaced. The second concern is one of circularity: if reasonable rejectability can be a function of some distinctively moral considerations, then it seems as if the account, when attempting to answer a moral question, presupposes the answer to that same question. But although Scanlon does allow that moral considerations can enter into the question of whether or not a principle is reasonably rejectable, those moral considerations need not themselves determine the course of action under investigation. For instance, when considering principles

[59] Scanlon 1998: 155.
[60] Ibid.
[61] Scanlon 1998: 153.
[62] Scanlon 1982: 273.
[63] Scanlon 1998: 213-221.

regarding our obligations to provide aid to those in need, we have to make a set of assumptions about the situation of those in need – where this, in turn, involves assumptions about what is morally the case. This is innocuous, provided that those assumptions do not themselves specify the extent of our obligations to provide aid. And where the initial moral considerations are provided by contractualist principles, there remains scope for reconsidering the status of those principles. As Scanlon puts it, '. . . a sensible contractualism . . . will involve a holism about moral justification: in assessing one principle we must hold many others fixed. This does not mean that these principles are beyond question, but just that they are not being questioned at the moment.'[64]

A stronger objection, however, is that 'reasonable rejectability' is simply too permissive as regards the principles which we may reject. Consequently, the range of non-reasonably-rejectable principles is too small. We might call this the 'gridlock' objection: if we are forced to take into account *everyone's* possible reasons for rejecting each principle, then we will be left with 'moral gridlock'.[65] As a test case, suppose that there were two equally efficacious possible principles open for consideration, either of which would form an acceptable solution to some co-ordination problem, but neither of which is without drawback. In light of the drawbacks, each of the principles may be reasonably rejected. But if both principles are rejected, the problem has no solution. And this, in turn, can be problematic – as with the case of Buridan's ass, who, being equidistant from two bales of hay, has no reason to choose one bale over the other and subsequently starves. What matters, in such cases, is not that one or more of the principles cannot be reasonably rejected – since, after all, they both can be – but that at least one of the principles is acceptable.

One way of defusing this problem is to claim that there must be some meta-principle, itself not reasonably rejectable, which requires that in these Buridanical cases we choose one or other of the two options. But although it is easy to see how such a principle (a mandate to toss coins wherever indecision between equally valuable options threatens) can solve Buridan-style dilemmas, it is not obvious how it would apply to the issue of reasonable rejectability. Whilst Buridanical cases involve intrapersonal comparison of options, the problems with reasonable rejectability involve an interpersonal conflict – more accurately, an interpersonal failure to co-ordinate. The intrapersonal case is soluble, because persons can decide to, for instance, toss a coin where necessary. But the interpersonal case is more difficult, because – on Scanlon's account – what determines the wrongness of an action is *whether or not* it is

[64] Scanlon 1998: 214.
[65] Scanlon 1998: 170.

reasonably rejectable by any of the relevant parties. This property is not affected by any coin-tossing, or any such mechanism.

Scanlon, discussing the issue of gridlock, suggests that the solution is to be found by considering ‘the question of an acceptable system of general principles of action’, where this requires considering representative persons (who are motivated to find principles to form the basis of a general, unforced agreement), rather than the actual responses of all persons. I agree with Scanlon in thinking that this is the solution to the gridlock problem for contractualists.[66] But acceptability and reasonable rejectability are distinct notions. If a principle is not acceptable, then it is reasonably rejectable. By the same token, if a principle is not reasonably rejectable, then it is acceptable. But acceptable principles may also be reasonably rejectable.

Acceptability is therefore a less demanding test than reasonable rejectability. Reasonable rejectability may lead to gridlock; acceptability need not do so. But using acceptability as a test also manages to capture the underlying thought that justifiability to others is a central element of our moral thought. Behaving in ways that are justifiable to others does not require that they *could not* reasonably reject the principle offered as justification (as the basis for a general, unforced agreement . . .), but rather that there is good reason for them to accept the principle. Scanlon denies this, claiming that ‘[t]o justify an action to others is to offer reasons . . . and to claim that they are sufficient to defeat any objections that others may have [and] also to defend a principle, namely one claiming that such reasons are sufficient grounds for so acting.’[67] But if these reasons are sufficient grounds, then it would seem as if understanding the nature of the reasons involved is incompatible with reasonable rejection of the principle on offer. However, the core interest – of behaving in ways that are *justifiable* to others – amounts to an interest in acting on reasons which *can be* held up as sufficient. Hence we need only present these reasons as *likely* to be sufficient – as reasonably considered to be sufficient, or suchlike. Consequently, this justificatory behaviour does not presuppose that the guiding principles *cannot* be reasonably rejected, only that they are unlikely to be so.

The suggestion I wish to offer, therefore, is that the core issue to be addressed in assessing whether or not a moral principle is justified is whether or not it is acceptable to all relevant, rational persons. Relevance is to be understood in terms of the parties whose observation of a moral principle will be mutually beneficial or deleterious. Rationality is to be understood in the standard, minimal sense of internal consistency.

[66] Scanlon 1998: 171.
[67] Scanlon 1998: 197.

Acceptability is to be understood in terms of taking the principle to be *justified*, where justification is to be understood in terms of taking the specified grounding features to be *reasons*. Acceptability is also a function of the consequences of acceptance. If universal (or widespread) acceptance of the principle would lead to the destruction of the underlying practice, then the principle is unacceptable. For instance, universal failure to keep promises would fatally undermine the practice of keeping promises. And if people didn't feel that they were under an obligation to keep their promises, then they would fail to keep them. Crucially, whether or not a principle is acceptable to the deliberating parties will depend on whether the general observation of the principle by the deliberating parties will be in the individual self-interest of each person.

Of course, current practice will impact on what people take to be reasons, and on which principles count as acceptable. This is often what determines which of multiple acceptable solutions to a problem should be followed. Given the current practice of property ownership, for instance, it becomes true that we ought not to appropriate others' possessions, even if communism provides a viable alternative method of structuring society. That is, even if communism and property-ownership each provide *ex ante* acceptable solutions to the general problem of structuring relations between persons, use, and objects, the entrenchment of one system may suffice to render the other unacceptable.

So moral deliberation is a matter of thinking (at least partly) about what *we* are to do (rather than simply about what I am to do). I can think about whether or not we ought to endorse a certain principle, or set of principles; which amounts to thinking about whether or not these principles are justified. And, if the story on offer is correct, a multiplicity of moral codes may turn out to be potentially viable. But we can explain why there is reason to adopt the current solution within a given society, given the difficulty of changing the moral code which is currently accepted by society, and given the consequences of individual defection.[68] For instance, it is an entrenched convention within the United Kingdom that we drive on the left hand side of the road. Another viable solution to the co-ordination problem would be to arrange things such that everyone drives on the right hand side of the road: this is implemented in America and on mainland Europe.[69] Either solution will suffice, but we rationally defer to whatever is currently implemented. Likewise with the constructivist part of morality. Suppose that we might structure society such that individual property ownership is permissible, and also such that no individual property ownership is permitted – and, furthermore, that each of these structures is satisfactory. But the success of these structures

[68] That is, given that the current solution is satisfactory.
[69] Driving in a zig-zag pattern, of course, is not a viable solution.

depends on the consistency of their implementation: the presence of a minority who disregard individual property rights, within a property-owning society, will have an undesirable impact; the minority will, justifiably, be punished.

This account runs close to Copp's 'society-centred' ethic. According to Copp, '[a] code is justified as a moral code in relation to a society just in case the society would be rationally required to select the code to serve in it as the social moral code, in preference to any alternative.'[70] But the account on offer differs from Copp's account in two ways. Firstly, the key justificatory element is acceptability to each person in that society *qua* widely-adopted code. This avoids the worry that the notion of societal rationality is, at best, poorly understood; it also guarantees that each person, insofar as they are reasonable, plays a determining role in which moral principles count as justified. Secondly, the process of justification focusses on principles which are offered as proposals for specific contexts. The extant moral code forms part of this context. So whilst there may be cases where the code comes in for wholesale revision (I am thinking in particular of large-scale religious conversion, revolution, and so on), the primary element in the constructivism on offer is the individual moral principle. Thirdly, and most importantly, I think that Copp's emphasis on optimising is mistaken. Although he does address the possibility of multiple optimal solutions, a moral principle which meets the criteria which I have set out – broadly grouped together under the aegis of 'universal acceptability' – may be suboptimal, from the point of view of the society, but nonetheless justified. Societies are compelled to adopt moral codes not because moral codes maximise the extent to which these societies flourish (although moral codes do, of course, facilitate flourishing), but rather because the wholesale absence of a moral code would be disastrous.

Although I think that this account is correct, it would clearly require a much more substantive work to defend the current proposal fully. However, what I am currently interested in is the metaethical plausibility of an account of this general form. Specifically, I am interested in whether an account which explains the property of wrongness in terms of the acceptability of a proposal for mutual co-ordination of action can meet certain key objections. Having discussed a range of particular constructivist positions, I now wish to turn to some more general issues: the scope problem for constructivists, and how the hybrid theory solves this; the categoricity problem, and how constructivists should approach it; and Horgan & Timmons' 'Moral Twin Earth' objection.

[70] Copp 1995: 104.

4. Objections

4.1. The Scope Problem

The scope problem applies to any form of constructivism or contractualism which makes the moral status of an action (or outcome, etc.) depend on some specified group of entities. For instance, a form of contractualism which made moral wrongness a function of mutually-unacceptable solutions to PD cases would have a limited scope. Specifically, it would be an account of impermissible actions for those agents involved in the PD case. Similarly, social contract theories of ethics have traditionally had difficulties accounting for obligations to those outside of society. Any theory which attempts to found ethics in mutual benefit will restrict the scope of ethics to those parties who are in a position to receive mutual benefit. Theories which interpret ethics as emerging from constraints on rational agency and relations between rational agents (as with Kantian theories) have difficulty in accounting for our obligations towards animals, or towards irrational humans (babies, young children, the severely mentally disabled, and so on).

For any constructivist or contractualist account which attempts to provide an account of ethics in its entirety, this is unpalatable, because it would seem to entail that we do not have obligations towards animals, irrational humans, the distant poor, and so on. One might try (as Kant does) to account for these obligations as somehow derivative: for instance, it might be that we ought to treat animals kindly because that is a good way to inculcate moral virtue; or we might think that we are justified in criticising those who treat animals badly because those who treat animals badly are liable to treat people badly. But this is suspect, for at least two reasons. Firstly, it is far from clear that moves of this sort can be made to work. It is quite possible to treat animals well but people badly, and vice versa. [71] Secondly, and more importantly, this falls prey to the charge of revisionism about justification discussed previously. It is implausible to suppose that animal cruelty is wrong because of its effect on our character (although this may be a relevant consideration). Similarly, it is implausible to think that our obligations not to harm animals have an entirely different metaethical status to our obligations not to harm people.

