Prescriptions and Universalizability:

A defence of Harean ethical theory

Daniel Y. Elstein

Gonville and Caius College, University of Cambridge, July 2013
This dissertation is submitted for the degree of Doctor of Philosophy.

To my grandma, Rose.

I declare that this dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

Summary

Prescriptions and Universalizability by Daniel Y. Elstein

R.M. Hare had an ambitious scheme of providing a unified account of meta-ethics and normative ethics by combining expressivism with Kantianism and utilitarianism. The project of this thesis is to defend Hare's theory in its most ambitious form. This means not just showing how the expressivist, Kantian and utilitarian elements are consistent, or that the three are each correct, but also that they are interdependent. The only defensible form of expressivism is Kantian; the only defensible Kantian theory is both expressivist and utilitarian; the only defensible utilitarianism is Kantian.

The thesis is divided into four chapters. Chapter 1 aims to show how expressivism can provide a coherent account of moral judgement and discourse. The argument for expressivism draws on Hare's thought that the main error of moral realism is to think of moral objectivity as requiring objects, moral properties which are really there in the world. It is shown, using an argument based on the Euthyphro and the Open Question Argument that realism is untenable because it makes this mistake, and this clears the path to expressivism.

Chapter 2 is a full account of the issues surrounding the Frege-Geach problem (often pressed against Hare), showing how it can be solved and how exactly the expressivist's embrace of minimalism about truth interacts with the solution to the Frege-Geach Problem. I include an explanation of how the expressivist is able to solve the most threatening version of the problem: Schroeder's discussion of negation.

Chapter 3 argues for the connection between expressivism and Kantianism. The argument (roughly following Korsgaard) is that Humean versions of expressivism run into a sceptical challenge of normative regress. Kant employed a transcendental argument to resolve this regress, deriving his Formula of Universal Law from the Categorical Imperative. This argument defended with expressivism playing a crucial role. This chapter thus explains how Hare is entitled to universalizability in a way that avoids the shmoralising objection: it is not justified merely by being derived from our moral concepts but rather from our inescapable nature as agents.

Chapter 4 illuminates the other connection, between Kantianism and utilitarianism. The largest part of the chapter is spent defending Hare's argument from universalizability to utilitarianism. Doing so shows how Hare's utilitarianism depends on his Kantianism, and so also how it indirectly depends on his utilitarianism. I then go on to defend Hare's distinctive two-level version of utilitarianism, especially against the objections of Bernard Williams. It is also argued that various difficulties for utilitarianism – utility monsters, interpersonal comparison, Korsgaard's objections – can be met by a form of utilitarianism like Hare's, which is Kantian, and thus that such a form of utilitarianism is indeed the most defensible.

Acknowledgements

My largest debt is of course to my supervisors, Simon Blackburn and Hallvard Lillehammer, who have given me their invaluable encouragement, time and comments, for all of which I am very grateful. This debt extends to my other, earlier teachers at Cambridge, Jimmy Altham and Ross Harrison, under whose guidance I was already working towards the ideas expressed in this thesis.

Over a period of many years now I have discussed my ideas with fellow students, other faculty and students at Cambridge, with colleagues and students at Leeds, Reading and UCL, with other research collaborators, and with audiences at other universities where I have presented parts of this work. I am grateful to all of these people, and particularly to Ben Colburn, Brian King, Carrie Jenkins, Neil Sinclair, Tim Button, Yoon Choi, Sasha Mudd, Gerald Lang, Pekka Väyrynen, Ulrike Heuer, Robbie Williams, Shane Glackin, Jussi Suikkanen, Richard Woodward, Tom Hurka, Wlodek Rabinowicz and Alex Gregory, with all of whom I have had particularly illuminating discussions.

I gratefully acknowledge the financial support of the Arts and Humanities Research Council and the Newton Trust in completing this research.

It has been a long struggle to finally bring this thesis to completion. Throughout I have been blessed with the love, support and patience of my friends and family, and especially of my parents, and I express the utmost thanks and appreciation to them.

Finally I am most grateful to Elizabeth Tepper for reading this work through and supporting me through the final stages of writing.

Table of Contents

CHAPTER 1: THE ARGUMENT FOR EXPRESSIVISM

1.0 INTRODUCTION

The goal of this work as a whole is to present a unified defence of R.M. Hare's views in meta-ethics and normative ethics, combining expressivism, Kantianism and utilitarianism. I will endeavour to show that not only was he mostly right about each of these areas, but that his views in each of them are mutually supporting. In this chapter I expound and defend a version of expressivism that would be unsurprising to Hare, though it differs from his presentationally. In Chapter 2 I defend this expressivism against the objection which has perhaps proved most troubling, the Frege-Geach Problem. In both of these chapters I try to connect Hare's views with the most up to date views in meta-ethics, but I say very little that is inconsistent with his own claims and arguments. In Chapter 3, however, I try to fill the main gap in Hare's argument, by replacing his argument for universalizability as a conceptual truth with a Kantian argument for it as a transcendental, unavoidable condition of agency. Doing so fills out the Kantian aspect of Hare's view that he somewhat neglected, solves very pressing problems in normative justification, and secures Hare's view as a whole against the important shmoralising objection. In Chapter 4 I proceed to defend Hare's argument from universalizability to utilitarianism, modifying it where necessary, and to defend his version of utilitarianism itself. In both Chapters 3 and 4 I provide evidence for the mutual dependence of the different parts of Hare's project: expressivism in meta-ethics, Kantianism about the foundations of morality, and utilitarianism in normative ethics. I begin, as Hare did himself, by considering the problems of meta-ethics.

It is not true that Western philosophy consists wholly of footnotes to Plato[1], but it is closer to the truth to say that meta-ethics is mostly commentary on the *Euthyphro*. In this chapter I aim to draw out an argument for expressivism, showing how a careful consideration of Socrates' famous dilemma closes off the alternatives. There is a historical dialectic which runs in parallel to this argument: it is for the most part a path of progress, with heroes such as Cudworth, Price, Moore, Stevenson and Hare; but reason was blown off course in the last half-century by some bad arguments whose errors I shall also illustrate.

The argument for expressivism proceeds initially by elimination of the alternatives; and the crucial alternative to expressivism is realism. I wish to be clear from the outset that I mean to use the term 'realist' in a slightly unusual way. Contemporary moral realists hold that moral judgements are primarily about some aspect of reality. They are thus descriptivists regarding moral utterances, holding that they describe the world, and cognitivists regarding moral judgements, holding that they are cognitive states (i.e. beliefs). I will call this kind of realism 'constitutive realism', because it views moral facts as constituted by some aspect of reality.[2]

But I wish to include another kind of view under the heading of 'realism', which I think preceded constitutive realism. Non-constitutive realists (e.g. Pufendorf and Hobbes) did not commit themselves to any view about the nature of moral judgement, but rather were interested in answering a somewhat different question: where do moral facts come from? We can see two ways in which this question might be answered by thinking about

[1] The phrase is from Whitehead (1929).
[2] The distinction I draw between constitutive and non-constitutive realism is inspired by Schroeder's (2005) distinction between constitutive and non-constitutive voluntarism.

the debate between voluntarists and intellectualists.[3] This was an argument between Christian philosophers over the interpretation of the idea of a morally perfect God. Voluntarists held that God's moral perfection was a trivial consequence of the moral truth being determined by God's command. Intellectualists held that the moral truth was independent of God's command, and that God's necessary moral perfection had to be explained by God's nature. Roughly then, voluntarists hold that there is a very close connection between moral rightness and being commanded by God. But this leaves two distinct options for voluntarism: either hold that moral rightness is dependent on being commanded by God, without the two being identical (non-constitutive voluntarism), or hold that moral rightness is explained by being commanded by God simply because the two properties are the same (constitutive voluntarism).

Now constitutive voluntarism appears to be just as much a form of realism as naturalism: thinking that moral rightness is identical to being commanded by God is not relevantly different from thinking that it is identical to being pleasant, or conducive to the survival of the species, or any of the other properties suggested by (naturalist) moral realists. And it would be strange to think of non-constitutive voluntarism as being motivated in a significantly different way from constitutive voluntarism; there is in both of them the distinctively realist motivation of wanting morality to be founded on and explained by the way things are (perhaps with the accompanying idea that it is only in this way that moral objectivity can be secured). Realists hold that the way to understand moral talk is to understand what makes it true or false; they see the problem of moral scepticism as a metaphysical one. Dostoevsky's Ivan Karamazov believes that if God does not exist then everything is permitted; what unites various forms of moral realism is thinking

[3] The history of this debate is covered by Schneewind (1997) in much more detail than I attempt.

similarly that there is some aspect or fragment of reality (like the existence of God) that morality depends on, and that this dependence is central to our moral concepts. The crucial point against realism is that moral scepticism cannot be given a metaphysical answer, and thus that it is a mistake to try to understand moral judgements by first considering the moral facts which they are about.

The way I see the relationship between constitutive and non-constitutive realism is that it is non-constitutive realism which is the purest manifestation of the realist motivation: it tells us that morality emanates from some portion of reality, and tells us little more than that. But as I shall argue in §1.1, Cudworth had a devastating objection to non-constitutive realism; I suspect (though I will not try to prove it) that it is because of that argument that non-constitutive realism is so out of fashion and even off the radar today. One major advantage of constitutive realism is that it avoids Cudworth's argument, but at the cost of falling foul of the famous Open Question Argument. As I argue in §1.2, forms of naturalistic realism (such as Cornell realism) which are commonly thought to escape the latter argument fail to do so. Once both non-constitutive and constitutive realism stand defeated, the whole idea of morality being "based on the facts" (insofar as that idea can be made sense of) is revealed as mistaken, and it is this realisation which leads the way to expressivism.

There are, however, some wrinkles to this picture. One important point, taken up in §1.3, is that expressivism properly understood does not deny many of the central claims of constitutive realism. Specifically, the expressivist concurs with the constitutive realist on matters of semantics and metaphysics. The difference between expressivism and constitutive realism lies in meta-semantics: what the Open Question Argument reveals is that even if moral predicates refer to the same properties which constitutive realists say

that they do, this reference does not come about in the normal way. Moral predicates refer to ordinary properties, but they do so by a quite different, practical mode of presentation. It is a mistake to see the problems in meta-ethics as problems at the level of semantics (reference) – the crucial question is why moral predicates refer to the properties they do refer to, and it is here that realists do not have a satisfactory answer. One subtlety here is that once we reach this point it is not obvious how realism is to be defined. If one sees the debate between realism and anti-realism as lying at the level of semantics, then it is tempting to say that since expressivists agree with constitutive realists about the reference of moral predicates, expressivism thereby counts as a form of realism.[4] I do not take this line, since I take the meta-semantic issue to be primary, and because it is more intuitive to describe expressivism as an anti-realist view. The expressivist holds that further explanation is required of why moral concepts appear to be practical; if we do not supplement the metaphysics and semantics of realism then we cannot distinguish morality from any other branch of science.

There are also variants of realism which do not fit so neatly into the framework laid out above, which I discuss in §1.4. Moorean realism, commonly glossed as the view that moral properties are non-natural, is interesting because it seems, like expressivism, to be a descendant of intellectualism, rather than voluntarism. What is wrong with Moorean realism is that it locates the problem with naturalism in the wrong place – in metaphysics rather than meta-semantics. There is another view which attempts to navigate between standard naturalist and non-naturalist versions of realism: non-reductive realism. According to this view, moral properties are shapeless, meaning that although they supervene on natural properties they are not connected to them in any law-like way. I

[4] Gibbard (1996, 2003) says that expressivism, in its most plausible form, ends up as 'sophisticated realism', in contrast to Blackburn's label 'quasi-realism'. Compare Hare 1989a: 87.

argue that although shapelessness is consistent, we can have no reason to believe in it, and it is also unclear how the resulting view is meant to be distinct from expressivism.