However, restricting the scope of constructivism, and combining it with a realist hedonist utilitarianism allows us to avoid these problems. The assumption of realist hedonist utilitarianism allows us to maintain that our obligations not to harm people have at least the

[71] Hitler was, as is often pointed out, vegetarian.

same meta-status as our obligations not to harm animals – but there may also be additional constructed reasons not to harm persons. More generally, I take it that revisionism about justification is unpalatable precisely in those cases where the realist account is correct. Where the content of morality concerns solutions to co-ordination problems, revisionism about justification becomes palatable. When explaining why it is wrong to break promises, for instance, the question 'what would it be like if everyone did that?' becomes relevant. This is so, because it must be appropriate to take the point of the practice into consideration when engaging in ethical deliberation – and the point of the practice is, at least in part, to co-ordinate action. So restricted constructivism can be made to avoid the scope objection. I discuss this issue in greater depth in Chapter Five.

4.2. Categoricity

The worry about categoricity runs as follows. It is a conceptual truth that moral requirements are categorical, i.e. if they apply to an agent, they do so independently of that agent's desires, interests, or ends.[72] Since constructivist morality is grounded in the need for social creatures to solve the co-ordination problems with which they are faced, the requirements of constructivist morality apply to all rational social beings appropriately situated. In this sense, it need not be binding on all rational agents as such, since there is at least conceptual room for rational beings who are not social beings relevantly situated (imagine a race of wholly self-sufficient, independent aliens). But constructivism (at least, as characterised thus far) looks like it will *not* yield categorical constraints; rather, they will be conditional (*if* you have a desire to behave in ways which . . . etc., *then* you ought to . . .).

One way to make sense of categoricity runs as follows. For a moral requirement to be categorical is for it to be binding on some set of entities independently of their accidental properties. Moral requirements might be understood to be binding on all rational agents as such, for instance. But given the peculiarly social grounding of constructivist morality, this would be implausible. Being a rational being does not, *per se*, guarantee that that entity will have any of the required desires, ends, or interests. A second possibility is to understand moral requirements as binding on all social beings as such. This supposes that all social beings will have some shared set of desires, ends, or interests. Understood in this manner,

[72] There is another sense in which moral requirements can be said to be categorical, namely the sense in which they are binding on all rational agents as such. But 'categorical' is properly contrasted with 'hypothetical', and the primary conception of a hypothetical imperative is one which refers to an agent's ends. Furthermore, I take it that the central thought behind the attribution of categoricity is that we cannot escape moral requirements simply by ceasing to care about them. I discuss this in more detail below.

categoricity is a matter of inescapability, rather than desire-independence. Social beings, by their nature (and possibly situation), are bound by constructivist moral requirements. In this sense, moral requirements are indeed categorical, i.e. binding on all social beings as such. But the phrase ‘social beings’ is, here, underspecified. In a minimal sense, the phrase includes all agents living within a society. On this construal, it seems hard to claim that there are any inescapable desires (for instance, a desire to behave in ways justifiable to others is widely shared, and prudent – but a particularly talented and capricious knave might very well lack this desire, and have no reason to acquire it). In a more robust sense, ‘social beings’ might be just those with a certain set of motivations and interests. However, this seems like an objectionably *ad hoc* stipulation. This being so, the claim that social beings share some set of properties which grounds constructivist moral requirements turns out to be either trivially true, or false. Bear in mind that the original worry was that moral requirements should not be understood as binding in virtue of one’s desires – they should be non-hypothetical. If the sensible knave claims that she lacks the required motivations, common-sense morality holds that she still ought to (for instance) keep her promises, respect the property of others, and so on. Even if we understand categoricity as a matter of inescapability, the stipulation that social beings share a certain set of motivations will fail to account for this: the sensible knave cannot escape the moral requirement by ceasing to be a social being in the stipulated sense. So the worry about categoricity remains.

We might attempt to tackle this worry head-on. For instance, Korsgaard assumes the existence of unconditional obligations, and then attempts to explain their categoricity:

> ‘It is our conceptions of ourselves that are most important to us that give rise to unconditional obligations . . . When an action cannot be performed without loss of some fundamental part of one’s identity, *and an agent could just as well be dead*, then the obligation not to do it is unconditional and complete.’[73]

The normative force of these unconditional obligations has to trump any other considerations (or else they would not be unconditional), and hence they must be grounded in something which has this kind of trumping potential; that is, the obligation must arise from something whose value is overriding. For Korsgaard, this is personal identity – because there is nothing worse than death, *except for the loss of personal identity*, which is itself tantamount to death. Now this is rather too fast: one might think that there are *pro tanto* obligations, which bind unconditionally, in the sense that they provide us with reasons for action no matter what, but

[73] Korsgaard 1996a: 102, my emphasis.

which are not overriding. That an action will harm another person, for instance, may provide us with a (moral) obligation to refrain from that action – but there might, for all that, be overriding reason to pursue that action. Korsgaard's analysis is an attempt to uncover categorical *and overriding* reasons for action, and this, I suggest, is too strong.

On the other hand, if the construction procedure depends on widespread but contingent desires – the desire to behave in ways which are justifiable to others, for instance – then it seems as if the resultant obligations will fail to provide reasons for those who lack the relevant desire. This leaves us with two problems: how to account for the appearance of categoricity within moral discourse, on the one hand, and the issue of whether moral constraints are necessarily categorical, on the other. Now it does seem that categoricity marks an important distinction between norms which are 'merely' conventional, and norms which are specifically moral. We ought to keep our promises even if we don't want to; we cannot escape our moral obligations by ceasing to care about them. There is a shared, strong, intuition that we cannot escape the demands of morality simply by failing to care about them: the thief who decides not to abide by prohibitions on theft still ought to keep to them. But note that this entails only that moral demands are independent of one's immediate desires – not that they are categorical. We may also think that we cannot escape the demands of morality simply by being appropriately positioned, but this intuition is much less strong: there seems to be ample scope within common-sense morality for a degree of relativism, such that the demands of morality may vary across societies.

So there is a sense in which categoricity is a necessary component of morality. Making moral obligations dependent on individual whims would defeat the point of the practice. But the existence of entirely amoral persons might nevertheless be thought to pose a problem for the constructivist; this is the familiar worry about external reasons discussed in the previous chapter. If it is true that we ought not to steal, then everybody has a reason to refrain from stealing. But the hardened amoralist might entirely lack the relevant motivations. Hence, for him, our moral discourse amounts to 'mere browbeating'.

Accepting this conclusion is, I think, innocuous. It is worth bearing in mind that hardened amoralists of the required sort are few and far between, hence moral discourse does not, in general, amount to 'mere browbeating'. Even those who do not care about morality *per se* still, by and large, have reason to behave morally; even if a 'sensible knave' sees nothing wrong with theft, it will still (in general) be in her interest to refrain from stealing, because thieves tend to be punished, if not formally (by the law) then informally (by the disapproval of others). She may have no reason not to steal where she can get away with it, but since people tend to be poor judges of when they can get away with it, she is best advised to adopt a policy of refraining from stealing. Given the existence of a widespread moral practice, knavish behaviour is generally (although not always) imprudent.

Nonetheless, let us allow that there are cases where agents have no reason to act in accordance with the constructivist part of morality. Since the truth of these moral principles is fixed by their acceptability as solutions to co-ordination problems, it is true to say of the knave that she ought to refrain from stealing. That this amounts to ‘mere browbeating’ is unproblematic, because the practice *as a whole* does not amount to ‘mere browbeating’. Constructivism need not be understood as committed to thoroughgoing categoricity, in the sense of the provision of reasons for action for all rational agents independently of their particular ends. But it is clear to see why the discourse should retain the appearance of thoroughgoing categoricity: after all, the principles under discussion are proposals for mutual co-ordination of action. As such, they are general, of the form ‘one should not break promises unnecessarily’, and are binding on agents *qua* social beings, appropriately situated. This, in turn, results in the appearance of categoricity. Strictly speaking, the reasons generated by these requirements will not be categorical: the highly sophisticated amoralist may have no reason to comply with their demands. But this is innocuous; we still have overwhelming reason to talk of these requirements as categorical.[74]

4.3. Moral Twin Earth

Constructivism, at least in the formulation set out by Scanlon, is not to be understood as an attempt to provide an analysis of the concept of moral wrongness – of what is meant by ‘morally wrong’. Rather, it is supposed to be an account of the property of moral wrongness. Given this, Moore’s ‘Open Question Argument’ does not apply. However, there is a successor argument to the OQA, namely Horgan and Timmons’ ‘Moral Twin Earth’ argument(s).[75] Initially designed to work against synthetic ethical naturalism (according to which moral properties can be identified *a posteriori*, yielding property identities akin to ‘water is H_2O’), the argument runs as follows. Suppose that Earthlings discover that moral wrongness, for instance, is to be identified with causing harm to innocents. Suppose also that on Moral Twin Earth it turns out that the term ‘moral wrongness’ picks out a different property: that of violating the Ten Commandments. Aside from this, Earthlings and Moral Twin Earthlings are alike in all relevant respects. Now if the synthetic account were correct, then Earthlings and

[74] This position is therefore close to the proposal outlined in Joyce 2001, which presents an error theory regarding the commitment to categoricity coupled with a non-revisionary proposal for moral discourse.
[75] Horgan & Timmons 1991, 1992a, 1992b, 1996, 2000, and elsewhere.

Moral Twin Earthlings would not be able to discuss the topic of moral wrongness: when Earthlings claim that a certain action is 'wrong', and Twin Earthlings claim that it is 'not wrong', they are talking past each other. But our intuition in these cases is that genuine disagreement *is* possible, and hence synthetic ethical naturalism cannot be correct. As far as the explanation of this intuition goes, Timmons suggests that the best explanation adverts to the specifically evaluative nature of the Earthlings' and Twin Earthlings' ethical discourse.[76] If the terms were purely descriptive, then the discovery that Earthlings and Twin Earthlings use the term 'wrong' in ways governed by distinct natural properties would rule out the possibility of genuine disagreement.

Unlike the thoroughgoing synthetic ethical naturalist, however, the hybrid theorist is at an advantage. The broad picture, remember, is that moral wrongness is to be explained in terms of universal acceptability, where universal acceptability in turn serves to track solutions to co-ordination problems, broadly construed. It is this feature that allows moral discourse to play the distinctively social role that it does. On Twin Earth, let us suppose, use of the term 'morally wrong' is causally regulated by the Decalogue. Now although I share Horgan and Timmons' intuition about 'genuine disagreement' in the case discussed, it is worth noting that this intuition does not extend to *any* possible regulatory property. The term '*tabu*' as used by the Polynesian islanders, for instance, may share the formal aspects of the term 'morally wrong' as used by contemporary westerners – but it is far from clear that the two terms are, for that reason, intertranslatable.

In the cases under discussion, however – the constructivist and the Divine Command Theorist, or Horgan and Timmons' example of consequentialist and deontologist properties – it is not only the formal properties of the discourse which are shared. There is one further and quite general feature of these discourses, being that they all serve to co-ordinate action (in a broad sense). The suggestion, therefore, is that this shared feature is enough to account for the intuition of disagreement. In a similar vein, Mark van Roojen suggests that, in the Moral Twin Earth cases,

> 'The two populations are both tracking well enough what makes sense to do. At most one population might be disposed to get it right in the long run . . . But both populations may be using "right" to refer to what makes sense to do.'[77]

[76] Timmons 1999: 62-69.
[77] van Roojen 2006: 178.

Since the population of Twin Earth is relevantly similar to ours (in terms of physical and social constitution), the relations between acceptability, mutually acceptable solutions to PD cases, and moral wrongness will also be relevantly similar. This, by itself, should provide enough common ground to permit genuine disagreement.

Possible counterexamples are given by cases where we would wish to adopt some form of ethical relativism. Now it is a standard problem with thoroughgoing ethical relativism that it has difficulty accounting for the possibility of inter-societal moral disagreement. This might seem problematic, since much of our ethical theorising proceeds on the assumption that there is a single right answer, that apparently contradictory ethical positions must indeed be contradictory, and so on.[78] But if I am correct in arguing that moral discourse often serves to locate satisfactory solutions to PD cases, then it is quite clear that there may sometimes be more than one such solution. This applies equally to the case of optimal solutions: there may often be more than one optimal solution to such cases. For the sake of argument, suppose that one satisfactory solution may be located by the Decalogue; another by a Kantian theory of ethics. In this case, although the natural properties which govern the content of the ethical theories differ (supposing that the content of the Decalogue and the universalisability of maxims are both natural properties), the general purpose of the discourse is the same in each case – hence the possibility of genuine disagreement, even in the face of equally viable solutions to the problem at hand. Disagreement, here, must be understood as disagreement about which solution to implement.

In any case, it should be borne in mind that the original problem for the synthetic ethical naturalist, as diagnosed by Horgan and Timmons, is a function of the 'purely descriptive' nature of the suggested property identity. But the constructivist account allows normative (and evaluative) properties to enter into the account. Suppose, then, that the account being given holds true for the inhabitants of Earth, but fails to account for the Twin-Earthlings' use of the term 'morally wrong'. Since the normative properties which are involved in the account relate not only to the content but also to the point of the practice, it follows that the *purpose* of talk of 'moral wrongness' on Earth is different from the purpose of the (orthographically similar) talk on Twin-Earth. Learning of this fact should, I think, be sufficient to convince Earthlings and Twin-Earthlings that their apparent disagreements are just that: they are external rather than internal to the practice. The analogy here with the case of 'moral wrongness' and 'tabu' is direct.

[78] See Chapter One.

5. Conclusion

The aim of this chapter was to motivate and outline one viable form of constructivism, with a view to considering whether this theoretical approach, combined with the realism of the previous chapter, can form a viable metaethical theory. The outlined form is intended as a test case, rather than a definite answer. I began with an attempt to motivate constructivism, using the ‘Prisoner’s Dilemma’ as a starting point. Our moral discourse and practice, at least in part, serves to solve a quite general co-ordination problem. This point has been correctly identified by metaethicists of various stripes, including social contract theorists such as Hobbes and Rousseau, expressivists such as Blackburn and Gibbard, and constructivists such as O’Neill. However, the metaethical position which most closely preserves our pre-theoretical commitments is, I have argued, constructivism: it allows us to make sense of moral claims as truth-apt, in some sense objective, and so on, whilst avoiding any commitment to spooky, non-natural properties. A moderately detailed discussion of Gauthier, Korsgaard, O’Neill and Scanlon led to an outline formulation of constructivism according to which actions are morally wrong for a group of agents if they are forbidden by principles the general adoption of which is universally acceptable to all reasonable members of that group. Acceptability in turn introduces substantive constraints: the principles must be able to be jointly implemented, and their joint implementation must be in the interest of each member of the group. The reason-giving force of this procedure is dependent on certain widespread but contingent motives. This leads to an error theory about the nature of the resultant moral principles: although the commonsense view of morality is that ‘sensible knaves’ ought not to steal *regardless* of their motivational set – and that this injunction provides them with reasons not to steal – there may be anomalous cases where those knaves have no such reason. Nonetheless, this appearance of categoricity is justified as an artefact of the construction procedure, since moral principles are *general* and *whim-insensitive* proposals for action. The resultant theory is therefore non-revisionary.

I concluded with a discussion of three key worries for constructivists. The scope problem, which applies to constructivist and contractualist accounts in general, is that these accounts appear to inappropriately restrict the scope of constructivism. This objection is, I argued, closely related to the problem of ‘revisionism about justification’ discussed in Chapters One and Two. Restricting the scope of constructivism, and incorporating realism about hedonic value, solves this problem: we have reasons not to harm other people just as we have reasons not to harm animals, and to aid the distant poor (although there may be additional reasons in the case of rational persons). I discuss this in greater detail in the next chapter. The categoricity problem is that morality purports to categoricity, but that our reasons for following constructivist morality are contingent. I suggested that the correct constructivist response is to adopt a non-revisionary error theory: although, strictly speaking, constructivist reasons for action are not categorical, the appearance of categoricity within constructivist morality is warranted, and hence ought to be retained. Finally, I addressed the Moral Twin Earth cases. Here, the possibility of genuine disagreement is secured by commonalities in the function of moral discourse and practice. Where these commonalities do not obtain, there is good reason to characterise the disagreement as merely apparent, rather than genuine.

CHAPTER FIVE

Applications

1. Introduction

The previous two chapters derived an account of morality which is partly realist, partly constructivist. The content of realist morality is, I have argued, a form of realism about intrinsic value – specifically, the value of hedonic (or anhedonic) states. This may be characterised as a form of hedonistic utilitarianism, although, unlike utilitarianism, I offer an account of *pro tanto* reasons for action, rather than of what we ought to do, all things considered.[1] As far as the constructivist part of morality is concerned, the account is broadly contractarian. Combining these theories yields theoretical benefits at the metaethical level, but also – as I argue below – at the applied level. I begin with a discussion of the moral status of animals, and our obligations towards them, before moving to a more general discussion of some hard cases (children, cognitively impaired humans, and so on). I consider briefly the relation between the scope of morality and its purpose, with particular reference to Roger Scruton, before turning to a discussion of international and intergenerational morality. I end with an examination of Derek Parfit's 'repugnant conclusion'. In all of these cases, allowing that there are two distinct sources of moral reasons yields an account which is intuitively plausible, and dissolves some potentially difficult problems. This counts as a further reason in favour of adopting the hybrid theory.

2. Animals and Moral Status

The issue of how we ought to treat animals is complex. On the one hand, current practice has animals being raised and slaughtered, often in conditions which cause significant suffering, for their meat, hides, fur, and so on. We use animals for pest control, clearing mines, and guiding the visually impaired. On the other, a large number of people attempt to eat only meat that is humanely reared and slaughtered; refrain from eating meat and consuming products which necessitate the killing of animals altogether (although, curiously, whilst moral reasons are often adduced for vegetarianism, vegetarians tend not to treat meat-eaters as morally reprehensible); or refrain from consuming products which involve the exploitation of animals.

[1] See Tannsjo 1998 for an account of a thoroughgoing hedonistic utilitarianism.

With such differences in practice and stance, the question is a pressing one. If there is any truth to the slogan, ‘meat is murder’, then the scale of the current atrocity warrants immediate action. The core of the question, however, is a metaethical issue: what is it that determines moral status and obligation? The issue is problematic, because for almost any feature which (putatively) determines moral status, there are human beings who lack this feature, and animals which possess it, at least to some extent. The exception, of course, is being (genetically) human: but this is problematically vague, and not obviously morally relevant. Whilst we can readily differentiate between humans and dogs, it is not clear where the dividing line between humans and dogs lies - if, indeed, there is a sharp line. So it is quite plausible to think that the notion of being genetically human is vague: rather than there being an identifiable point beyond which incremental changes to the organism’s DNA turns a human into a dog, our species concepts simply are not sensitive to such incremental changes. Worse, the concept of a ‘species’ is itself problematic.[2] In any case, there does not seem to be good reason to take genetic make-up *per se* to be morally relevant: the stock cases here are instances of human beings who lack moral status (the permanently brain-dead, for instance), and the possibilities of non-human creatures (angels or aliens) who possess moral status but are not genetically human. Nor would we exclude an otherwise normal person from membership of the moral community if it turned out, on closer inspection, that some unusual mutation rendered them non-human. So the simple view, on which the only creatures which have moral status are human beings, seems implausible. This leaves even sophisticated constructivist or contractualist accounts at a disadvantage. On Kant’s view, for instance, our reason to refrain from torturing animals for fun is derivative from the negative impact that such activity would have on our characters. But this is entirely the wrong kind of reason. Torturing animals for fun is wrong regardless of whether or not it would have a negative impact on the torturer. And there is something much worse about the case where a torturer *actually* inflicts wanton suffering on an animal, as opposed to the case where a torturer *merely believes* that she is doing the same. Hence what is morally the case does, after all, depend on the recipient.[3] This strategy also shows that the wrongness of the action cannot (in such cases) be purely a function of whether or not it expresses the vice of cruelty.

[2] See O’Hara 1993. The core of the problem is that we have no principled way of distinguishing between major intra-species variation and minor inter-species variation. O’Hara speculates that this may be dissolved by a historical approach, on which ‘species are the ultimate terminal taxa in evolutionary trees: within species relationships are reticulate, and between species relationships are branching.’

[3] See Bernstein 1997. Bernstein discusses the possibility of extending Rawlsian contractualism to ‘marginal humans’, and concludes that such humans can be incorporated into the moral framework, but at the cost of attributing instrumental, rather than intrinsic, value to them.

At the opposite end of the spectrum, the claim that all organisms have moral status seems equally implausible. We have no qualms about damaging or killing plants; likewise for bacteria, moulds, and so on. Torturing sentient creatures for fun, on the other hand, seems morally suspect, so we might think that what matters is the capacity to suffer. But this fails to accommodate the thought that there is something worse, *ceteris paribus*, about inflicting suffering on persons, rather than animals. Those who think that suffering (and related states) are the only relevant considerations are required to deny this claim.