In §1.5 I explain arguments for expressivism related to judgement internalism and what I call the antinomy of normative judgement, which consists in the appearance of compelling arguments for each of the apparently contradictory conclusions that moral judgements are desires and that they are beliefs. I explain in §1.6 how exactly quasi-realism ends up resolving this antinomy: it does so by distinguishing between two standpoints, and holding that each conclusion of the antinomy holds from only one standpoint.

1.1 FROM THE *EUTHYPHRO* TO THE OPEN QUESTION ARGUMENT

We begin the attack on realism with the *Euthyphro* dilemma: do the gods approve an action because it is pious (intellectualism), or is an action pious because the gods approve it (voluntarism)?[5] To understand the point of this question, we need to consider the underlying assumptions. Euthyphro takes it for granted that there is some connection between the property of being pious and that of being approved by the gods. That connection is, in modern terms, that they are (necessarily?) co-extensional. Socrates is prepared to grant this co-extensionality, at least for the sake of argument; but he assumes that we need to say something more if we want a satisfactory account of piety. This is because he distinguishes between essence and attributes, and holds that being approved by the gods is a mere attribute of pious actions, rather than the essence of piety. Socrates denies that piety can simply be the same as being approved by the gods (thus he

[5] I do not insist that the reading of the dilemma which I discuss is faithful to Plato, so I will not be engaging with any Plato scholarship. Whether or not Plato intended something different, it is the reading which I discuss in the text which has been so influential in meta-ethics (since at least the Early Modern period).

repudiates what I call constitutive voluntarism). His argument for this is that the gods approve a pious action because it is pious; whereas an action is approved by the gods because the gods approve it. If being pious was the same as being approved by the gods, then we would have to say that an action was pious because the gods approved it. Since being pious and being approved by the gods bear different explanatory relations, they cannot (by Leibniz's law) be identical.

Alas, this argument is questionable; its weak point is the idea of explanation which it employs. For one thing, we might hold that explanations are epistemically sensitive: to Lois, who does not know that Clark Kent is Superman, Superman's appearance will be a better explanation of how disaster was averted than Clark Kent's appearance, even though the *explanans* is really the same in each case. There are two ways of using this point to attack Socrates' argument. One is to say that it would be a mistake for Lois to argue that Clark and Superman are distinct by pointing to their different explanatory powers; for she is wrong about those explanatory powers on account of her mistake about their identity. What this suggests is that apparent differences in explanatory powers are only defeasible evidence of non-identity, for good evidence of identity may convince us that we were mistaken about the explanatory powers. But perhaps this is not a serious threat to Socrates's argument, since it is hard to see what evidence there could be for the identity of being pious with being approved by the gods. Another way of elaborating the point about the epistemic sensitivity of explanations is to hold that it is strictly true that Superman's appearance explains how disaster was averted, and strictly false that Clark's appearance explains how disaster was averted. This is consistent with Leibniz's law if we hold that explanatory claims generate opaque or intensional contexts in which intersubstitutability *salva veritate* of co-referring terms fails. If so, then Socrates' argument would fail, since

being pious might be the same as being approved by the gods despite the truth of different respective explanatory claims.

At this point, however, we need to pay careful attention to the distinction between predicates and properties. When we talk about intersubstitutability, we are thinking of linguistic items – names and predicates – rather than what they refer to – objects and properties. So even if the terms 'is pious' and 'is approved by the gods' can feature in different true explanations, that does not mean that being pious and being approved by the gods have different explanatory powers. Superman and Clark Kent, being identical, must have the same explanatory powers, even if claims using the term 'Superman' strike us as better explanations. But when Socrates asks for an account of piety, is he asking about a property or a concept? His argument is more convincing as an attack on the view that the concepts of being pious and being approved by the gods are identical.[6] What is ultimately most revealing about the Euthyphro dilemma is that it can be posed at all. If we can sensibly ask whether an object's satisfying one predicate is responsible for its satisfying another predicate, then the predicates cannot mean the same, and must express distinct concepts. But arguably this point does not apply at the level of properties. As Kripke showed, there can be *a posteriori* property identities, such that two predicates can refer to the same property without meaning the same. This seems to create room for constitutive explanations, where the fact that the property is instantiated under one mode of presentation (i.e. predicate) can explain its instantiation under another, where the mechanism of explanation is not causation but constitution. This strategy becomes clearer

[6] Though it is dubious even interpreted in this way, since Socrates is apparently committed to a conceptual distinction between an action being approved by the gods and the gods approving that action. Even if such a distinction is credible, it is hard to accept that the conceptual distinction between being pious and being approved by the gods is as weak as that. In the text above I therefore move away from the letter of his argument.

in discussion of Ralph Cudworth's elaboration of the Euthyphro dilemma. But as we shall see in §1.2, there is room to wonder how this talk of *a posteriori* identity can really answer Socrates' question about what we mean when we talk about piety.

Cudworth's insight is to note that voluntarists who hold that what is right is what God commands, must hold that it is right to do what God commands. But it is hard to see how God's command could explain why we ought to do what God commands:

> And if this [doing what is commanded by God] were not morally good and just in its own nature before any positive command of God that God should be obeyed by his creatures, the bare will of God himself could not beget an obligation upon any to do what he willed and commanded, because the natures of things do not depend on will, being not things that are arbitrarily made (γιγνόμενα) but things that are (ὄντα). [Cudworth 1996/1731: 19]

Thus the divine command theorist is committed to there being at least one moral fact which is prior to any command of God: that it is right to do what God commands. But this is fatal to non-constitutive voluntarism, since such a theory aspires to explain **all** moral facts on the basis of divine commands. Hume, drawing on Clarke (2003/1706: 179-80), and targeting Hobbes, makes a structurally identical argument against a social contract theory of morality: the social contract theorist holds that we are obliged to do whatever we promised to do in the social contract; but this cannot account for the obligation to keep one's promises, since that is what makes it obligatory to keep to the contract. Since the obligation to honour contracts must be prior to the bindingness of any particular contract, no contract can explain it:

> It has been asserted by some, that justice arises from Human Conventions, and proceeds from the voluntary choice, consent, or combination of mankind. If by *convention* be here meant a *promise* (which is the most usual sense of the word) nothing can be more absurd

> than this position. The observance of promises is itself one of the most considerable parts of justice, and we are not surely bound to keep our word because we have given our word to keep it. [Hume 1975/1777: §257]

The general form of the argument is as follows: a non-normative fact (that the gods command something, that a promise has been made) can only have normative force in virtue of some antecedent principle (that we should do what the gods command, or what was promised); so any attempt to cite a non-normative fact as the ultimate source of normative principles will fail, because it is only in virtue of some other principle that such a fact can have normative potency. In Jerry Cohen's (2003: 227-8) recent formulation: '*if* any facts support any principles, then there are fact-insensitive principles that account for that relationship of support'.

This is why the issue of identity comes to assume such importance. Some philosophers have been wedded to the idea that there can be an ontological answer to the question of where normativity comes from. But Cudworth's argument makes any such move dubious. Suppose that we think there is a problem in accounting for normativity on a naturalistic framework (with whatever ontology natural science is committed to). Will it help to postulate additional entities? It seems not, because the fundamental normative principles will be insensitive to the fact that these entities exist, and thus it will make no fundamental difference whether they do or not. The only response to such an argument, as recognised by Schroeder (2005: 17), is to say that the principle holding is constituted by the relevant fact: e.g. 'For God to have commanded *X* to do *A* is *just what it is* for it to be the case that *X* ought to do *A*.' In this example, we have an alleged case of two modes of presentation of the same thing: it is presented both under the description 'what *X* ought to do' and under the description 'what God has commanded *X* to do'. This raises the question

of whether normative facts can be explained using this constitutive move. The Open Question Argument (OQA) is an attempt to show that it is not plausible, and thus to shore up the loophole in Cudworth's argument against voluntarism and realism in general.

What this reveals is that there are really two broad types of realism floating around: non-constitutive realism, found in the voluntarists and in Hobbes, which thinks of some set of circumstances as having a privileged status in justifying moral principles, was decisively refuted by Cudworth; contemporary moral realists tend to agree with Schroeder in espousing a constitutive form of realism according to which the moral facts are simply constituted by apparently non-moral facts. When one realises that this is the genuine origin of modern moral realism, the position becomes much less appealing. Suppose that one were in an argument with a divine command theorist, who claimed that all moral obligations depended on the wishes of the gods. One would make Cudworth's point, that there must on this view be a more basic obligation to do as the gods wish; but suppose that the divine command theorist then held that the reason why moral obligations depended on the wishes of the gods was that there was no difference between being morally obligatory and being wished by the gods. Surely at this point one would feel that one's interlocutor was cheating, retreating from an interesting position to a sterile one. The reason why this kind of realism seems sterile is that there now appear to be no rules of the game; anyone can assert that moral rightness is identical to whatever non-moral property she likes, and there will be nothing to choose between the different candidate properties. Of course, realists think that they can answer this concern, and I will show (in §1.2) why they are mistaken about this.

But the first step towards seeing why constitutive realism is mistaken is to follow the OQA, which draws on the intuitions above. The most influential formulation of the

OQA is undoubtedly that of G. E. Moore (1903: 16-21), but the argument first appears explicitly much earlier. Here is Richard Price's version, which is at least as clear as Moore's:

> Right and wrong when applied to actions which are commanded or forbidden by the will of God, or that produce good or harm, do not signify merely, that such actions are commanded or forbidden, or that they are useful or hurtful, but a sentiment [i.e. opinion] concerning them and our consequent approbation or disapprobation of the performance of them. Were not this true, it would be palpably absurd in any case to ask, whether it is right to obey a command, or wrong to disobey it; and the propositions, obeying a command is right, or producing happiness is right, would be most trifling, as expressing no more than that obeying a command, is obeying a command, or producing happiness, is producing happiness. [Price 1948/1787: 16-17]

Price's point is this: if 'right' applied to an action just meant that that action was commanded by god, then it would silly to ask whether it was right to do what god commanded, and it would be tautological to say that it was. To complete the argument, we note that of course it is not silly to ask that question, and that the claim that it is right to do as god commands is a substantive one, so by *modus tollens* it cannot be the case that 'right' just means 'commanded by god'. And the argument is meant to generalise: think of any non-normative predicate you like, and consider the question whether actions satisfying that predicate are right. The idea is that the question is bound to be open, and the corresponding proposition is bound to be substantive. So 'right' cannot mean the same as any non-normative predicate.

It has become common to hold that the OQA makes the mistake of assuming that all analytic truths are obvious. The thought is that for a question to be open is just for the answer to the question not to be obvious. If that were so, then the inference from the openness of the question 'Is producing happiness right?' to the conclusion that it is not

analytic that producing happiness is right would indeed be an inference from non-obviousness to non-analyticity. And non-obviousness does not entail non-analyticity. If it did, then conceptual analysis would be pointless, because we would have to know all the answers before we started (that is the so-called 'paradox of analysis').[7] Perhaps Moore thought that openness and non-obviousness were the same, but I do not; so what is it for a question to be open? When we ask 'Is producing happiness right?', it strikes us that the answer to this question cannot be settled by definition. For moral questions are practical questions, and the meanings of words cannot provide practical guidance (except when the task is e.g. the compilation of a dictionary). Someone who doubts whether producing happiness is right has an intelligible doubt, and if we say that she is merely ignorant about the meaning of 'right' then we misconstrue her thought.

As Price says, to say that an action is right signifies one's approbation of its performance, and this is inconsistent with thinking that it is analytic that some action is right. Hare explains why:

> The meaning of expressions like "A puppy is a young dog" is preserved by expanding them into overt definitions like "The English sentence 'If anything is a puppy it is a young dog' is analytic". [...] On the other hand, a sentence of the form "An A which is C is good" cannot without change of meaning be rewritten "The English sentence 'An A which is C is good' is analytic". For a sentence of the latter type certainly could not be used for commending, whereas sentences of the former type can be and are. [Hare 1952: 91]

The sentence 'It is right to do what God commands' is used to commend doing what God commands. But it could not be used to do this if it were analytic that it is right to do what God commands – so it is not analytic. This makes clear why the OQA only applies to moral

[7] See for example Smith (1986: 293-4).