The hybrid view has, I will argue, a significant advantage when it comes to addressing these questions. It accommodates the thought that we should avoid inflicting unnecessary pain on sentient beings, whilst allowing room for the thought that there is a morally relevant distinction to be drawn between persons and animals. Furthermore, it allows that the obligations to refrain from harming animals, and to refrain from harming persons, have (at least in part) the same metaethical status. There may be a further (constructed) obligation to refrain from harming persons; but there is in both cases an obligation grounded in the intrinsic disvalue of suffering. The purpose of this chapter is to explore the implications of the hybrid view for our obligations towards animals, and other hard cases - infants, the severely mentally impaired, and so on – and to argue that the resulting picture is plausible, and so lends support to the hybrid view.

3. Consequences of the Hybrid Theory

According to the hybrid view, suffering (and similar states) are intrinsically disvaluable, and pleasure (etc.) intrinsically valuable. This is the realist domain, and the value is taken as being derivative from the phenomenology. There are also obligations which obtain between persons: these address solutions to co-ordination problems (broadly construed), and are mediated via moral principles – and so depend on each party being able to moderate their conduct accordingly. So understood, this 'constructivist' domain ranges only over rational, social animals. It seems plausible (although I do not wish to argue for this claim here) that rational animals must be language-using animals, and must also have a sense of self. Certainly, creatures able to bind themselves by moral principles must be language-using; and one might think that in order to act intentionally, creatures must be able to conceive of *their* aims and goals, and this requires having a sense of self. But this is peripheral to the thesis.

The core claim is that there are two sources of moral reasons: one which is shared by all sentient beings, and one which is specific to rational, social animals. But (of course) each of these admits of degrees: pain is not an all-or-nothing affair, and neither is rationality. As regards sentience, the more vivid and intense the pain, the stronger the reason to avoid inflicting it. Less clear, however, is the question of rationality; it is not obvious that the constructivist account allows for degrees, despite the fact that the capacity for rationality (or sociability, or – presumably – language use) *does* come in degrees. The problem is that whether or not we have obligations towards other entities depends in part on whether or not they are rational, but the *strength* of these obligations is not a function of the degree to which they are rational (contrast the case where we ought to revere persons insofar as they are rational: *this* obligation does admit of degrees).[4] To borrow a Kantian metaphor: whether or not we have constructivist obligations to others depends on whether they are members of the Kingdom of Ends. But membership of the Kingdom of Ends is an all-or-nothing affair.

We might, therefore, worry about our obligations towards dolphins, chimpanzees, and the like. But a central element of the constructivist account is that it applies only to those who can deliberate over, and are capable of moderating their conduct by, moral principles. So the real issue, here, is whether membership of this category comes in degrees, and, if so, whether that matters. Now the extent to which persons are able to deliberate over, and moderate their conduct by, moral principles does vary: some people engage in careful thought, and display self-control; others do so to a lesser extent. But the vast majority, at least, have the relevant capacity, even if their success in exercising it varies. Again, the problem can be refocussed: on the assumption that this capacity is not an all-or-nothing matter (is a capacity with fuzzy edges, so to speak), we might respond by claiming that, although some people have more of this capacity than others, the mere fact that they have this capacity grounds their membership within the constructivist sphere. It is clear that normal adults have the relevant capacity; it is equally clear that inanimate objects do not. If the constructivist account does not allow for degrees, then there must, at some point, exist a threshold. It is hard to see how to specify this threshold in a non-arbitrary fashion. Nonetheless, our inability to specify this threshold does not mean that the threshold does not exist; rather, our existing conceptual framework may be too coarse-grained to specify it (there may be vagueness at the conceptual level), or we may simply be unable to know where this threshold is (that is, this vagueness may amount to an epistemic difficulty). The specification problem by itself does not pose a serious threat to constructivism. It would become problematic if the number of borderline cases were large in comparison to the number of determinate cases; fortunately, this is not so. We are therefore left with a number of worrisome cases where the capacity is either barely present, or far from fully developed. This, however, need not be a problem, insofar as we can specify (non-

[4] C.f. Scruton 1996: 28-29. Scruton claims that membership of the moral community requires rationality, freedom, the desire to obtain co-operation, the willingness to co-operate, and the ability to understand and accept obligations.

arbitarily) a way of handling these cases. The most obvious way of doing this would be to simply extend the benefit of the doubt, and treat them as if they were full moral agents. Given the importance of the threshold in question, there is good reason to treat all borderline cases as above the threshold.

There are actual instances of such borderline cases: most obviously, children, and the severely mentally impaired. On the assumption that we have (realist) obligations towards each of these cases, what are we to make of their relation to the constructed part of ethics? Talk of their inclusion in the constructivist sphere needs, here, to be clarified. There are at least three issues: whether we have constructivist obligations towards them (for instance, whether we have obligations to respect their personal property); whether they have constructivist obligations towards us (for instance, whether they have obligations to respect our personal property); and what role they play in determining our moral obligations (whether they are to be counted in the assessment of constructivist moral principles).

If we assume (reasonably) that we do not have constructivist obligations towards animals, then it is natural to conclude that our having these obligations depends on the nature of the recipient. Since the constructed sphere counts, in the determination of its principles, only those who share the capacities for motivation, and standards of justification (as with the Rawlsian suggestion that the principles should be acceptable to all under certain conditions, or the Scanlonian that they should not be reasonably rejectable by anyone motivated to find principles which all could accept as the basis for an unforced, general agreement). We therefore only have constructivist obligations towards other moral agents (in contrast with the realist case, where no agency is required).

So we have obligations to, for instance, keep our promises to children precisely to the extent that they share the relevant concepts and capacities, and to the extent that they are able to be motivated to respect the promises which they have made to others. Similarly with the severely mentally impaired: to the extent that they are able to be active participants in the moral community, we have constructivist obligations towards them. Other obligations, such as those concerning property, are less clear-cut: although it is not unreasonable to think that we ought to respect children's property rights over their toys (those gifted to them by other adults, in particular), we might also conceive of the child's toys as being *owned* by the parent (or guardian, etc.), but given over for use by the child.[5] Children's bank accounts, also, are taken

[5] There is an independent legal issue: persons in certain situations may not be able to own property, or this right may be subject to waivers (as when, for instance, the police confiscate the property of

to be held in trust for them by their parents or guardians, with ownership transferring to the child on maturity: here the child acts as the beneficial owner of the money held in trust, while the legal owners, the parents or guardians, are under a duty to keep the account for the child's (future) use. In the limit, extreme youth or mental impairment may result in an absence of moral agency; this would be problematic if all moral obligations depended on the moral agency of the recipient. But again, as with animals, our obligations may also derive from the recipient's sentience.

There is an immediate problem with this suggestion, namely that there may be – indeed, are, and have been – cases where people lack the relevant concepts, but where we nonetheless think that we ought to respect their property. Suppose that a boatload of colonists arrives on an island where the indigenous population lacks an institution of property altogether: do the colonists not have an obligation to treat the indigenous population as though they had some title to the land? On the one hand, on the account being given, it does seem that the natives' incapacity gives the colonists a pass to behave as they will. The colonists, then, may annex the land, put up walls and fences, and exclude the indigenous population from their habitat (although not from their means of subsistence, on account of the suffering which would ensue). This seems problematic. But, conversely, the natives are under no obligation to respect the colonists' property: indeed, the colonists may have property rights over certain goods *against other colonists*, but not against the natives. So whilst the colonists may establish smallholdings without wrongdoing, their property rights over these smallholdings yield obligations for other colonists, but not for the natives.

However, property rights do not only entail obligations on behalf of others; they also bear on the issue of whether, for instance, we are entitled to forcibly exclude others from the use of certain goods. In the case of the colonists, the issue is not only whether the colonists can make certain demands on the natives, but whether they are entitled to displace them. The question deals with how we are to treat others, as well as how we are to go about engaging in processes of moral discourse and justification. We need to know what kind of behaviours could be justified – but that raises the question of what justification, in this context, amounts to.

There are four possibilities here. The first is that justification is a relative matter; the actions of the colonists may be justifiable-to-the-colonists but unjustifiable-to-the-natives. For various reasons, this possibility is unpalatable – but in any case, can be put to one side, since the relevant issue is whether or not the colonists' behaviour is justifiable to the colonists (or,

criminals). But this is a distinct issue from the question of when we may take ourselves to have obligations (derivative from property rights) regarding others.

closer to home, whether there is any difference between how we are to treat contracting and non-contracting parties). The second possibility is that the colonists' property rights justify enforcing a monopoly on those goods only against other colonists. The third is that the colonists' property rights justify enforcing a monopoly on those goods against all other parties, and the fourth is that no special justification obtains.[6]

To recap: the initial worry was that if property rights are dependent on some contingent feature of persons (e.g. their conceptual framework), then there will be cases where this feature is absent, and where those persons would consequently lack property rights. Conversely, I suggested, there will be a range of obligations which those persons also lack – in the case under consideration, regarding others' property rights. This lessens, but does not remove, the initial worry: if talk of property rights, and corresponding obligations, does not apply to the natives, then the colonists cannot assert these rights or obligations against the natives, and hence cannot claim the natives' land for themselves. But there still seems to be something unpalatable about a situation where colonists displace natives from their traditional hunting grounds and relocate them to, for instance, a specially-designated reserve – even where that reserve contains sufficient resources to meet their needs.

The obvious route to capturing this unpalatability is via a constructed principle which enshrines the value of freedom, or autonomy; that is, one which stipulates that persons ought not to have their freedom restricted by others, absent some special justification. If the natives object to the way in which they are treated by the colonists, then there will be terms in which these objections are couched. Consequently, on the assumption that there is some shared conceptual framework, there will be a principle couched in these general terms which gives rise to obligations on behalf of both colonists and natives. There are two concerns with this suggestion, which run parallel to the previous discussion: firstly, this assumption stands in need of justification; secondly, even without this shared framework, we might think that we would nonetheless have special obligations towards persons, obligations which the constructivist account hopes to capture.

On the question of shared conceptual schemes, one could take a Davidsonian line, and claim that interpreting creatures as persons requires interpreting them as thinking, which in turn requires interpreting them as using language, which, in turn, requires supposing the possibility

[6] There is a notional fifth possibility, which is that some special justification obtains but is trumped by an overarching principle. But I take it that this is equivalent to the fourth possibility – that no special justification obtains.

of understanding – and therefore translating – their language.[7] The conclusion of Davidson's argument is that all conceptual schemes are, in principle, intertranslatable: there is no such thing as a radically different 'alternative' conceptual scheme. Even if the argument from treating creatures as thinking beings to supposing the translatability of their conceptual scheme goes through, however, this would only establish that we are constrained to suppose that there are no alternative conceptual schemes.

However, it is far from clear that Davidson's argument succeeds even in this respect. There are at least two dubious premises in the argument: the first being that thought requires language, and the second being that interpreting a set of noises as language presupposes that this language is translatable into our own. Each is, I think, too strong. If thought requires language, then what are we to make of creatures who appear to deliberate about some problem before finding a (possibly creative) solution? Crows, for instance, have been observed forming pieces of wire into hooks in order to retrieve food from cylindrical containers.[8] This is not a piece of behaviour which occurs in the wild (there is, after all, a shortage of wire in their natural environment), and nor need it be taught. Prior to extracting the food, however, there is a period where the crow devotes its attention to the cylinder and to the wire. If crows, as non-linguistic animals, cannot think, then we are denied the resources to describe what is happening when the crow inspects the cylinder and wire, or to explain how the crow manages to generate a solution to the problem - unless, of course, we maintain that the process of generating a solution does not require thought. But that seems objectionably *ad hoc*.

Davidson does have reason for thinking that taking someone to be a language-user presupposes that we take their language to be (potentially) translatable into our own: to treat an utterance as linguistic is to treat it as meaningful, and to treat it as meaningful is to suppose that there is a meaning which can be attributed to it - and therefore that it is translatable into our own language. If we don't presuppose translatability, then we have no grounds for taking the noises which others generate to be part of a language. There is here the same difficulty as before: there is a difference between being compelled to assume that p, and having some proof that p. Furthermore, I think that Davidson is wrong to say that translatability is a necessary presupposition of thinking of some practice as a language. There are other reasons open to us for thinking of something as a language - for instance, the observation that a practice consists of systematic utterances, appears to be used for co-ordinating action or transmitting information, and so on.

[7] Davidson 1974.
[8] Weir, Chappell, & Kacelnik 2002.

If we suppose that there can exist languages which are entirely untranslatable into our own, or which lack cognate terms for central elements of our moral discourse (rights, obligations, etc.), then there is the possibility of agents who are rational, social creatures, much like ourselves, but incapable of sharing with us a conception of the moral life. This is problematic if we think that we would nonetheless have duties or obligations towards such creatures above and beyond those of refraining from harming (etc). It isn't clear that such a supposition would be warranted, since rationality requires being able to respond to (theoretical or practical) reasons. Since to think of a creature as rational is to think of it as responsive to universal considerations of a certain kind, there is an element to all rational mentality which is common, and hence shareable. But again, the worry with such cases is that we would (on the theory being offered) have no constructed obligations towards such creatures. Were constructed obligations to constitute the whole of morality, this would be deeply problematic: but it seems entirely plausible to suppose that, were we to encounter creatures who were alien to us in this respect (i.e. unable to play an active part in our moral community), we would have no obligations towards them beyond those which require us to avoid harming them unnecessarily.

Similarly with the case of the colonists and the natives: the realist sphere will also significantly restrict the extent to which the colonists may use force in displacing the colonists; absent some special justification, the disvalue of inflicting harm on others provides reason not to engage in the relevant course of action (importantly, the required justification is justification *to* the colonists: what is at issue is whether any reason can be given to the colonists to moderate their behaviour). But those considerations will be too weak: firstly, I take it that we are entitled to use force to prevent animals from interfering with our smallholdings (for instance), and the valence of pain and pleasure holds equally for animals as for humans; secondly, there are methods of coercion which do not involve causing pain, but which obviate agency nonetheless. As suggested earlier, therefore, we should assume that the obligations to refrain from displacing others derive from more general principles to respect individual autonomy and interests, and that these in turn are a function of the extent to which colonists and natives are capable of shared deliberation.

4. Scope and Purpose

Roger Scruton, in his *Animal Rights and Wrongs*, objects to thoroughgoing utilitarianism (in particular, to Singer's account, on which there are no intrinsic moral differences between humans and non-humans), partly on the grounds that part of the purpose of morality is to enable the kind of high-level social behaviour that human beings characteristically engage in, and also on the grounds that such an approach neglects salient (albeit contingent) differences between humans and animals.[9] Not only do humans characteristically have capacities which animals lack (language, rationality, etc.), but we clearly form moral communities in a way which animals do not: we use moral language to express conceptions of rights, obligations, and so on, and are willing to moderate our behaviour in accordance with these conceptions. Insofar as this is true, the hybrid theory has an advantage over both thoroughgoing utilitarianism and thoroughgoing constructivism.

There is, however, an instability in Scruton's position, which derives from his attempt to hold that part of morality depends on these complex capabilities, but also that infants and the severely mentally impaired are members of this peculiarly human moral community. Scruton claims that the key difference between infants and animals is that infants are *potential* persons – and hence that treating infants as such enables them to realise this potential. As far as non-moral adults are concerned, they are (he thinks) abnormal instances of a distinctively moral *kind* – and it is their status as (defective) members of the relevant kind that qualifies them as members of the moral community.

Neither of these claims is particularly plausible. The first claim is plausible only insofar as it can handle cases where the infant will never develop into a full member of the moral community (at least, with respect to its moral capacities). That, in turn, depends on the second claim: that abnormal instances of a certain kind may be treated according to the kind, rather than according to their intrinsic properties. But there are at least two problems with this claim: firstly, membership of kinds is indeterminate. That is, it is not clear whether we are to treat the severely mentally impaired as abnormal instances of the kind 'human being', or as normal instances of the kind 'severely mentally impaired', or as instances of the kind 'non-moral-agent'. Secondly, and more seriously, it is simply not true that we ought to treat abnormal instances of kinds according to the kind of which they are an instance. Poisoned meat is not to be treated as food on the grounds that it is merely an abnormal instance of the kind 'meat';

[9] Scruton 1996.

nor do chimpanzees qualify for membership of the moral community simply in virtue of being instances of the kind ‘primate’.

Any account, then, on which moral obligations derive from complex, specifically human, capabilities, is committed to treating a large set of marginal cases as falling outside of the moral community. The hybrid account is committed to treating these marginal cases as falling outside of the constructivist sphere of morality. To the extent that we think of human life – all human life – as being sacred, or as human beings as intrinsically valuable, or deserving of respect (in the Kantian sense), this account seems worrying. But there are few grounds for taking this thought at face value, especially where we can provide a debunking account of why things might seem this way. What would be problematic, I have suggested, is if our metaethical theory implied that we had *no* moral obligations towards the severely mentally impaired. The hybrid theory allows for the existence of moral obligations towards the severely mentally impaired; the cost of assuming this theory is that we are obliged to treat the value of human life as derivative from, rather than foundational to, the structure of morality. This cost is, I think, minor.

Nonetheless, Scruton is correct to highlight the relation between the point of a practice, on the one hand, and its content. Proposals for rules governing games, for instance, may be judged according to (amongst other criteria) the difference which they make to the pleasure derived from the game. Similarly, it is clear that our moral practice and discourse has an instrumental value: there are various goods, many of a peculiarly social nature, which are secured by morality. This is an observation which is secured by the constructivist part of the theory on offer.

5. Distant Others: Future Generations, Foreign Nations

Any moral theory which takes as basic the notion of a contract or agreement, hypothetical or otherwise, will have difficulty in accounting for obligations towards those who fall outside the scope of the contract. If morality is conceived of as the outcome of a contract between members of society in order to avoid conflict, then members of a given society have no obligations to those who are not part of that society. This is problematic, because it is commonly accepted that we have at least some obligations to other nations – if not to provide aid, then, minimally, not to harm unnecessarily. Furthermore, it is not obvious how membership of a society is to be determined. Future generations pose even greater difficulties for such moral theories: future generations are not part of any current society, nor can any current society be extended to incorporate such generations. Worse still, the very existence of future generations will depend on the action of present individuals; this makes it difficult to assess our actions in terms of the harm or benefit caused to future individuals, since, had we acted otherwise, those future individuals would not have come into existence. Parfit calls this the 'non-identity problem'.[10] Not only do future persons fall outwith the scope of any potential contract, but, at least from our point of view, there are no identifiable future persons towards whom we might bear moral obligations.

These reasons, among others, count strongly against any such moral theory.[11] But these considerations count only against a theory which is *purely* contractarian in form. I now wish to outline, briefly, how the hybrid view might bear on these issues.

As regards the international question, there is scope – within constructivist morality – for making sense of international moral obligations. Just as individual persons within society are bound by moral obligations given their social nature, so this relation is mirrored at the international level, where shared interest gives reason for two or more parties to mutually constrain their behaviours. The more interesting question, however, concerns our individual obligations to members of other societies, particularly where those other societies involve widespread (absolute) poverty, and where our capacity to provide aid is substantial. There is disagreement within the existing philosophical literature as to whether we have stringent obligations to the needy others in such communities, merely moderate obligations, or no obligations whatsoever.[12] Even once the existence of obligations to such needy others is accepted, there remains a further question as to the nature of these obligations – whether they are obligations of justice, or of common humanity, and so on.

On the hybrid theory, we do not have constructivist obligations to needy members of other societies: the absence of interaction, and the vast asymmetry in capacity for action, entails that there is no co-ordination problem to be solved here.[13] Nonetheless, there remain reasons for action provided by the realist part of morality: the suffering of distant others provides reason for us to alleviate this suffering, dependent on our capacity to provide aid, and regardless of our particular relation to those distant others. There will, however, be *further* reasons (constructivist reasons) to provide aid to members of our own society. On this

[10] Parfit 1984: 350.

[11] See Mulgan 2006: 24 – 54 for a detailed treatment of these issues.

[12] For 'stringent' theories, see Unger 1996, Singer 1993; for a minimal theory, see Nozick 1974.

[13] There is one possible exception here, which is that a society of benevolent individuals may construct principles mandating aid to needy but distant others. For simplicity, I will avoid discussion of this exception.

analysis, our reasons for providing aid to members of other societies are a function of compassion, rather than of justice.

As regards the question of future persons, the analysis here follows the same lines. Since future generations are not involved with the current co-ordination problem, and since there are no determinate future persons to whom we might have obligations, the constructivist part of morality does not apply. Nonetheless, the realist part of morality, which maintains that reasons derive from the intrinsic value of hedonic or anhedonic states, indicates that we have reason to act in such a way as to promote this value in future, as well as present, generations. We have, on this account, realist reasons to make happy people, as well as to make people happy, although the latter derives in part from the former. The fact that our actions will affect which persons exist in future is irrelevant, here, and hence the non-identity problem ceases to be a problem – at least, not given the nature of realist morality as outlined in this thesis. I discuss an objection to this account of reasons in the next section.

6. The Repugnant Conclusion

The hybrid theory as outlined in this thesis also has a distinct advantage over any straightforward utilitarian theory, as it enables us to provide a solution to the 'repugnant conclusion' objection generated by Derek Parfit. This is a widely-discussed challenge to any form of aggregate maximising (or 'total') utilitarianism, and it runs as follows. Suppose that aggregate maximising utilitarianism is correct. This commits us to what Parfit calls the 'impersonal total principle':

'*Impersonal Total Principle*: If other things are equal, the best outcome is the one in which there would be the greatest quantity of whatever makes life worth living.'[14]

But since the Impersonal Total Principle ignores the way in which personal value is distributed, then for any distribution A we can create a comparatively better distribution B which contains a much greater population of individuals, a lower mean level of value, and a greater net level of value.[15] In terms of hedonistic utilitarianism, for any group of very happy individuals, it would be better to create a larger group of slightly less happy individuals, as

[14] Parfit 1984: 387.