(and other normative) terms: it is only these terms whose application signifies approbation or disapprobation or similar. So there is no danger that the OQA will commit us to a general paradox of analysis.

There is another, less impressive, objection to the OQA – that is, unlike the previous objection, it cannot be understood as a fair reaction to some confusion or lack of clarity on Moore's part. The objection is that the OQA is questionbegging. This claim was originally made by Frankena (1939); here Alexander Miller explains Frankena's point:

> We can appeal to our conviction that there is an open question [...] only if that conviction is well-founded. But if analytical naturalism is correct, that conviction is not well-founded [...]. So we can appeal to the open question [...] only if we have already established that analytical naturalism is incorrect. [Miller 2003: 15-6]

I reject the first conditional; on the contrary, our conviction that there is an open question is evidence that analytical naturalism (the view that 'right' or 'good' means the same as some naturalistically acceptable predicate) is incorrect. When we are investigating meanings, our semantic intuitions are the only evidence that we have to go on, so the idea that we can only appeal to them if we know that they are correct would make the investigation impossible. Consider Kripke's famous argument (1980: 6-7) that names are rigid designators. That argument depends on the intuition that the truth value of 'Aristotle was fond of dogs' in describing counterfactual situations does not depend on the truth value of 'The greatest philosopher of antiquity was fond of dogs' in describing counterfactual situations. It is meant to follow that 'Aristotle' does not mean 'The greatest philosopher of antiquity'. But if 'Aristotle' does mean 'The greatest philosopher of antiquity' then Kripke's intuition is not well-founded. So does Kripke's argument beg the

question against the description theory of names? Evidently not, but Frankena must suppose that it does. The OQA is in the same boat as Kripke, which is not a bad place to be.

It is now quite common to run the line that what the OQA does is elicit linguistic intuitions, and thus to reject Frankena's charge of question-begging.[8] But this usually goes along with some remarks about how this forces us to be more modest about the strength of the OQA, and to admit that these intuitions might be explained away by the views they conflict with. I agree that in principle the OQA is only a defeasible consideration against naturalism, but only in the way that Kripke's argument for rigid designation is defeasible. Even defeasible arguments like Kripke's can be decisive in showing that the positions they conflict with are deeply unpromising. And what evidence could there be in favour of the view that a normative term like 'right' means the same as some non-normative term? I suspect that the only possibility is an argument by elimination of the alternatives; but since I will show how expressivism represents a wholly coherent explication of moral language, such an argument will effectively be refuted by what follows.

1.2 The Open Question Argument meets Cornell Realism

The possibility that there could be a synthetic identity between moral and natural properties has already been raised. It has become common to attach an almost mystical significance to the work of Kripke and Putnam on semantic externalism, as if meta-ethics was on the wrong track in virtue of a lack of knowledge or understanding of their work.[9] Kripke (1980) showed that there are necessary truths which are not analytic or *a priori* (e.g. that Hesperus is Phosphorus); Putnam (1975) showed that property identity claims,

[8] E.g. Snare (1975), Darwall, Gibbard & Railton (1992: 117), and Miller (2003).
[9] One prominent culprit here is Soames (2003).

particularly identities between natural kind terms, are like that (e.g. that water is H_2O).[10] A group of moral realists based at Cornell (Richard Boyd, David Brink and Nicholas Sturgeon) thought that these observations could be harnessed so as to neutralise the OQA, because the OQA could only (they thought) challenge alleged analytic identities. Here Brink explains what he thinks follows from the work of Kripke and Putnam:

> The naturalist can concede that there are neither synonymies nor meaning implications between moral and non-moral, for instance, natural, terms and still maintain that moral facts and properties are identical with, or constituted by, natural and social scientific facts and properties. The naturalist's identity or constitution claims can be construed as expressing synthetic moral necessities. [Brink 1989: 166]

According to Brink (1989: 163), the OQA contains as a premise the claim (refuted by Kripke) that all necessities are analytic, so it is clear to him that it is hopeless against a position involving synthetic moral necessities. This line (that Cornell realism dodges the OQA) has become orthodox even amongst those who are not committed to Cornell realism. Here are Darwall, Gibbard and Railton attempting an impartial gloss on Cornell realism:

> If moral properties are to be viewed as irreducible natural properties akin to natural kinds, then this sort of naturalism would run afoul of no Moorean argument: no a priori analysis is being offered, and full respect is being paid to Moore's Butlerian motto: "Everything is what it is and not another thing." [Darwall, Gibbard and Railton 1992: 171]

By my lights the mistake here is simple enough: an *a priori* analysis **is** being offered, and that is what I shall argue below. The Cornell realists have a point. You can't refute naturalism just by refuting analytic naturalism. But they are still wrong: switching to

[10] It is worth noting that they did not show that such necessities do not depend on analytic or *a priori* truths. It is analytic that names are rigid designators; so it is analytic that if Hesperus is Phosphorus then necessarily Hesperus is Phosphorus.

synthetic naturalism just adds an epicycle which gives the illusion of avoiding the OQA. To see why, we have to examine Putnam's account of natural kind terms more closely.

What does 'water' mean? According to Putnam, natural kind terms are (approximately) indexicals[11]: to say that something is water is to say that it is stuff of the same kind as the watery stuff around **here**. More slowly, we identify water by its apparent physical and chemical properties. But having those properties does not suffice (and may not even be necessary) for something to count as water: it has to be the same stuff as the stuff around here that has those properties. So XYZ, which is the stuff on Twin Earth which has the same apparent properties as H_2O, does not count as water, because it is not the same kind of stuff as H_2O. The reference of an indexical depends on *a posteriori* questions; but there is an *a priori* aspect to the meaning of the term.[12] For example, it is an *a posteriori* matter where, on any particular use of the term, 'here' refers to; but I know *a priori* that 'I am here' is true, because it is *a priori* that 'here' refers to the location of the person using the term. In the same way, it is *a posteriori* that 'water' refers to H_2O, but it is *a priori* that the watery stuff around here (or, perhaps, the watery stuff which my linguistic community's use of the word 'water' is regulated by) is water. So if it is truly analogous to Putnam's natural kind view, Cornell realism will be committed to certain *a priori* truths involving moral terms. A simple way of putting the point is that inserting an indexical element into a proposed analysis of a moral term will not make the OQA go away. For instance, if the view that 'right' does not mean 'commanded by God' is refuted by pointing

[11] It is largely correct to say that Putnam simply thinks that natural kind terms **are** indexicals. The only difference would be that uses of 'water' to refer to different kinds will not be uses of the same word, whereas uses 'here' to refer to different places will be uses of the same word. This is because we will not want to say that people on Twin Earth speak the same language as people on Earth. The reason for this is presumably that a language is associated with a linguistic community, and Earth and Twin Earth comprise separate linguistic communities. So 'water' is only indexical in a view which sees English and Twin English as two dialects of the same language, in a somewhat loose sense of 'language'.

[12] The distinction is sometimes put as that between broad and narrow content respectively.

to the open question 'Is what is commanded by God right?', then the view that 'right' means 'commanded by God **around here**' will be defeated by the open question 'Is what is commanded by God **around here** right?' So the Cornell realists appear to think that indexical analyses are better able to avoid the OQA, but they have no good reason for thinking so.

Since it is not usually clear exactly how the Cornell realists think that the reference of moral terms gets determined, it is similarly unclear which truths involving moral terms they take to be *a priori*. But as soon as things get more precise, the problem becomes more obvious. One influential formulation of Cornell realism is that of Richard Boyd (1988: 195), who claims that a moral term such as 'right' refers to whatever (natural) property causally regulates its use. This is a philosophical thesis about the meaning of the term 'right' (so it belies Darwall, Gibbard and Railton's claim that Cornell realists are not in the business of providing *a priori* analyses). And as an analysis of 'right', it immediately falls foul of the OQA in exactly the way one would expect: it is an open question whether actions possessing whatever property causally regulates the use of the word 'right' are in fact right. It is indeed quite easy to imagine societies in which the causal regulation occurs in some bad way so that the actions possessing the causally regulating property are not in fact right.[13] And more importantly, if Boyd were right, then in saying that it is right to perform actions which have whatever property causally regulates use of the word 'right'

[13] This is the strategy pursued by Timmons and Horgan (1991, 1992) with their Moral Twin Earth cases. In my view the use of actual counterexample cases is distracting, since it encourages attempts to explain away the counterexamples. The OQA does not need thought experiments to prove its point: if somebody cannot just see that it is an open question whether pleasure is good, then it will not help to imagine a scenario in which pleasure is not good, because she will bite the bullet and hold that pleasure is good even in that case. And if a Cornell realist cannot see that it is an open question whether the property which causally regulates our use of 'right' is rightness, then she will simply bite the bullet when presented with a counterexample. Instead it is better to follow Hare's dialectical path of asking whether she minds constraining herself from commending pleasure, or whether she thinks that someone who doubts that pleasure is good does not understand what the word 'good' means. Since the OQA applies in just the same way to Cornell realism, it is similarly superfluous (and potentially distracting) to support it with thought experiments.

one would not be commending such action. In light of these rather obvious facts, it is mysterious that anyone could think that we give a plausible account of the meanings of moral terms by holding that they refer to whatever causally regulates their use. I suspect that there is some tendency to think that **all** terms must refer to whatever causally regulates their use; but there are clear counterexamples – 'unicorn' does not (and never did) refer to narwhals, and 'hello' does not refer to meetings.

Cornell realism is thus hopeless as a response to the OQA if it simply claims that moral terms are indexical natural kind terms, or that their reference is determined by causal regulation. But the specific details of the proposal are in fact irrelevant, since there is a perfectly general argument which can be given against any possible version of the view. The crucial point which Cornell realists are committed to defending is that there is some function from a moral term and a possible world to the extension of that term at that world, and that this function is somehow embedded in the meaning of that moral term. So it is analytic that the actual extension of 'good' is that of the property f('good',@), and thus analytic that whatever satisfies f('good',@) is good. But since f(x,y) is a naturalistic specification of a property, like 'whatever has the property which causally regulates use of x at y', this runs into the OQA: it is an open question whether everything satisfying f('good',@) is good. One reply to this would be to deny that the relevant function is analytic (though the actual Cornell realists do not do this). But then it suddenly becomes unclear what the fuss was about, because the claim that there is some naturalistically statable function from terms and worlds to extensions amounts to nothing more than supervenience: there is such a function iff the moral supervenes on the natural. This is because what the supervenience of the moral on the natural requires is that there is no variation in moral properties without variation in natural properties, which means that

there is a function from configurations of natural properties to configurations of moral properties, because a function is simply a relation which maps each element of the domain (the configurations of natural properties) to exactly one element of the range (the configurations of moral properties). Since everyone (or at least everyone in the debate) agrees to supervenience, it cannot be a complete statement of Cornell realism. Cornell realists have to claim that the relevant function is determined by the meanings of moral terms, otherwise they are saying nothing more than that they believe in supervenience. Supervenience is not in and of itself a view about the meaning of moral terms or the nature of moral judgement.

Now it may seem that the relevant function need not be analytic, since it is not in the case of natural kind terms which the Cornell realists think are analogous to moral terms. Perhaps is not analytic that 'water' is a natural kind term, because water might not have been a natural kind (i.e. it was epistemically possible that water was not a natural kind). At some stage the epistemic situation with respect to 'water' was the same as that with respect to 'jade'; but 'jade' is not a natural kind term because it turned out that there were two minerals (nephrite and jadeite) which were both counted as jade by the relevant experts.[14] If we interpret 'jade' as e.g. 'the actual stuff called "jade" around here' then we get reference failure, since there is no such single stuff. So if it had turned out that what was called 'water' was really two chemically disparate substances, then 'water' would not have been a natural kind term. But this does not mean that there is no reference function assigned to 'water' as part of its meaning. There is, and could always have been, a difference between the meanings of 'water' and 'watery'. Perhaps the meaning of 'water'

[14] Nephrite and jadeite are not chemically similar: nephrite is rich in calcium and magnesium, whereas jadeite is rich in sodium and magnesium. Yet this distinction was not known until the nineteenth century, after both minerals were classed as jade.

is 'the actual watery stuff around here, if there is any such unique stuff, or all watery stuff, if not'; that would still be a naturalistic function, and could be known to be the meaning without knowing that water is a natural kind. But it will not help the Cornell realist to say that 'right' refers to the property which causally regulates our use of 'right' *if there is any such property*, because the open question arises even when that condition is satisfied. So an analogous analysis of 'good' would still fall foul of the OQA, by the argument above.