[15] Of course, we may have reasons to refrain from bringing the repugnant situation into existence, because that would impinge on our welfare – and principles which would seek to bring about the repugnant situation are consequently unacceptable. However, these reasons only exist insofar as the repugnant situation impinges on us. There still seems to be something unpalatable about the suggestion that the existence of a vast number of people with mediocre lives is better than the existence of a slightly less vast number of people with more-than-mediocre lives.

long as the total amount of happiness is increased. But this can be iterated until the individuals in question have lives which are barely worth living – this, Parfit calls the 'repugnant conclusion':

RC: 'For any possible population of at least ten billion people, all with a very high quality of life, there must be some much larger imaginable population whose existence, if other things are equal, would be better, even though its members have lives that are barely worth living.'[16]

Certainly, it seems that aggregate maximising utilitarianism is committed to the repugnant conclusion. Utilitarians who are unwilling to admit the repugnant conclusion might choose some function other than aggregate maximisation; one might, for instance, think that we ought to do is to maximise the average level of utility. Equally, we might argue that there is some critical level of welfare at which it becomes valuable to bring that person into existence. Broome, for instance, distinguishes between the level of well-being at which it is neither good nor bad to add an individual with that level of well-being, and the level of well-being which is neutral *for that person* – that is to say, 'barely worth living'.[17] The individuals in the 'repugnant' situation, according to Broome, may have a fairly decent quality of life, and hence the conclusion is not as repugnant as it seems. But I will avoid a full discussion of these issues. I think, however, that we should be cautious about trusting our intuitions in the case of the repugnant conclusion.

As Broome points out, our intuitions regarding very large numbers are to be treated with some suspicion. Broome discusses the suggestion that dying from AIDS is so bad that saving one person from AIDS is better than saving any number of people from a very mild headache, and observes that there are many occasions where imposing some mildly frustrating rule would save lives, but where we do not think it best to impose such a rule.[18] For instance, we might save many lives by reducing the speed limit by five miles per hour across the board; this would lengthen journey times, but would also (presumably) save lives. Nonetheless, it is far from clear that reducing the speed limit is justified; we might think that a fractional increase in the risk of death is outweighed by the potential gains. This thought is open to the aggregate maximising utilitarian, although we can also make sense of it along constructivist lines: as the example indicates, there are cases where we are willing to accept additional risk in order to marginally increase individual welfare. But that suggests that, were we to find

[16] Parfit 1984: 388.
[17] Broome 2004: 199-212, 233-234.
[18] Broome 2004: 54-57.

ourselves in the repugnant situation, we would have reason to attempt to engage in population reduction, if that would marginally increase individual welfare.

It is also not clear why, exactly, we find the repugnant conclusion unpalatable. One (to my mind very plausible) suggestion is that the repugnant situation omits many factors which we take to be particularly valuable. That is, the repugnant situation replaces a certain number of Mozart-like elements with a large volume of 'muzak and potatoes'.[19] But although it is certainly true that Mozart-like elements are more valuable than muzak and potatoes, I take it that they are both valuable for the same reason – and, ipso facto, commensurable. Consequently, sufficient quantities of mild pleasures can outweigh smaller quantities of more complex pleasures. Now it would be entirely appropriate to bemoan the absence of Mozart-like elements in the repugnant situation: Mozart-like elements, after all, provide a great deal of pleasure for a great number of people. But it is (as Broome says) much more difficult to adequately take into account the mild happiness of the many.

There is another confounding factor, which is that when we imagine the repugnant situation it is very difficult not to imagine ourselves as part of some Orwellian scenario, where not only are we prevented from pursuing certain goods which we find particularly valuable, but our lives are generally dull and joyless. The conclusion seems to us repugnant because it is clear that we would prefer to live in a small but generally happy society, rather than a large, Orwellian one. But what is at issue is not individual preference, but rather questions of value. The question of what one would prefer is distinct from the question of what there is reason to bring about, although the two are easy to conflate.

So there is some explanatory work to be done as regards our intuitions concerning the repugnant conclusion. If we maintain that the repugnant conclusion is indeed repugnant, we need to give some reason for *why* it is so. It does seem that the conclusion is only genuinely repugnant if it involves the denial of something which is peculiarly valuable – meaningful friendships, experiences of Mozart, etc. – but the hedonistic utilitarian is committed, I have suggested, to the claim that such experiences are quantitatively, rather than qualitatively, superior to baser pleasures.[20] Consequently, the value of such experiences can be 'traded off' against the value of other experiences. Hence, it is hard to see how the utilitarian can treat the repugnant conclusion as genuinely repugnant.

[19] Ryberg 1996, esp. 208.

[20] There is scope for further discussion here – however, I take it that an axiology which introduces an irreducible distinction between qualities of pleasure thereby introduces values other than pleasure. But see e.g. Riley 200, Fletcher 2008.

Here the hybrid theory is at a distinct advantage. According to the hybrid theory, both hedonic states *and* constructivist principles generate reasons for action. So although the hybrid theory agrees with the hedonistic utilitarian in taking certain mental states to be the sole candidates for intrinsic value, there may nonetheless be reasons to hold that the repugnant conclusion is, indeed, repugnant. One way in which such reasons could be generated is as follows. Of all the proposed ways in which to structure a society, one which might turn out to be unacceptable to most people is one which offers a life which, being filled with 'muzak and potatoes', they find deeply undesirable. The repugnant character of the situation outlined by Parfit will therefore turn out to be sufficient to generate reasons to refrain from promoting such a situation. This will hold in any situation where increasing the number of happy persons will leave the existing group, on average, objectionably worse off. The construction procedure allows individual preferences to come to bear on the question of what we have reason to do, without denying that the only bearers of intrinsic value are hedonic states. This is, I think, a theoretical merit: it explains why the repugnant conclusion seems repugnant, thus accounting for our intuitions, whilst avoiding a commitment to the conclusion itself. And note that, although I take it that the construction procedure deals primarily with ethical principles, there is no reason why these cannot justify talk of peculiar values which are not instantiated within the repugnant conclusion – provided that these values are understood as derivative, rather than intrinsic.

7. Conclusion

I have attempted, in this chapter, to provide a brief illustration of how the hybrid theory is at a distinct advantage over two independently plausible theories – hedonistic utilitarianism (which attaches to a realist metaethics), on the one hand, and contractarianism (which attaches to a constructivist metaethics), on the other. Problems that are distinctive for hedonistic utilitarianism can be solved by restricting the scope of that theory, and by combining it with contractarianism. The realist part of morality accounts for our generic reasons to refrain from harming sentient others, whilst constructivism yields additional reasons to refrain from harming persons, as well as addressing the more complex machinery of justice, rights, and so on. Where constructivist reasons run out – in the case of distant or future persons, or for sentient non-rational creatures – we still have reasons deriving from the value of hedonic states. We therefore have reasons to treat animals well, regardless of whether or not they are moral agents, and reasons to treat the environment in such a way that the lives of future generations are not impoverished. Lastly, the combination of realism and constructivism serves to block the 'repugnant conclusion' which straightforward aggregative utilitarianism seems driven towards. The benefits of adopting a hybrid theory at the applied level, that is, mirror the benefits of adopting a hybrid theory at the metaethical level. There is, therefore, strong reason to adopt a position of the form outlined in this thesis.

References

Altham, J. E. J. 1986. *The Legacy of Emotivism*, in G. W. Macdonald, C. Wright (eds.) *Fact, Science and Morality: Essays on A. J. Ayer's Language, Truth and Logic*. Oxford: Blackwell.

Altman, A. 2004. “Breathing Life into a Dead Argument: G.E. Moore and the Open Question”. *Philosophical Studies* 117(3): 395-408.

Aristotle. 1925. *The Nichomachean Ethics.* Trans: Ross, W. D. Oxford: Oxford University Press.

Arkway, A. 2000. "The Simulation Theory, the Theory Theory, and Folk Psychological Explanation". *Philosophical Studies* 98: 115-137.

Audi, R. 2004a. *The Good in the Right: A Theory of Intuition and Intrinsic Value*. Oxford: Princeton University Press.

Audi, R. 2004b. "Reasons, Practical Reason, and Practical Reasoning". *Ratio* 17(2): 119-149.

Austin, J. L. 1962. *Sense and Sensibilia*. Oxford: Clarendon Press.

Ayer, A. J. 1946. *Language, Truth and Logic*. London: Gollancz.

Baehr, J. 2003. "Korsgaard on the Foundations of Moral Obligation". *Journal of Value Inquiry* 37(4): 481-491.

Baggini, J. 2002. “Morality as a Rational Requirement”. *Philosophy* 77(3): 447-453.

Batson, C., Fultz, J., & Shoenrade., P. A. 1987. “Adults' emotional reactions to the distress of others”. In Eisenberg, N. & Strayer, J. (eds.) *Empathy and its development.* Cambridge: Cambridge University Press.

Bavelas, J. B., Black, A., Lemery, C. R., & Mullett, J. 1987. “Motor mimicry as primitive empathy”. In Eisenberg, N. & Strayer, J. (eds.) *Empathy and its development*. Cambridge: Cambridge University Press.

Bernstein, M. 1997. "Contractualism and Animals". *Philosophical Studies* 86(1): 49-72.

Binmore, K. 1994. *Playing Fair: Game Theory and the Social Contract I*. Cambridge, MA: MIT Press.

Binmore, K. 2005. *Natural Justice*. Oxford: Oxford University Press.

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

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

Blackburn, S. 1998. *Ruling Passions: A Theory of Practical Reasoning*. New York: Clarendon Press.

Blackburn, S. 2002. "Précis of *Ruling Passions*". *Philosophy and Phenomenological Research* 65(1): 122-135.

Bloomfield, P. 2001. *Moral Reality*. New York: Oxford University Press.

Bloomfield, P. 2006. “Opening Questions, Following Rules”. In Horgan, T. & Timmons, M. (eds.) *Metaethics after Moore*. Oxford: Oxford University Press**.**

Boyd, R. 1988. “How to be a Moral Realist”. In Sayre-McCord, G. (ed.) *Essays on Moral Realism.* London: Cornell University Press**.**

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

Broome, J. 1999. “Normative requirements”. *Ratio* 12: 398-419.

Broome, J. 2002. “Practical Reasoning”. in Bermudez, J. & Millar, A. (eds.) *Reason and Nature: Essays in the Theory of Rationality*. Oxford: Oxford University Press.

Broome, J. 2004. *Weighing Lives*. Oxford: Oxford University Press.

Broome, J. 2008. "Reply to Southwood, Kearns and Star, and Cullity". *Ethics* 119: 96-108.

Brown, R. & Ladyman, J. 2009. "Physicalism, Supervenience, and the Fundamental Level". *The Philosophical Quarterly* 59(234): 20-38.

Campbell, R., & Woodrow, J. 2003. "Why Moore's Open Question is Open: the Evolution of Moral Supervenience". *The Journal of Value Inquiry* 37: 353-372.

Carlson, G. R. 1990. "Pain and the Quantum Leap to Agent-Neutral Value". *Ethics* 100(2): 363-367.

Canfield, J. V. 1996. "The Community View". *The Philosophical Review* 105(4): 469-488.

Chagnon, N. A. 2000. "Yanomamö: the Last Days of Eden". In Gowans 2000.

Cohen, G. A. 1996. "Reason, Humanity, and the Moral Law". In Korsgaard 1996b.

Copp, D. 1995. *Morality, Normativity, and Society*. New York: Oxford University Press.

Cullity, G. & Gaut, B. (eds.) 1997. *Ethics and Practical Reason*. Oxford: Oxford University Press.

Dancy, J. 2004. *Ethics without Principles*. Oxford: Clarendon Press.

Darwall, S. 1992. "Internalism and Agency". *Philosophical Perspectives* 6: 155-174.

Darwall, S. 1998. "Empathy, Sympathy, Care". *Philosophical Studies* 89(2): 261-282.

Darwall, S., Gibbard, A., & Railton, P, (eds.) 1997a. *Moral Discourse and Practice: Some Philosophical Approaches*, New York: Oxford University Press.

Darwall, S., Gibbard, A., & Railton, P. (eds.) 1997b. "Towards *Fin de Siècle* Ethics: Some Trends". In Darwall, Gibbard & Railton 1997a. Originally published in *The Philosophical Review* 101(1): 115-189, January 1992.

Davidson, D. 1973. "Radical Interpretation". *Dialectica* 27: 313-328.

Davidson, D. 1974. "On the Very Idea of a Conceptual Scheme". *Proceedings and Addresses of the American Philosophical Association* 47: 5-20.

Davidson, D. 1984, *Inquiries into Truth and Interpretation*, Oxford: Oxford University Press.

Dawkins, R. 1995. *River out of Eden*. London: Phoenix.

de Brigard, F. 2008. "If you like it, does it matter if it's real?". University of North Carolina: Chapel Hill. Online at http://homepage.uab.edu/angner/SWB/DeBrigard.pdf.

Dennett, D. C. 1978. *Brainstorms*. Hassocks, Sussex: The Harvester Press Limited.

Dennett, D. C. 1991. *Consciousness Explained*. London: Penguin.

Doyle, J. 2000. "Moral Rationalism and Moral Commitment". *Philosophy and Phenomenological Research* 60(1): 1-22.

Dreier, J. 2000. "Dispositions and Fetishes: Externalist Models of Moral Motivation". *Philosophy and Phenomenological Research* 61(3): 619-638.

Eisenberg, N. & Strayer, J. (eds.) 1987. *Empathy and its development,* Cambridge: Cambridge University Press.

Eklund, M. 2007. "Fictionalism". *The Stanford Encyclopedia of Philosophy*, Zalta, E. N. (ed.), online at http://plato.stanford.edu/entries/fictionalism/.

Elithorn, A., Glithero, E. & Slater, E. 1958. "Leucotomy for Pain". *Journal of Neurology, Neurosurgery & Psychiatry* 21(4): 249-261.

Enoch, D. 2005. "Why Idealize?". *Ethics* 115(4): 759-787.

Fantl, J. 2006. "Is Metaethics Morally Neutral?". *Pacific Philosophical Quarterly* 87: 24-44.

Feinberg, J. 1970. *Doing and Deserving: Essays in the Theory of Responsibility.* Princeton: Princeton University Press.

Feldman, F. 1997. *Utilitarianism, Hedonism, and Desert: Essays in Moral Philosophy*. Cambridge: Cambridge University Press.

Feldman, F. 2000. "Basic Intrinsic Value". *Philosophical Studies* 99(3): 319-346.

FitzPatrick, W. J. 2004. "Reasons, Value, and Particular Agents: Normative Relevance without Motivational Internalism". *Mind* 113(450): 285-318.

FitzPatrick, W. J. 2008. "Robust Ethical Realism, Non-Naturalism and Normativity". In Shafer-Landau, R., (ed.) *Oxford Studies in Metaethics: Volume III*. Oxford: Oxford University Press.

Fletcher, G. 2008. "The Consistency of Qualitative Hedonism and the Value of (at Least Some) Malicious Pleasures". *Utilitas* 20(4): 462-471.

Foot, P. 2002. *Virtues and Vices and Other Essays in Moral Philosophy*. Oxford: Clarendon Press.

Fox, I. 1989. "On the Nature and Cognitive Function of Phenomenal Content - Part One". *Philosophical Topics* 17: 81-117.

Gauthier, D., Ed. 1986. *Morals by Agreement*. Oxford: Clarendon Press.

Geach, P. T. 1960. "Ascriptivism". *The Philosophical Review* 69(2): 221-225.

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

Gibbard, A. 2003. *Thinking How to Live*. London: Harvard University Press.

Goldman, A. H. 2007. "The Case Against Objective Values". *Ethical Theory and Moral Practice* 11(5): 507-524.

Goldman, A. I. 1992. "Empathy, Mind, and Morals". *Proceedings and Addresses of the American Philosophical Association* 66(3): 17-41.

Goldstein, I. 1989. "Pleasure and Pain: Unconditional, Intrinsic Values". *Philosophical and Phenomenological Research* 1(2): 255-276.

Goodwin, G. P. & Darley, J. M. 2008. "The psychology of meta-ethics: Exploring objectivism". *Cognition* 106: 1339-1366.

Gowans, C. W. (ed.) 2000. *Moral Disagreements: Classic and Contemporary Readings*. London: Routledge.

Grahek, N. 2007. *Feeling Pain and Being in Pain*. Cambridge, MA: MIT Press.

Haldane, J. & Wright, C. (eds.) 1993. *Reality, Representation and Projection*. Oxford: Oxford University Press.

Hare, R. M. 1981. *Moral Thinking: Its Levels, Method and Point*. Oxford: Clarendon Press.

Hare, R. M. 1991. *The Language of Morals*. New York: Oxford University Press.

Hare, R. M. 1997. *Sorting Out Ethics*. New York: Clarendon Press.

Harman, G. 1997. "Ethics and Observation". In Darwall, Gibbard & Railton 1997.

Hoffman, M. L. 1987. “The contribution of empathy to justice and moral judgement”. In Eisenberg and Strayer 1987.

Hooker, B. 2000. *Ideal Code, Real World: A Rule-Consequentialist Theory of Morality*. Oxford: Oxford University Press.

Hooker, B. & Stratton-Lake, P. 2006. “Scanlon versus Moore on Goodness”. In Horgan & Timmons 2006.

Hooker, B. & Little, M. O. 2000. *Moral Particularism*. Oxford: Clarendon Press.

Hooker, B. & Streumer, B. 2004. “Procedural and Substantive Rationality”. In Mele, A. and Rawling, P. (eds.) *The Oxford Handbook of Rationality*. Oxford: Oxford University Press**.**

Horgan, T. & Timmons, M. 1991. “New Wave Moral Realism Meets Moral Twin Earth”, *Journal of Philosophical Research* 16: 447-472.

Horgan, T. & Timmons, M. 1992a. “Troubles on Moral Twin Earth: the ‘Open Question Argument’ Revived”. *Philosophical Papers* **21**: 153-175.

Horgan, T. & Timmons, M. 1992b. “Troubles on Moral Twin Earth: Moral Queerness Revived”. *Synthese* 92(2): 221-260.

Horgan, T. & Timmons, M. 1996. “From Moral Realism to Moral Relativism in One Easy Step”. *Critica* 28: 3-39.

Horgan, T. & Timmons, M. 2000. “Copping out on Moral Twin Earth”. *Synthese* 124(1-2): 139-152.

Horgan, T. & Timmons, M. (eds.) 2006. *Metaethics after Moore*. Oxford: Oxford University Press.

Huemer, M. 2005. *Ethical Intuitionism*. Basingstoke: Palgrave Macmillan.

Hume, D. 1975. *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*. Selby-Bigge, A. (ed.). Oxford, Oxford University Press.

Hume, D. 2000. *A Treatise of Human Nature*. Norton, D. F. & Norton, J. M. (eds.). Oxford: Oxford University Press.

Hurley, S. L. 1989. *Natural Reasons*. Oxford, Oxford University Press.

Irwin, G. 1989. "Pleasure and Pain: Unconditional, Intrinsic Values". *Philosophical and Phenomenological Research* 1(2): 255-276.

Jackson, F. 2000. *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford: Oxford University Press.

James, S. 1998. "The Virtue of Justice: Onora O'Neill's Towards Justice and Virtue: A Constructive Account of Practical Reasoning". *International Journal of Philosophical Studies* 6(2): 253-263.

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

Joyce, R. 2002. "Expressivism and Motivation Internalism". *Analysis* 62(4): 336-344.

Kalderon, M. E. 2004. “Open Questions and the Manifest Image”. *Philosophy and Phenomenological Research* 65(2): 251-289.

Kalderon, M. E. 2005. *Moral Fictionalism*. Oxford: Clarendon Press.

Kawall, J. 2005. “Moral Realism and Arbitrariness”. *The Southern Journal of Philosophy* 43: 109-129.

Kearns, S. and Star, D. 2009. “Reasons as Evidence”. Forthcoming in *Oxford Studies in Metaethics*, Volume IV, Shafer-Landau, R. (ed.) 2009. Oxford: Oxford University Press. Online at http://bu.academia.edu/documents/0020/1354/ReasonsEvidence.pdf, accessed 23rd August 2009.

Kennett, J. 2002. "Autism, Empathy and Moral Agency". *The Philosophical Quarterly* 52(208): 340-357.

Korsgaard, C. M. 1996a. *Creating the Kingdom of Ends*. Cambridge, Cambridge University Press.

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

Kraut, R. 2007. *What is Good and Why: the Ethics of Well-being*. Cambridge, MA: Harvard University Press.

Lance, M. & Little, M. 2005. “Defending Moral Particularism”, in Dreier, J. (ed.) *Contemporary Debates in Moral Theory*. Blackwell: London.

Lenman, J. 1999. "The Externalist and the Amoralist". *Philosophia* 27(3-4): 441-457.

Lenman, J. 2003. "Disciplined Syntacticism and Moral Expressivism". *Philosophy and Phenomenological Research* 66(1): 32-57.