1.3 EXPRESSIVISM AND THE METAPHYSICS OF MORAL PROPERTIES

Expressivism is the view that moral claims express conative attitudes: it is thus equivalent to the view (often called 'non-cognitivism'[15]) that moral judgements are conative states.[16] Hare called his view 'prescriptivism'; it is equivalent to expressivism, though couched in different terms, and I will henceforth use 'expressivism' inclusively of Hare's views.[17] The term 'expressivism' perhaps originates with Gibbard (1986)[18], but was

[15] This label 'non-cognitivism' fell out of fashion largely because expressivists rejected it, concerned that it led to the confusion that they were committed to rejecting the possibility of moral reasoning (which is a cognitive activity). If it just means the view that moral judgements are non-cognitive states, and 'non-cognitive' is just a synonym for 'conative', then it is not far wrong, though even this may confuse the issue by seeming to rule out quasi-realism. For these reasons I will use the term 'expressivism' almost exclusively; readers may with care substitute 'non-cognitivism'.

[16] *Pace* Kalderon (2005), as discussed in a later footnote, and Horgan & Timmons (2006), who call their view 'cognitivist expressivism'. Since it is fairly clear that what Timmons & Horgan have in mind is not a rival to quasi-realism, I do not trouble to argue against them, and simply treat their view as a notational variant of the one I argue for (with them using the term 'cognitive', or possibly the term 'belief', in a different way.)

[17] I take Hare (e.g. 1999: 90) to have licensed this. Because Hare was (originally) writing at the height of the Linguistic Turn, he emphasised the nature of moral language over that of moral psychology. So for him, the crucial idea was that moral claims are really (in their deep logical form) prescriptions, kindred to imperatives rather than indicatives (see e.g. Hare 1952). This is hardly inconsistent with expressivism, however, since the point about prescriptions, just like conative states, is that they aim to make the world fit them rather than to describe it. Thus it is quite in order to see prescriptions as the linguistic correlates of conative states, and the former as expressing the latter. The only real point of departure from Hare is that in explaining why the quasi-realist view is superior to realism, I focus on the need to give an explanation of moral psychology, whereas Hare might think that moral language is explanatorily prior and want a different style of argument. I do not consider that resolving the large issues concerning which of language and thought has priority is possible here, so I follow the recent trend in meta-ethics, which I agree with, of focusing on moral judgements.

[18] Gibbard in turn cites Blackburn's (1984) use of the phrase 'expressive analyses'. The competing view, which is to cognitivism as expressivism is to non-cognitivism, is called 'descriptivism', following Hare.

even there used inclusively to refer back to the work of Barnes (1933), Ayer (1946) and Stevenson (1937). What those authors had done was to emphasise the possibility of making conative states (such as approval and disapproval) central to ethical thought without lapsing into subjectivism. Subjectivism is the view that moral claims report conative states: e.g. 'Torture is wrong' reports the speaker's disapproval of torture.[19] It has very unattractive consequences: moral truth is too cheap and two speakers can make claims that appear to contradict each other but turn out to be consistent, making a mockery of moral disagreement.[20]

What Barnes, Ayer and Stevenson noted was that there is a crucial difference between saying that moral claims **report** the speaker's attitudes, and saying that the claims **express** the attitudes. They gave examples of ejaculations such as 'Ow!'; this expresses rather than reports being in pain, because we do not say that it is **true** when the speaker is in pain, but rather **sincere**. The expressivist view is that likewise 'Torture is wrong' expresses and (does not report) the speaker's disapproval of torture; disapproval is the sincerity condition of the utterance without being its truth condition, and of course sincerity does not entail truth.[21] Because of this, expressivism does not face the same problems as subjectivism: it does not make the truth of moral claims trivial, nor does it undermine moral disagreement.[22] It was long thought (though neither Stevenson nor Hare

[19] A more sophisticated subjectivism can have it that it is the attitudes of the speaker's group which are reported; hence subjectivism is not readily distinguishable from meta-ethical relativism.

[20] If A says, 'Torture is wrong,' and B says, 'Torture is not wrong,' according to subjectivism, so long as A disapproves of torture and B does not, they both speak truly.

[21] Gibbard (1990: 85-6) holds that having the state of mind that you express constitutes sincerity for the expression. Ridge (2006) holds instead that you are sincere just in case you believe that you have the state of mind you express. In the text I write as if Gibbard is correct, but adjusting to Ridge's view would not make a substantial difference to the argument, which turns on distinguishing between sincerity and truth, not on any precise account of sincerity.

[22] If A says, 'Torture is wrong,' and B says, 'Torture is not wrong,' according to expressivism if A disapproves of torture and B does not, all that follows is that both claims are sincere. See also Hare 1989a: 14-32 and 1999: 87-95. Unfortunately, expressivism's separation from subjectivism has not been clear to everyone.

agreed), that expressivism avoided subjectivism at the cost of precluding moral truth altogether. This issue will be discussed later in the chapter, and at greater length in §2.2, with the conclusion that this is a mistake. Suffice it to say now that the expressivist approach is to understand the meanings of moral claims in terms of their sincerity conditions, which are conative states. Saying this does not preclude the existence of truth-conditions and the provision of a truth-conditional semantics as well.

At this point it is important to explain the difference between semantic and meta-semantic issues, and how they relate to Cornell realism. I do not claim that my usage of the terms 'semantic' and 'meta-semantic' is completely standard; indeed, it is common in the meta-ethics literature to use the term 'semantic' in a capacious sense which glosses over the distinction I intend to make. But since there is a real distinction here which needs to be highlighted, and since these terms are the best available, readers should take what I say about the terms here as stipulative. Semantics concerns the truth conditions of sentences, and the references of names and predicates. Meta-semantics concerns the facts which determine these truth-conditions and references; these facts can still be thought of as facts about the meanings of words in a broad sense.

It is important that the mistake in Boyd's view concerns meta-semantics rather than semantics. Suppose that the property which in fact causally regulates our use of the term 'right' as applied to actions is Fness. Then, on Boyd's view, 'right' refers to Fness (this

---

Jackson & Pettit (1998, 2003) argue that expressivism entails subjectivism because insofar as having the relevant attitude makes a claim sincere, it is ok to make it, and if it is ok to make a claim then it is true. This is a poor argument (as Ridge 2006 points out), because sincerity does not imply truth; the sense in which an assertion is ok if sincere is simply different from the sense in which it is ok if true, and Jackson & Pettit effectively equivocate on this point. There are obviously many uncontroversial examples of sincere but false assertions unconnected to expressivism. Suikkanen (2009) worries that the expressivist is still in trouble absent an account of the accuracy of moral claims (i.e. moral truth), because it is only if we have such an account that the contrast between truth and sincerity can really be drawn. But the expressivist will say that it is the function of normative ethics to flesh out moral truth (by fleshing out ethical theory itself), meaning that so long as expressivism does not somehow get in the way of doing normative ethics, there is nothing to worry about here. Perhaps some of the discussion in Chapter 3 would alleviate Suikkanen's concerns here.

is a semantic claim), and rightness is Fness. It is important to note that expressivists may not want to disagree with this part of the view in the end. After all, it is for normative ethics rather than meta-ethics to decide whether rightness necessarily coincides with Fness. For the expressivist then, the fact that Fness causally regulates 'right' is no reason to think that Fness is rightness; but they might coincide all the same. Boyd's real mistake is to think that it is part of the meaning of 'right' to refer to whatever property causally regulates the use of that term, and this is a mistake about the meta-semantics. We might conclude then that meta-ethical questions correspond to meta-semantic questions, whereas semantic questions about moral terms fall under normative ethics. The reason that it is important to make this distinction is that it helps to snuff out a common misunderstanding of the disagreement between expressivists and Cornell realists. The misapprehension is that since expressivists disagree with Cornell realists about the meanings of moral terms, they must be disagreeing with the Cornell realist semantics, specifically with the idea that moral predicates refer to natural properties. But this is not the aspect of Cornell realism with which expressivists disagree, because it is not this aspect which is targeted by the OQA.

Indeed, Allan Gibbard (2006) has argued that on plausible assumptions expressivists are committed to moral properties being natural properties, and on even weaker assumptions expressivism is compatible with this kind of identity.[23] The gist of Gibbard's argument can be understood by thinking about (i) what normative ethical theory has to say about the extensions of moral predicates, and (ii) what metaphysics has to say about property identity. Regarding (i), an ethical theory such as utilitarianism tells us which actions are right not just in the actual world but in all possible worlds: it provides

[23] A similar argument was given by Frank Jackson & Philip Pettit (1996), and defended by Jackson (1998: 122-3 and 2001: 655), though without having expressivism in mind.

not just the actual extension of 'right' but its function from worlds to extensions. Moreover it tells us which actions are right in terms of which natural properties those actions instantiate (because e.g. maximising preference satisfaction is a natural property). So if some ethical theory such as utilitarianism is correct (i.e. any theory which gives moral verdicts on the basis of descriptions of situations in terms of their natural properties), then rightness is necessarily co-extensive with some natural property (whose definition is as complex as that theory).

Regarding (ii), it is at least very plausible that necessarily co-extensive properties are identical. The kinds of alleged counterexamples that are sometimes given, e.g. (for a closed plane figure) having three vertices (triangularity) vs. having three sides (trilaterality) or being equiangular vs. being equilateral, are more convincingly dealt with as cases where there is a single property with more than one potential mode of presentation and hence more than one conceptual correlate.[24] Brad Majors, responding to Gibbard's (and Jackson's) reliance on (ii), says the following about the equiangular/equilateral case:

> The plausible view of the situation is surely that we have here two different properties, rather than two different ways of conceiving the same property. One is the property (more precisely, the relation) of having *sides* all the same length as *x*, and the other is the property (relation) of having *angles* all the same size as *x*. Manifestly, sides and angles are not the same things. [Majors 2005: 488]

So Majors is effectively arguing that since sides and angles are not the same ('side' and 'angle' do not co-refer), neither are being equilateral and being equiangular (i.e. 'equilateral' and 'equiangular' do not refer to the same property). This is an invalid argument because it would prove too much if it were valid. 3 and -3 are not identical, but

[24] There are theories of properties in which a property just is a function from worlds to extensions; but that strong claim is unnecessary here.

3^2 and $(-3)^2$ are. Why should the property of having equal *sides* be distinct from the property of having equal *angles*, just because sides and angles are not the same, any more than the square of *3* is distinct from the square of *-3*? Sometimes we have two names for the same object, and we discover that they co-refer, and sometimes we have two predicates (or concepts) for the same property and we discover that they co-refer; in each case this co-reference may come as a surprise, but no one (at least after Frege) should think that all identities have to be trivial.[25]

What we can see, from putting (i) and (ii) together is that ethical theory plausibly delivers some natural property which is necessarily co-extensive with rightness, and metaphysics tells us that in that case that natural property **is** rightness. And we reach this conclusion without any distinctively meta-ethical premises, but rather simply from making observations about normative ethics and metaphysics which are entirely untouched by expressivism. We could well conclude then that expressivists are committed to seeing moral properties as natural properties. But of course not everyone agrees that ethical theory will end up being anywhere near as systematic as utilitarianism. Some philosophers think that although rightness (like other normative properties) supervenes on natural properties, any specification of all the possible right actions in terms of natural properties would be unsurveyably disjunctive. I discuss such views in more detail in §1.4 under the banner of non-reductive realism.[26]

It is not clear that even the truth of such views would make much difference to the metaphysics, because it might still be argued that even if the only naturalistic definition of

[25] For further helpful discussion of this issue, see Streumer (2008: 541-5).

[26] I identify non-reductive realism with the view that moral properties are shapeless (from the non-moral point of view). This is meant to be a charitable reading, because I agree with Streumer (2008) that any stronger form of non-reductive realism is ruled out by the Jackson/Gibbard argument; but it does not seem that Streumer includes the shapelessness view when he says that the argument rules out non-reductive realism.

rightness was unsurveyably disjunctive the property would still be natural.[27] Even supposing that the non-reductive realist has this kind of loophole, and we cannot conclude quite so easily that moral properties are natural properties, there is still a very important upshot for the consideration of expressivism. It is yet a further demonstration that arguments about the metaphysics of moral properties are irrelevant to the dispute between expressivists and Cornell realists. Insofar as Cornell realists insist that the OQA poses no obstacle to the identity of moral and natural properties, they are right, but they are missing the point in supposing that accepting such identities would count as a victory for realism against expressivism.

The overall verdict on Cornell realism must be that it is a dead-end, which only seemed tempting because of a misunderstanding of the point of the OQA. The OQA does not make the mistake of holding that all necessities are analytic. What the OQA targets is the idea that it is reference-fixing (or reference-borrowing) intentions which are primary when we use moral terms. All that the Cornell realists contribute to the discussion is the idea that the external world could help to fix the reference of moral terms; but that does not mean that there is no reliance on our intentions (not even Kripke thought so), and the reference-fixing ingredient needed in our intentions is enough to get the OQA going. But this does not refute the idea that moral predicates refer to natural properties, or show that it is incompatible with expressivism.

That leaves a few questions to explore. First, we will need an explanation of how this kind of naturalism can be compatible with expressivism (even given the argument above), and why in agreeing to that kind of naturalism the expressivist does not simply concede all the ground to the realist. I have already given some explanation of this, but the

[27] This is Gibbard's (2006) view.

task will be taken up more fully later in this chapter. Second, we can at this point ask whether, if the expressivist accepts that normative properties are constituted by natural properties, she must also accept Schroeder's reply to Cudworth. Recall that, according to Schroeder, this constitutive solution allows for normative facts to be explained by natural facts, because the explanations will be completed by property identities that do not themselves require explanation. The answer is that, perhaps surprisingly, Schroeder's reply does not go through given expressivism, because then the relevant property identity cannot be treated as a brute fact in need of no further explanation. But this issue falls outside the scope of this chapter, since it concerns the consequences of expressivism for normative explanation, rather than the justification of expressivism itself. So the story of Cudworth's argument will be taken up again in Chapter 3.

1.4 'GOOD' IS NOT THE NAME OF A CHARACTERISTIC

I have said that this chapter would argue for expressivism. But the arguments up to this point have been against naturalism, and the main proponents of those arguments (Cudworth, Price, Moore) did not take them to establish expressivism. Indeed Moore is perhaps the most famous proponent of non-naturalism in meta-ethics, and Cudworth, Clarke and Price held rather similar positions. So in this section I want to explore the non-naturalist response to the arguments discussed above, and to explain why this response either fails to overcome the OQA or else cannot be distinguished from expressivism.

An important preliminary point here is that the terms 'natural' and 'non-natural' are potentially misleading. We have already seen that the target example which Price uses when formulating his version of the OQA is analytic (constitutive) voluntarism: the view that claims about what is right can be analysed as claims about what God commands. The

OQA is evidently just as powerful against analytic voluntarism as against the analytic utilitarianism discussed by Moore; so if the OQA distinguishes between naturalism and non-naturalism then voluntarism must be understood as a naturalistic view, despite involving theism. So if invocation of the paradigm supernatural entity does not count as non-naturalist, what does the distinction between natural and non-natural properties amount to? The answer lies in how the OQA works. What it does is show that no normative predicate means the same as any non-normative predicate. But of course the non-normative predicate can involve supernatural objects. So when we talk about properties rather than predicates, the crucial distinction (which Moore is looking for) must be between properties of which we have an independent grasp *a priori,* and those of which we have no such grasp.[28] This is, I imagine, what leads Moore to insist that goodness is a simple property: it cannot be analysed, and all that can be said is that it is the property which 'good' refers to, and this seems to make his view immune to the OQA. We cannot point to an open question which uses some description of the property to ask whether it is good, because no such description is forthcoming.

The point about the redundancy of the natural/non-natural distinction, and thus the OQA, becomes telling against Moore once we extend it to the synthetic versions of naturalism considered above. As has already been argued, on such views there is always some independently graspable fact in virtue of which normative terms refer to natural properties, and this leaves them open to the OQA. Take the claim that 'good' refers to the property that it does in virtue of that property regulating our use of the word 'good'; that was ruled out because it is an open question whether what regulates our use of the word 'good' is good. But clearly it does not matter here that what regulates our use of the word

[28] Broad (1933) makes the distinction in roughly this way.

'good' is a natural property. In propounding his doctrine that good is simple and non-natural, Moore suggests that our only grasp of it is via our moral intuition. But if this is to say that 'good' refers to whatever property regulates our moral intuitions, then Moore's theory is subject to the very same objection as Cornell realism: it is an open question whether what regulates our intuitions is good.[29]

Perhaps Moore did not really mean to say that goodness is the property which regulates our intuitions. But if he does not say this it is altogether unclear what grasp we have of the good at all, for example what distinguishes it from other simple non-natural properties, or why we are all talking about the same property when we use the word 'good'. It will not, for instance, help to say that good is marked out from the other simple, non-natural properties by being prescriptive; for that is just to say that it is the property which we ought to pursue, and in order for that to be illuminating we must have some prior grasp of what is involved in this 'ought'. That line would thus prevent Moore's non-naturalism from being a positive account of the most basic normative or ethical concepts. If Moore is not giving such a positive account, then we should ask whether what he says is really a realist view at all. After all, shorn of the intuitionism which makes it vulnerable to the OQA, all Moore's position amounts to is that there are moral properties, and that they are not independently identifiable *a priori*. But this seems to be just what expressivists think: they think that some actions are right and some are not right, so that there is a property of rightness in some sense, and it is unclear in what way Moore's belief in moral

[29] Of course I do not deny that our intuitions ought to conform to the good. But that is a normative sense of regulation; I mean it is an open question whether they are causally regulated by the good. It is perhaps also worth noting that Moore's view is vulnerable to a Cudworth-style argument. For Moore holds that goodness is never an intrinsic property of objects. If this is because an object is always made good in virtue of some intrinsic property or properties which it possesses, then we can ask why that property is good-making. If Moore answers that it is because that property is good, then he is embarked on a regress; but if he holds that what makes a property good-making is something other than that property being good then he seems not to have given an account of the most fundamental normative property.

properties is any stronger. Of course, Moore is a cognitivist, and expressivists are not. But since the whole point of Moore's arguments is that moral terms have no primary descriptive meaning, it is a mystery what the content of moral 'beliefs' would be on his view. Surely the simplest option is just to abandon cognitivism as a last vestige of realism that makes no sense when separated from the rest of the edifice. The verdict on Moorean realism must be that whilst it is the logical endpoint of the realist who realises the strength of Cudworth's argument and the OQA against other forms of realism, it strips away so much from realism that it collapses into expressivism.

There is another brand of realism which shares the Moorean dissatisfaction with straightforward naturalism, but also avoids the pointless metaphysical extravagance of postulating non-natural properties: non-reductive realism.[30] According to this view, moral properties supervene on natural properties, without being reducible to them (see e.g. McDowell 1981: 144-5). What this means is that there is no systematic moral theory specifying law-like connections between naturalistic descriptions of situations and moral verdicts on those situations, i.e. no ethical theory as traditionally conceived. Moral properties are thus thought of as 'shapeless' (highly disjunctive and unnatural, but not spooky) from the scientific point of view.

It can be hard to get to grips with the motivations for non-reductive realism, or to understand exactly where it parts company from expressivism.[31] The most illuminating angle to approach here is to consider two phenomena which the non-reductive realists point to in trying to explain why expressivism must be wrong. The first phenomenon

[30] Here I am summarising the views of McDowell (1979, 1981, 1987), Dancy (1993: 84-6) and Kirchin (2010), though they are far from alone.

[31] I do not intend, for example, to wade into the thicket of McDowell's (1981) attempted use of Wittgenstein's rule-following considerations to attack expressivism, though for explanation of where McDowell goes wrong see Blackburn (1981) and Lang (2001).

concerns so-called 'thick' terms and concepts[32], such as 'courageous' and 'cruel', and the impossibility of disentangling them.[33] Hare (1952: 121) had argued that a thick concept effectively involved the conjunction of a description and an evaluation, with the description fully determining its extension, and the evaluation only then entering in as a verdict applying to things falling within the extension. Whilst this might be plausible for some terms, such as derogatory epithets, McDowell (1981: 144) argued, followed by Williams (1985: 130, 141-2), Dancy (1996), and Putnam (2002: 34-40) *inter alia*, that with other thick terms the extension is driven by the evaluation, so that the evaluative and descriptive components of meaning could not be disentangled in the neat way envisaged by Hare. People disagree about the extension of 'courageous', in large part because they disagree about how to evaluate different actions and characters. It turns out, however, that although Hare may have been hasty in committing himself to an inadequate analysis, it is perfectly possible for expressivists to accept analyses of thick concepts which allow the evaluation to drive the extension. For example, we could say that being courageous involves accepting lesser risks or harms to oneself for the sake of a greater good, and it is good for doing so; nothing in such an analysis troubles the expressivist, because it allows that fundamentally evaluative and descriptive concepts are distinct, entangled as they may be within more complex concepts.

[32] It is notoriously difficult to give a precise characterisation of thick concepts (rather than simply pointing to examples). From the point of view of the expressivist, they are concepts which are impurely evaluative, with some mixture of evaluation and description built into their meaning. Such a definition is controversial, though it can serve well enough here as it does not beg any questions against the non-reductive realists. But see Eklund 2011 and Väyrynen 2012 for discussion.

[33] In the discussion of this paragraph I draw on Elstein & Hurka 2009. I therefore specifically acknowledge the co-authorship of Tom Hurka with respect to this material (though he is of course not responsible for any errors in my presentation here).

The other idea, even more central to the non-reductive view, is that of shapelessness, which is supported by pessimism about ethical theory.[34] Non-reductivists claim that moral concepts outrun any theory, because however complex the theory some counterexample is always possible. This is because however complex a situation is (in terms of its morally relevant features), it is possible for some additional complexity to be added which changes the moral verdict.[35] Now expressivists will tend to simply deny this alleged data, and there is a danger of a simple clash of intuitions. If moral concepts really were shapeless, expressivists would have a problem, because they see evaluations as reactions to patterns of similarity between the objects of evaluation that are there prior to the evaluation itself. In contrast, the non-reductive realist thinks of these patterns as being imposed onto the world by our sensibilities.[36]

Fortunately for the expressivist, neither the outrunning argument nor any other can give us reason to believe in shapelessness. The non-reductive realist faces a dilemma: are moral properties causally significant or not? The expressivist can cheerfully answer in the affirmative, given the argument of Gibbard (2006) discussed in §1.3 above that expressivists can think of moral properties as natural properties.[37] But causally significant properties are paradigmatically ones which are shapely from a non-evaluative perspective: we are able to form non-evaluative concepts of such properties on the basis of their causal

[34] Elstein & Hurka 2009: 532-3 argue that the arguments concerning disentangling and shapelessness (called there 'uncodifiability') are not really connected, and that shapelessness does not have anything particularly to do with thick concepts as opposed to other evaluative ones. For a contrary view, see Roberts 2011.

[35] See McDowell 1979: §4 and Kirchin 2010: 5-6.

[36] For this reason we can think of non-reductive realism as really being a more radical kind of projectivism about morality than expressivism is.