Lewis, D. 1988. "Desire as Belief". *Mind* 97(387): 323-332.

Lewis, D. 2001. "Redefining 'Intrinsic'". *Philosophy and Phenomenological Research* 63(2): 381-398.

Lillehammer, H. 2000. "The Doctrine of Internal Reasons". *Journal of Value Inquiry* 34(4): 507-516.

Lillehammer, H. 2003. "Smith on Moral Fetishism". *Analysis* 57(3): 187-195.

Loeb, D. 2007. "The Argument from Moral Experience". *Ethical Theory and Moral Practice* 10(5): 469-484.

Mackie, J. L. 1990. *Ethics: Inventing Right and Wrong*. London: Penguin Books.

McDowell, J. 1979. "Virtue and Reason". *The Monist* 62: 331-350.

McDowell, J. 1995. “Might there be external reasons?”. In Altham, J. E. J. & Harrison, R. (eds.) *World, Mind, and Ethics: Essays on the Ethical Philosophy of Bernard Williams*. Cambridge: Cambridge University Press**.**

McDowell, J. 1998. *Meaning, Knowledge and Reality*. London: Harvard University Press

McNaughton, D. 1988. *Moral Vision: An Introduction to Ethics*. Oxford: Blackwell.

Moore, G. E. 1922. *Principia Ethica*. London: Cambridge University Press.

Mulgan, T. 2001. *The Demands of Consequentialism*. Oxford: Clarendon Press.

Mulgan, T. 2006. *Future People - A Moderate Consequentialist Account of our Obligations to Future Generations.* Oxford: Oxford University Press.

Mulgan, T. 2007. *Understanding Utilitarianism*. Stocksfield: Acumen.

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

Nagel, T. 1986. *The View from Nowhere*. New York: Oxford University Press.

Neill, A. 1996. “Empathy and (Film) Fiction”. In Bordwell, D., and Carroll, N. (eds.) *Post-Theory: Reconstructing Film Studies*. Madison, Wisconsin: The University of Wisconsin Press.

Nichols, S. & Stich, S. 1993. "Folk Psychology: Simulation or Tacit Theory". *Philosophical Issues* 3: 225-270.

Nietzsche, F. 1973. *Beyond Good and Evil*. London: Penguin.

Nietzsche, F. 1990. *Twilight of the idols; and, The Anti-Christ*. Harmondsworth: Penguin Books.

Nozick, R. 1974. *Anarchy, State, and Utopia*. Oxford: Blackwell.

O'Connor, T. 1994. "Emergent Properties". *American Philosophical Quarterly* 31(2): 91-104.

O'Hara, R. J. 1993. "Systematic Generalization, Historical Fate, and the Species Problem." *Systematic Biology* 42(3): 321-246.

Olson, J. 2002. "Are Desires De Dicto Fetishistic?". *Inquiry* 45(1): 89-96.

O'Neill, O. 1996. *Towards Justice and Virtue: a Constructive Account of Practical Reasoning*. Cambridge: Cambridge University Press.

O'Neill, O. 2003. "Constructivism vs. Contractualism". *Ratio* 16(4): 319-331.

Papineau, D. 1993. *Philosophical Naturalism*. Oxford: Blackwell.

Parfit, D. 1984. *Reasons and Persons*. Oxford: Oxford University Press.

Rachels, S. 2002. "Nagelian Arguments Against Egoism". *Australasian Journal of Philosophy* 80(2): 191-208.

Railton, P. 1984. "Alienation, Consequentialism, and the Demands of Morality". *Philosophy and Public Affairs* 13(2): 134-171.

Railton, P. 1997. “On the Hypothetical and Non-Hypothetical”. In Cullity, G. & Gaut, B. (eds.) *Ethics and Practical Reason*. Oxford: Clarendon Press**.**

Rakowski, E. 2001. “Taking and Saving Lives”. In Harris, J. (ed.) *Bioethics*. Oxford: Oxford University Press.

Rawls, J. 2005. *Political Liberalism*. New York: Columbia University Press.

Richardson, H. 1999. “Towards Justice and Virtue by Onora O’Neill”. Review of O’Neill 1996. *Mind* 108: 598-601.

Riley, J. 2003. “Interpreting Mill’s Qualitative Hedonism”. *The Philosophical Quarterly* 53(212): 410-418.

Rosati, C. S. 1995. “Naturalism, Normativity and the Open Question Argument”. *Noûs* 29(1): 46-70.

Roskies, A. 2003. "Are ethical judgements intrinsically motivational? Lessons from ‘acquired sociopathy’". *Philosophical Psychology* 16(1): 51-66.

Ross, D. 1930. *The Right and The Good*. Oxford: Clarendon Press.

Ryberg, J. 1996. “Parfit’s Repugnant Conclusion”. *Philosophical Quarterly* 46(183): 202-213.

Sayre-McCord, G. 1988. *Essays on Moral Realism*. Ithaca: Cornell University Press.

Scanlon, T. 1982. “Contractualism and Utilitarianism”. In Sen & Williams 1982.

Scanlon, T. 1998. *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

Schroeder, M. 2008. "How Expressivists Can and Should Solve Their Problem with Negation". *Noûs* 42(4): 573-599.

Schueler, G. F. 1998. “Modus Ponens and Moral Realism”. *Ethics* 98(3): 492-500.

Schweder, R. A. 2000. “The astonishment of anthropology”. In Gowans 2000.

Scruton, R. 1996. *Animal Rights and Wrongs.* London: Metro Books.

Searle, J. R. 2001. *Rationality in Action.* London: MIT Press.

Sen, A. 1988. *On Ethics and Economics*. Oxford: Blackwell.

Sen, A. & Williams, B. (eds.) 1982. *Utilitarianism and Beyond*. Cambridge: Cambridge University Press.

Shafer-Landau, R. 2003. *Moral Realism: a Defence*. Oxford: Clarendon Press.

Shafer-Landau, R. (ed.) 2006. *Oxford Studies in Metaethics Volume 1.* Oxford: Oxford University Press.

Shafer-Landau, R. (ed.) 2007. *Oxford Studies in Metaethics Volume 2.* Oxford: Oxford University Press.

Shafer-Landau, R. (ed.) 2008. *Oxford Studies in Metaethics Volume 3*. Oxford: Oxford University Press.

Shaver, R. 2004. "The Appeal of Utilitarianism". *Utilitas* 16(3): 235-250.

Sidgwick, H. 1930. *The Methods of Ethics*. London: Macmillan.

Simpson, E. 1999. "Between Internalism and Externalism in Ethics". *The Philosophical Quarterly* 49(195): 201-

214.

Sinclair, N. 2007. "Expressivism and the Practicality of Moral Convictions". *The Journal of Value Inquiry* 41: 201-220.

Sinclair, N. 2009. “The Pretensions of Moral Realism”. Unpublished. Via personal correspondence, 10th June 2009.

Singer, P. 1993. *Practical Ethics*. Cambridge, Cambridge University Press.

Sinnott-Armstrong, W. 2006. *Moral Scepticisms*. Oxford, Oxford University Press.

Smith, M. 1987. "The Humean Theory of Motivation". *Mind* 96(381): 36-61.

Smith, M. 1994. *The Moral Problem*. Oxford: Blackwell.

Snow, N. E. 2000. "Empathy". *American Philosophical Quarterly* 37(1): 65-78.

Strandberg, C. 2004. “In Defence of the Open Question Argument”. *The Journal of Ethics* 8: 179-196.

Stratton-Lake, P. 1998. “Internalism and the Explanation of Belief/Motivation Changes”. *Analysis* 58(4): 311-315.

Stevenson, C. L. 1945. *Ethics and Language*. London: Oxford University Press.

Street, S. 2006. “Constructivism about Reasons”, paper presented at Metaethics Workshop, University of Madison-Wisconsin. Online at http://philosophy.wisc.edu/info/2006%20Metaethics%20Workshop/street.doc, accessed 23rd August 2009.

Sturgeon, N. L. 1988. “Moral Explanations”. In Sayre-McCord, G. (ed.) *Essays on Moral Realism*, 229-256. London: Cornell University Press.

Svavarsdottir, S. 1999. "Moral Cognitivism and Motivation". *Philosophical Review* 108(2): 161-219

Tannsjo, T. 1998. *Hedonistic Utilitarianism*. Edinburgh: Edinburgh University Press.

Timmons, M. 1999. *Morality Without Foundations: a Defence of Ethical Contextualism*. Oxford: Oxford University Press.

Timmons, M. 2003. "The Limits of Moral Constructivism". *Ratio* 16(4): 391-423.

Trogdon, K. 2009. "Monism and Intrinsicality". *Australasian Journal of Philosophy* 87(1): 127-148.

Unger, P. 1996. *Living High and Letting Die: Our Illusion of Innocence*. Oxford: Oxford University Press.

Unwin, N. 1999. "Quasi-Realism, Negation and the Frege-Geach Problem". *The Philosophical Quarterly* 49(196): 337-352.

Van Roojen, M. 2000. “Motivational Internalism: a Somewhat Less Idealized Account”. *The Philosophical Quarterly*, 50: 233-41.

Van Roojen, M. 2002. “Humean and Anti-Humean Internalism about Moral Judgements”. *Philosophy and Phenomenological Research* 65(1): 26-49.

Van Roojen, M. 2006. “Knowing Enough to Disagree”. In Shafer-Landau 2006.

Van Willigenburg, T. 2002. "Shareability and Actual Sharing: Korsgaard's Position on the Publicity of Reasons". *Philosophical Investigations* 25(2): 172-189.

Watkins, E. & FitzPatrick, W. 2002. "O'Neill and Korsgaard on the Construction of Normativity". *Journal of Value Inquiry* 36(2-3): 349-367.

Wedgwood, R. “The Moral Evil Demons”. Forthcoming in Feldman, R. & Warfield, T. *Disagreement*. Oxford: Oxford University Press.

Weir, A. S., Chappell, J., & Kacelnik, A. 2002. "Shaping of Hooks in New Caledonian Crows". *Science* 297(5583): 981.

Westermarck, E. 1906. *The Origin and Development of the Moral Ideas, Volume I*. London: Macmillan.

Westermarck, E. 1908. *The Origin and Development of the Moral Ideas, Volume II*. London: Macmillan.

Wiland, E. 2000. “A Fallacy in Korsgaard’s Argument for Moral Obligation”. *The Journal of Value Inquiry* 34(1): 103-104.

Williams, B. 1966. "Consistency and Realism". *Proceedings of the Aristotelian Society Supplementary Volume* 40: 1-22.

Williams, B. 1981. *Moral Luck: Philosophical Papers 1973-1980*. Cambridge: Cambridge University Press.

Wright, C. 1992. *Truth and Objectivity*. London: Harvard University Press.