[37] Given this, the literature on moral explanations looks fundamentally misguided. Sturgeon (1986, 1991) argued that there are genuine moral explanations, and that expressivism could not account for this; Blackburn (1991a, 1991b) tried to show, in defence of expressivism, that such explanations were dispensable. But if Gibbard is right, and he is, then expressivists have no more trouble accounting for moral explanations than Cornell realists like Sturgeon.

roles. So the non-reductive realist, who insists on shapelessness, must deny that moral properties are causally significant.

The problem here is that this threatens our knowledge of moral properties, and so the evidence for shapelessness. Recall that the evidence for shapelessness – outrunning – is meant to be the accumulation of instances of the following kind: we can see that a certain action is right even though it is not predicted to be by the best pre-existing theory of rightness. But it is deeply mysterious how we could see that an action is right if rightness is not a causally significant property. Perception is causal; indeed it is hard to see how there can be any a posteriori knowledge of a property's presence or absence that does not ultimately depend on the causal powers of that property. Hence the problem with the second horn of the dilemma: we could not have the evidence that the argument for shapelessness relies on if moral properties were not causally significant. The only plausible argument for shapelessness thus undermines itself, and it seems that any such argument would have the same problem. If we can have no evidence for shapelessness it seems that non-reductive realism is ruled out as an option.

The ultimate moral of Cudworth's argument and the OQA is that any attempt to explain our moral concepts in terms of properties ends up being incoherent. The key claim of expressivism on this view is a negative one: that 'good' (like right' etc.) is not the name of a characteristic at all. In putting forward this doctrine we must be careful not to say more than we mean. It is not part of the expressivist view that 'good' has no extension. After all, expressivists will want to continue to say that some things are good; and thus they hold that those things fall within the extension of 'good'. And indeed that extension picks out a (natural) property, which 'good' will refer to. But grasping that extension is not the way to understand what moral terms mean, and what kind of concepts moral concepts

are. Although moral concepts refer to natural properties, they are not primarily referential. This gives the expressivist work to do: what does it mean to say that moral concepts refer but are not primarily referential?

The key here is to emphasise that the point of the expressivist variety of anti-realism at least is not to deny the claims of realism on the whole but to add to them. We do not get to expressivism by thinking that the metaphysics or semantics of realism are wrong, or too ambitious. For example, it is not that the realist is wrong that there are moral facts[38], but that the realist simply takes this for granted without explanation. What expressivism claims to be is the best explanation of how morality can be both factual and practical; that is as the best resolution of the puzzle discussed in the next section.

1.5 THE ANTINOMY OF NORMATIVE JUDGEMENT

Up to this point the argument given for expressivism has been rather unusual. The standard approach is to think about the difference between cognitivism and non-cognitivism, and argue for non-cognitivism, and thus expressivism, on the basis of judgement internalism.[39] So let us now backtrack to see how the argument against realism given above links up with that approach. Cognitivists hold that moral judgements are cognitive mental states (i.e. beliefs); non-cognitivists hold that moral judgements are conative mental states (i.e. desires).[40] The distinction between cognitive and conative

[38] Ayer (1946) famously does deny that moral claims are factual or truth-apt, but this is because he is a verificationist, and of course expressivism is not committed to verificationism. Even given Ayer's verificationism, his argument that moral claims are not truth-apt fails. Insofar as Ayer himself makes moral claims (which express his attitudes) he takes moral predicates to have extensions and so there is no obstacle to a theory of truth applying to those claims.

[39] The view that moral judgements are intrinsically motivating, and so like desires rather than beliefs.

[40] There is a risk of confusion concerning the definition of 'cognitivism': one might think that cognitive states are the only states capable of being true or false, or of rational appraisal. These views are perhaps plausible, though I will argue against them, but they should be recognised to be controversial, and not part of the definition of 'cognitive'. Otherwise we will end up begging the question against non-cognitivism.

states is to be understood via Anscombe's (1957: 56) direction of fit metaphor: cognitive states have mind-to-world direction of fit, and conative states have world-to-mind direction of fit.

Expressivists (including Hare) have long argued that moral judgements must be conative states, by an argument based on moral judgement internalism and the Humean Theory of Motivation as follows[41]:

(MJI) Moral judgements are intrinsically motivating, meaning that they are capable of motivating action in combination with (other) beliefs, and not requiring (other) desires.

(HTM) Action is always explained by a combination of beliefs (cognitive states) and desires (conative states): beliefs alone cannot motivate action without desires, and beliefs and desires are distinct and never necessarily connected.

(CON) (Therefore) Moral judgements are conative states (like desires).

The argument is clearly valid, but its premises are controversial. (HTM) is probably the majority view in meta-ethics, and is even more predominant in philosophical assumptions about psychology; the most common reason for denying it is precisely to avoid expressivism (i.e. by internalist cognitivists). There is a very simple argument in favour of (HTM) from the direction of fit account of beliefs and desires. Insofar as beliefs are characterised only by their being correct insofar as they fit the world, they do not alone make any call for changing the world, i.e. to action. In contrast, desires are characterised precisely by demanding that the world fits them, so that a desire (in concert with the belief to the effect that it is not satisfied) calls out for something to be done.

[41] This argument is central to Hare's (1952) way of arguing for expressivism, and it is influentially discussed by Smith (1994). See also Hare 1999: 96-108.

Regarding (MJI), at least some of the disagreement appears to stem from a simple clash of intuitions – although it seems very intuitive to me[42], this is apparently far from universal. For this reason, I accept that the above argument will not convince everyone, and I think there are more convincing observations to be made which lead directly to (CON), without going via (MJI). Before explaining this, however, I want to deal with some of the purported counterexamples to (MJI) in the literature. The reason for doing so is that even if we do not rely on (MJI) in arguing for expressivism, it is clearly entailed by expressivism, so if there really were counterexamples to it they would equally threaten expressivism itself.

The most famous kind of counterexample involves the so-called 'amoralist', who makes moral claims but professes not to care about morality: she has views about what morality requires, but considers those requirements to have no relevance to her.[43] Hare (1952: 124-6, 163-5) responded to this challenge by claiming that if such amoralists make moral claims sincerely, they must do so in an 'inverted commas' sense. Hare's idea is that once a group or society has an established moral system, even those who do not believe in moral requirements can talk about morality, meaning not what **is** morally required, but simply what **their group holds to be** morally required. Since judgements about what a group holds to be morally required are not moral judgements, the amoralist is not making moral judgements, and it is no surprise (or problem for MJI) therefore that she is not

[42] I find Hare's (1952: 148-9) Missionaries and Cannibals example particularly compelling here. The idea is that if we had to explain the meaning of moral terms to those who shared neither our language nor our moral views, it would be hopeless to say that moral terms referred to the properties picked out by our moral system. What we should do instead is explain that to use a moral term is to guide action in the way you think appropriate.

[43] Of course there really are amoralists, but they are people who straightforwardly deny that there are any moral requirements, and they will not be counterexamples to (MJI). There is something absurd in using the term to refer to people who people who make moral claims rather than those who deny morality. In the text I use 'amoralist' simply as a term of art, following the usage in the literature, e.g. Stocker (1979) and Brink (1989: 46), however misguided.

motivated. Notice that the amoralist may not realise that she uses moral terms in an inverted commas sense; she may think that moral terms are simply descriptive, referring to the established group standards directly (perhaps because she has received her moral education from wrong-headed philosophers).[44] In that case she simply misuses moral terms – she is sincere, but confused. We are all capable of using evaluative terms in an inverted commas sense, to indicate precisely when we are talking about standards which we do not subscribe to; the amoralist is just someone who does this all the time. Clearly such a person's practice could only be parasitic on a prior, morally committed usage, so as Hare says the descriptive meaning of moral terms is secondary to the evaluative meaning (which is morally committed).[45]

The other important kind of counterexample offered is that of depression[46]. The basic idea is that when we are depressed our moral motivation is reduced or eliminated, so that although our moral judgements may stay the same, we are no longer motivated by them. This is a quite different case from that of amoralism, and an inverted commas response would be unconvincing here: it would be implausible that the onset of depression would remove a genuine moral judgement and replace it with a matching inverted commas one. And yet what the depression case indicates is not that anything is wrong with (MJI), but rather that we need to be careful about how internalism is formulated. The key thing to notice here is that even desires lose their motivational

[44] Hare (1952: 125) calls this the 'conventional' use.

[45] Hare (1952: 118, 170-2). Whilst I find Hare's response to the amoralist problem compelling, Bromwich (forthcoming) has further helpful thoughts for the internalist here. One objection to Hare's account that may occur to the reader is that of Svavarsdóttir (1999: 189), which is that we can imagine an amoralist making moral claims which (she knows) are not in fact shared by her group (so cannot be construed simply as descriptions of what they think). But presumably in such cases she deduces those claims from others which are accepted by the group, and in that case she is simply describing what the group are committed to thinking, even if they do not actually think it – this still falls short of making genuine moral judgements.

[46] Sometimes called 'accidie' in the literature. For discussion of such cases see Stocker (1979), Smith (1994), Mele (1996), Audi (1997: 231) Svavarsdóttir (1999: 164-5) and Shafer-Landau (2000: 273-4).

efficacy when we are depressed.[47] The onset of depression may very well not change our preferences at all, but simply lead us to irrationally fail to act on them. Since what the expressivist is interested in is the thesis that moral judgements are conative states, depression cannot be a counterexample to expressivism, as it displays no asymmetry between moral judgements and conative states: both have their motivational efficacy weakened by depression.

It is not then compulsory for us to consider exactly how to formulate internalism so as to deal with the depression case, but as it happens the above formulation of (MJI) is adequate. The crucial point is that moral judgements are **intrinsically**, but not **necessarily** motivating.[48] As we have observed, even desires and other conative states may not motivate under depression; but all this shows is that those states are not necessarily motivating. It is platitudinous that desires are intrinsically motivating; I have no idea what else it could be for a state to be intrinsically motivating. Of course, if an opponent insists for some reason that a state cannot be intrinsically motivating without being necessarily motivating, there is another formulation of internalism to fall back on, which is the one expressivists ultimately believed in anyway: that moral judgements have the same motivational role as conative states, whatever that turns out to be.[49] I conclude therefore

[47] As pointed out by Arkonovich (2001), though not as part of a defence of internalism. One point to emphasise: whilst it is debatable whether desires survive depression if their motivational force does not, the evidence here is on a par with the case of moral judgements: just as people may report their moral judgements being unchanged when depressed but not motivating them, they may do the same concerning their desires. This means that non-motivating moral judgements and non-motivating desires seem to stand and fall together, just as expressivism predicts.

[48] In saying this I echo Dancy (1993: 23-4), though he is interested in defending the kind of besires view which goes along with non-reductive realism, rejecting (HTM).

[49] No discussion of formulations of internalism would be complete without mentioning that of Smith (1994), which is that moral judgements are motivating so long as the agent is rational. The rationality qualification is introduced because of the depression case: depression is or induces a form of irrationality which blocks normal motivation. Superficially at least the Smith formulation does not fit well with expressivism, because conative states intuitively motivate regardless of rationality. I have argued in the text, however, that forms of irrationality such as depression can and do interfere with the motivation supplied by conative states. Given this (and an appropriately narrow understanding of rationality), Smith's formulation is both acceptable and

that (MJI) is at least defensible enough that it is a strength not a weakness of expressivism, even if it is not the best path to argue for expressivism.

There are platitudes about moral claims which point more directly to expressivism, bypassing internalism. For example, consider the following claim:

(T) Torture is wrong.

Then the following are platitudes:

(NEC) It is a necessary condition for sincerely asserting (T), knowing what it means[50], that one disapproves of torture.

(SUF) It is a sufficient condition for sincerely asserting (T), knowing what it means[51], that one judges that torture is wrong.

Taken together, (NEC) and (SUF) imply that judging that torture is wrong involves disapproving of torture, because a sufficient condition for X entails all of the necessary conditions for X. Since disapproval is a conative attitude (it has desire-like direction of fit), we can conclude that judging that torture is wrong involves a conative attitude, and thus, generalising, that moral judgements involve conative attitudes. The simplest way this could be true (and the only way compatible with the Humean Theory of Motivation) is if moral judgements **are** conative states. Of course there are also platitudes which make things more complicated here, and seem to count against expressivism. It would be wrong to take expressivism as established without confronting the real difficulties in moral psychology that any view has to overcome. Rather than simply seeing expressivism as the conclusion of a direct argument, we have to think of it (in its quasi-realist form) as the best

---

compatible with expressivism.

[50] This qualification is required because there could be cases where someone asserts (T) on the basis of testimony, despite not understanding it: a case of sincere assertion not requiring disapproval.

[51] Here the qualification is required because a non-English-speaker could, through testimony, believe that (T) was false without understanding it, and yet still judge that torture was wrong. In asserting (T) one would thus be insincere despite making the judgement expressed (in English).

explanation of puzzling features of moral thought and discourse, which I now begin to explain.

There are two platitudes about normative claims that end up creating an antinomy for normative judgement. First, normative claims have correctness conditions: a claim that torture is wrong, or that simpler theories are superior, can be correct or incorrect.[52] Second, they have compliance conditions: the claim that torture is wrong is complied with if people do not torture; the claim that simpler theories are superior is complied with if we accept theories that are simpler, *ceteris paribus*. These two conditions are not usually the same: the claim that torture is wrong may be correct whether or not it is complied with, and it may be complied with whether or not it is correct. This is, I think, what Hare (1999: 1-18) was aiming to express by saying that moral claims are objective prescriptions.[53] A claim is prescriptive just in case it has compliance conditions. To say that a claim is objective means at a minimum that it has correctness conditions and, going slightly beyond what is said above but still within the realm of platitude, that those correctness conditions are not trivial or subjective i.e. not simply dependent on the claim being made or accepted.[54] Indeed, it is the fact that moral claims are prescriptive, that they have compliance conditions, that Hare used to explain how the OQA works, as we saw above. When Hare says that moral claims are used for commending, unlike claims involving non-

[52] I talk of 'correctness' rather than 'truth' here simply to make the claim more clearly platitudinous. This is not to deny that correctness and truth are the same.

[53] Whether Mackie (1977: ch1) intended the idea of objective prescriptivity in the same way when he originated it is less clear, because Mackie seems to build realist assumptions into what he means by 'objective', and he also sometimes seems to think of the prescriptivity of the moral facts in terms of the facts themselves motivating as opposed to beliefs about the facts. But if we read Mackie charitably we might well think that he has the same basic problem in mind that I have: the most plausible way in which moral facts and judgements are objectively prescriptive is that they have both correctness and compliance conditions. See also Hare (1999: 1) on Mackie's view.

[54] What else might be required by objectivity is discussed in Chapter 3.

normative predicates, what he means is that moral claims have compliance conditions which claims about the meaning of moral terms lack.[55]

So far the existence of compliance and correctness conditions on moral claims creates no problem: we can readily admit that normative claims are associated with both kinds of condition, because there is no reason to think that the two conditions conflict. Acknowledging that both conditions are present, at least intuitively, seems necessary for saying anything sensible about meta-ethics. The trouble starts when we ask about the connection between normative claims and normative judgements. For it seems that if normative claims have both correctness conditions and compliance conditions, so must normative judgements, because we take normative claims and corresponding normative judgements to have the same contents (in a loose sense of 'content'). The claim that torture is wrong and the judgement that torture is wrong must have the same content, and thus the same correctness and compliance conditions, since otherwise we would not link the two with the clause 'that torture is wrong'.

The problem which emerges is that to claim that normative judgements have both correctness and compliance conditions, and that the two can come apart, is tantamount to claiming that normative judgements are both beliefs and desires. This is because we understand the distinction between beliefs and desires in light of Anscombe's direction of fit metaphor. We say that beliefs have mind-to-world direction of fit, because when the representational content of a belief fails to accurately represent ('fit') the world, the thing to do is to change the belief: belief is meant to track the world. And desires have world-to-mind direction of fit, because when the representational content of a desire fails to fit the

[55] Note that I do not mean to deny that claims about meaning are normative, just that normative claims which use normative terms do not have the same compliance conditions as claims about the meanings of those terms.

world, the thing to do, *ceteris paribus*, is to change the world: desire suggests changing the world so that the world tracks desire. And now it appears that a mental state which has a correctness condition must also have mind-to-world direction of fit and thus be a belief, and that a mental state which has a compliance condition must also have world-to-mind direction of fit and thus be a desire. For a state has a correctness condition only insofar as it can be correct or incorrect (and thus subject to correction) in relation to something else, which is just what is involved in mind-to-world direction of fit. And a state has a compliance condition only insofar as something else can comply or fail to comply with it (and thus be subject to being made to comply) which is just what is involved in world-to-mind direction of fit.[56] Yet famously Hume held that beliefs and desires are distinct

[56] I would like to say that all of this is completely uncontroversial, but in fact the close link that I insist on between normative claims and normative judgements has recently been questioned. Kalderon (2005) tries to defend a form of descriptivist non-cognitivism, according to which moral judgements have (only) world-to-mind direction of fit, but moral claims have mind(word)-to-world direction of fit. On this view the claim that torture is wrong has a correctness condition which the judgement that torture is wrong lacks. This does not seem plausible, because it means that there is a difference between judging that the claim that torture is wrong is correct, and judging that torture is wrong. Indeed, it seems that Kalderon's official position is to withhold judgement on the correctness of all moral claims, whilst still making moral judgements. In other words, Kalderon holds that moral claims do not express moral judgements. He concedes that there is some relation between the two: that moral claims **convey** moral judgements (so that we can correctly infer that the person who claims that torture is wrong also judges that torture is wrong, if she is sincere). But this still leaves the problem that competent speakers (who are not opinionated philosophers) take moral judgements and claims to be equivalent: they do not recognise grounds for rejecting moral claims that are not also grounds for abandoning the corresponding moral judgements. In light of this, if meaning is determined by use then since moral claims are used exclusively to express moral judgements that must constitute their meaning. But even if meaning is not directly determined by use it is mysterious how anyone could learn what moral claims mean (if that meaning is, as Kalderon claims, something other than expressing moral judgements), because the only evidence for meaning is use, and so knowledge of meaning must be constrained by use. Since I find Kalderon's position of dubious coherence I shall not discuss it further, and I will thus henceforth take it for granted that expressivism and non-cognitivism are equivalent. But there is one final point to mention concerning the motivation of Kalderon's view. As I understand it, his non-expressivist non-cognitivism is intended as a response to the Frege-Geach problem (which I discuss in detail in Chapter 2); Kalderon's thought is that the Frege-Geach problem refutes expressivism, but not non-cognitivism. But this also seems mistaken, because the Frege-Geach worry is about the validity of inferences, and that applies just as much to inferences between judgements as inferences between claims. If the validity of the move from the claim that lying is wrong and the claim that if lying is wrong then getting little brother to lie is wrong to the claim that getting little brother to lie is wrong depends on those claims having 'robust' truth conditions in a way incompatible with expressivism (as Kalderon presumably believes), then the validity of the corresponding inference amongst moral judgements must also depend on those claims having such truth conditions. Thus it is hard to see how non-cognitivism, but not expressivism, can be defended against the Frege-Geach objection.

existences – that nothing is both a belief and a desire (a 'besire' in the terminology of Altham 1986: 284, see also Smith 1994: 119) – and the general acceptance of this view makes the foregoing platitudes about normative judgement seem like an antinomy.

Normative judgements seem like besires, but there are good reasons to think that there are no besires. The problem with besires is that if there were a mental state which apparently had both directions of fit, it could hardly have both with respect to the same content. That is to say that it could not be both a belief that P and a desire that P, because that would leave it unclear whether it should disappear or not when it is not the case that P. So a besire would have to be both a belief that P and a desire that Q, where P and Q are distinct. Now the problem is that it is unclear why this counts as a unified state: what is the glue that holds the belief that P and the desire that Q together?[57] Note that it is no good just to say that the instantiation of a normative judgement consists in the co-instantiation of a belief (with correctness conditions) and a desire (with compliance conditions). Intuitively, the very **same** state that is correct iff torture is wrong also has a compliance condition (involving the absence of torture, or some kind of action with respect to torture). The idea that the state with the correctness condition can in any way be separated from the state with the compliance condition is misguided. This is, I think, enough reason to dismiss out of hand externalist versions of cognitivism, which hold that moral judgements are simply beliefs with no necessary link to desires (though I say more about externalist arguments below). And versions of internalism which take the necessary connection between the belief and the desire to be casual – a psychological or physiological mechanism of some kind – also look to be ruled out. The link between the correctness condition and the compliance condition is conceptual: a state with that correctness

[57] This is Altham's (1986) main objection to besires.

condition just has to have the corresponding compliance condition. This makes it look as though normative judgements have to be besires in a strict sense: unitary mental states with properties of both belief and desire. But would we really want to incorporate besires into any psychological theory? It seems unlikely that the postulation of this additional kind of state will do any better at predicting behaviour than a theory which takes normative judgements to be desires and explains behaviour by talking only of desires and non-normative beliefs.

Whilst normative judgements seem to be besires, there is little enthusiasm for actually postulating such besires as a genuine psychological category.[58] The reason that the issue over cognitivism and non-cognitivism has proved so hard to resolve is that intuitively moral judgements have **both** directions of fit. Given that a Humean folk psychology of beliefs and desires seems worth hanging on to, we have to decide which of the two conditions – correctness or compliance – is primary. In fact, it appears that we have to do more than this: we have to take whichever condition we take to be primary and explain away the appearance of the other condition. This is the strategy that an early emotivist like Ayer (1946) would have to adopt: claim that normative judgements are simply desires, and that the appearance that moral claims and judgements can be correct or incorrect is misleading. That emotivist line is unattractive, however, because it runs counter to our understanding of what we are doing when we make normative judgements, particularly moral judgements. It matters to us that our moral judgements are correct; to adopt emotivism is to apparently concede that moral judgements are arbitrary. And it seems that a thoroughgoing Moorean realism (which I have shown above is the only

[58] McDowell (e.g. 1979: 346) does endorse besires, which are part of the non-reductive realist package. Since his ultimate reason for abandoning the Humean Theory of Motivation stems from believing in shapelessness, I consider that my discussion of that issue in §1.4 suffices.

remotely tenable form of realism) would have to go in the opposite direction: simply deny that moral claims have compliance conditions. The problem with that would be that this seems to leave out normativity: the presence of compliance conditions is what distinguishes normative claims from other kinds of claims.

Mackie (1977) supplied Moorean realists with a different alternative: locate the prescriptivity of morality in the world itself, and use that to explain the prescriptivity of moral judgements without denying that they are simply beliefs. For Mackie, the only way that moral realism can live up to the relevant platitudes is by postulating facts, ways that the world is, which themselves have compliance conditions. Of course, Mackie thought this idea so incomprehensible that he considered it to be a decisive objection to moral realism (the objection that moral facts are too 'queer' to exist).[59] To get a handle on this queerness, we can compare the prescriptivity of moral facts to the prescriptivity of commands and desires. The fact that somebody issues a command, or has a desire, introduces something with a compliance condition into the world. But the judgement that such a command has been made does not itself have that compliance condition; rather, it is a judgement that something with that compliance condition exists. Thus it would be misleading to say that the fact that a command has been issued itself has a compliance condition: it is the command that has a compliance condition, not the fact. But the case of moral facts (on the picture Mackie considers) would have to be different: the moral facts themselves would have compliance conditions, so that the judgement that the facts were thus and so would itself have compliance conditions. Now Mackie is surely right that we have no idea what it would be like for facts to have compliance conditions, and this makes

[59] Mackie perhaps have thought that additional things followed from prescriptivity, for example that moral properties were such as to cause a kind of attraction or revulsion in those who observed them. Since it is not here important what exactly Mackie meant, I simply note that the version in the text is the interpretation of Mackie which I consider most charitable.

the claim that they do unhelpful as an explanation of why moral judgements have compliance conditions.

That is why Moorean realism is useless as an explanation of our moral practice (i.e. as a solution to what Jackson (1998) calls the 'location problem' for ethics, and as an answer to what Korsgaard (1996) calls the 'explanatory question'). But this does not mean that we have no idea at all what it means for normative facts themselves to be prescriptive; rather, any account of the prescriptivity of normative judgement will serve to explain the prescriptivity of normative facts. What is wrong with Moorean realism is that it gets the order of explanation the wrong way round, and this leads to the thought that the problem requires a metaphysical (non-naturalist) solution. But this immediately creates a puzzle, and hence a clue about what a solution should look like. It is a natural thought that judgements should reflect in some way the properties of the facts which they are judgements of. So the direction of explanation which Mackie considers looks like the right one (except that it engenders queerness). Going the other way looks impossible: the way that the facts are cannot be influenced or determined or explained by the way that judgements of those facts are, because the facts are prior to the judgements. Unless, that is, normative facts are **not** in every way prior to normative judgements. Insofar as we are involved in the explanatory project, we have to see moral facts (and moral properties) as projections or constructions of normative judgements, as Hume and Kant did.

1.6 QUASI-REALISM

We can explain what it could mean to say that moral facts are projections or constructions through the expressivist solution to the antinomy. Stevenson and Hare both held, against Ayer, that moral claims could be true or false. So they were not interested in

resolving the antinomy by denying the platitudes on which it is based. Nor were they prepared to accept a besire-based solution. So if they rejected the obvious ways of dealing with the antinomy, what response to it could they have? The answer begins with their implicit minimalism about truth (or at least about ethical truth): being committed to the truth of an ethical claim is not to be any more committed than one is by accepting that ethical claim. The details of this minimalism are deferred until §2.2, where they can be more appropriately be considered in the light of the Frege-Geach Problem – ultimately it is the solution to this problem which allows the expressivist to earn the right to talk of truth in ethics. So I will assume at this point that the expressivist has no difficulty with truth, and see what follows. Even if there is nothing contradictory in the idea that desires can be true, the problem remains that given that normative judgements can be true they also seem to have to **aim** at truth, which makes them too much like beliefs. Desires which can be true, as expressivism requires, look like desires that are also beliefs, making it seems that expressivists are committed to besires.

The key to Blackburn's quasi-realist response to the antinomy is the insight that the desire-like and belief-like aspects of normative judgements make their presence felt in different ways. Alien psychologists, sociologists or anthropologists, might investigate the role which normative judgement plays in human life without ever finding reason to think that those judgements were truth-apt. They would have no problem in identifying compliance conditions for normative judgements, but although they would notice that humans took normative judgements to have truth conditions they would be tempted to think this a mistake. From that perspective Ayer's emotivism would seem quite appealing. And thus it starts to appear that normative properties are projections, that moral rightness is something with which we produce by 'gilding and staining' the world, rather than part of

the way the world is independently of us.[60] But if we take the metaphor of projection strictly, it involves two viewpoints. What it suggests is that we are both doing the projecting and seeing the projection; yet it is a basic part of the phenomenology of normativity, particularly morality, that it does not seem like a projection. One way of seeing this involves the idea of mind-independence: we do not believe that what is morally right always depends on what we think (either individually or collectively) is right. And yet it seems that if morality is a projection it must be mind-dependent, which would make the projection a kind of *trompe l'œil*. What the metaphor of projection suggests then is that the phenomenology of normative judgement is an illusion, and thus projection seems like a debunking of normativity. To understand projection in a different, less revisionary way, we have to engage more closely with the perspective from which we make normative judgements to show that what it is subject to is not properly thought of as illusion.

Whilst the project of understanding normative judgement as a facet of human behaviour, and thus as part of the natural world susceptible to causation and explanation, is indeed one which we are engaged in, it is not the only relevant perspective on normative judgement. For we are not merely passive observers: we make normative judgements ourselves. One of the key planks in Blackburn's quasi-realist programme is to question the natural assumption that projectivism about normative properties entails their mind-dependence.[61] What is it, Blackburn asks, to say e.g. that torture would be wrong even if nobody thought that it was wrong? The answer is that this counterfactual is itself a normative claim. It clarifies the compliance conditions of the claim that torture is wrong: we can only comply with the moral prohibition on torture by not torturing even in

[60] Hume (1975/1777: 294/246). Hume seems to commit himself to the view that moral properties are not really there, and this in particular is against the spirit of quasi-realism, though closer to that spirit than Ayer.
[61] See e.g. Blackburn 1993a: 153, 181-2 and 1998a: 311-2.

situations where everyone tolerates torture. To put it another way, the claim is that people's moral opinions are not relevant to the moral status of torture. Now given expressivism, according to which all normative claims express attitudes, this counterfactual claim also expresses an attitude. So since the claim that normative properties are projections does not express an attitude, but rather a straightforward belief, it cannot conflict with mind-independence.[62] This removes the temptation of thinking that accepting projectivism would necessarily mean accepting that morality was some kind of illusion or otherwise not all that it seemed to be. That is not to say that projectivism creates no sceptical worries. As I argue in Chapter 3 there is a serious question (which I attempt to answer) concerning how any norms can be justified given projectivism; but that is a not the same problem as that of mind-dependence.

Although the sceptical worry stands defeated, we are left with a mystery: how can normative facts be projections and yet still mind-independent? This is really another way of putting the earlier question lying at the heart of the antinomy: how can normative judgements be both desire-like and belief-like? When Blackburn discusses the mind-independence issue, he emphasises that the solution is to understand that we must give the relevant counterfactuals an **internal** (to morality) reading: they are themselves moral claims, just concerning counterfactual rather than actual circumstances (and this is hardly

[62] Jenkins (2005: 207-8) has argued that although expressivism is not committed to normative facts depending modally (i.e. counterfactually) on the mental, that still leaves room for another objectionable kind of mind-dependence. That is what she labels 'essential' dependence: the suggestion is that for expressivists the essence of normative facts, what it **is** (rather than what it **takes**) for a normative fact to obtain, is something mental. But this suggestion can be ruled out straightforwardly, for essential dependence entails modal dependence. For instance, if being water essentially depends on being H_2O, then it is not possible for anything to be water without being H_2O, so being water also modally depends on being H_2O. So if Blackburn's argument shows that expressivists can deny that moral facts modally depend on mental facts, then it also shows that they can deny that there is any essential dependence either. The real mistake that Jenkins is making here is to assume that the expressivist denies that normative properties can simply be identical to moral properties, but as discussed above we should not understand what distinguishes the quasi-realist from the realist in such metaphysical terms (but rather in meta-semantic terms). The quasi-realist therefore can just say that if it turns out e.g. that utilitarianism is correct, then what it **is** for an action to be right is that it maximises utility etc.

unusual – normative moral theories concern themselves with all circumstances, actual or counterfactual). We distinguish between the questions we ask about the psychology of morality – how it fits into the human sciences – which are external to morality itself, with those questions which are themselves moral.[63] The mistake that critics of expressivism make is in thinking that the expressivist answers to the external questions commit them to particular answers to internal questions, without explaining why this should be so. Indeed, the assumption that the expressivist must deny that moral properties are natural properties is another example of this: the question of what it is for an action to be right is really just a question about which moral theory is correct, which is obviously an internal question.

Blackburn's use of the internal/external distinction (and Dworkin's even more so) sometimes suggests that it applies in a mutually exclusive way to questions (and claims), so as to partition them into disjoint sets: e.g. mind independence questions are moral counterfactuals, so they are internal and not external. But when we come to consider moral judgements, it seems as though they can figure in both internal and external inquiry. If I ask whether a particular moral judgement is correct, I am asking an internal question. If I ask how somebody will behave, given their moral judgements, I am asking an external question. The former question treats moral judgements as beliefs; the latter (at least according to expressivism) treats them as conative attitudes. So as well as talking about internal and external **questions**, it seems that we have to talk about internal and external

[63] For more on the internal/external distinction, see Dworkin 1996 and Blackburn 1998b. Presumably the distinction between internal and external questions is inherited from Carnap (1956), who tries to show that various apparently ontological questions, e.g. "Are there numbers?" are really internal to the relevant discourse (e.g. mathematical discourse). Note that although Dworkin correctly judges that the apparently metaphysical questions we can ask about morality are internal, and so that morality objectivity requires an internal rather than external justification, he goes too far in seeming to deny that there are any legitimate external, meta-ethical questions to ask. Since he has really nothing to say about moral psychology, he fails to engage with or provide an alternative to expressivism.

perspectives or **standpoints**.[64] From the internal standpoint we are engaged in moral reasoning and trying to reach moral verdicts on that basis, meaning that our moral judgements appear belief-like in their direction of fit: we are trying to get them right. From the external standpoint we are trying to explain and predict the behaviour of others (and ourselves at times other than the present), and here it is the desire-like direction of fit that is relevant in connecting up moral judgements to our behaviour.

The upshot is that quasi-realists do not see moral judgements as besires in the original straightforward sense, in that there is no standpoint from which they have both directions of fit at once. The besires discussed by Altham (1986) violate the Humean Theory of Motivation[65] because they are effectively beliefs and desires somehow glued together, and so not distinct existences. In contrast, the quasi-realist is able to hang on to the Humean Theory of Motivation, because from each standpoint beliefs and desires are distinct existences; the two directions of fit as different aspects of a single, unified state. All this allows the quasi-realist account to have the advantages of besires without the associated costs.[66]

---

[64] Blackburn (1998b) uses both terms. It is very tempting to trace the idea of two standpoints back to Kant. Although interpreting Kant's entire Critical system in these terms is a minority view (though see Allison 2004), Kant does explicitly rely on the idea of two standpoints in Section III of the *Groundwork*, in explaining how freedom is possible even though the phenomenal world is causally closed. Kant's idea, roughly, is that we can take two different standpoints towards an action: the practical (i.e. internal) standpoint, from which it is treated as something to be justified (normative reasons), and the empirical (i.e. external) standpoint, from which it is treated as something to be causally explained (explanatory or motivating reasons). Since when considering normative reasons a variety of options must be considered as open (we deliberate "under the idea of freedom"), it makes sense to say that our actions are free from the practical standpoint and causally determined from the empirical standpoint, without contradiction. The strategy that I pursue in the text below is obviously indebted to Kant, and perhaps my reading of Blackburn's quasi-realism owes more to Kant than was really intended. It is also unclear whether Blackburn is consciously echoing Kant in his use of the two standpoints, or whether Carnap is transmitting Kant's influence; though note that Carnap would have been thinking of internal as corresponding to phenomenal/empirical, and external as corresponding to noumenal. The correct interpretation of Kant or indeed Blackburn is not paramount here.

[65] For more explanation of the Humean Theory of Motivation, see Smith (1987 and 1994: 92-125).

[66] There is an alternative to expressivism which resolves the Antinomy and likewise does not resort to besires in any objectionable way: the 'Desire as Belief' (DAB) view (see Lewis 1988, 1996, Oddie 2005, Bradley & List 2009 and Gregory ms for discussion). On this view desires are reduced to evaluative beliefs, so that there is no problem with saying that moral judgements have both directions of fit. The difference between DAB and

An obvious question to ask about this two standpoints view is: which standpoint gets us closer to how things really are, and so are moral judgements **really** more like beliefs or desires? Now the quasi-realist project is in danger of collapse if we accept that a straightforward answer can be given here; saying that moral judgements are **really** desires (which is the more obvious option for expressivists) because the external standpoint is primary, looks like admitting that the internal, moral standpoint involves some kind of illusion.

An alternative answer, however, is that all that could be involved in the internal standpoint being non-illusory, and so just as valid as the external standpoint, is that things are as it represents them as being. But all that the internal standpoint tells us is that we are bound by moral requirements, and this is something that can only be evaluated from the internal standpoint itself. There is no way of stepping outside the internal standpoint so as to throw doubt on its verdicts, because the external standpoint has nothing to say about internal questions. So the quasi-realist should simply insist that the internal and external standpoints do not stand in any kind of hierarchy. On the other hand, it may be that questions about reality are disguised forms of the location problem (Jackson 1998), and so concerned with how ethics fits into the scientific worldview. Since such explanatory questions are clearly external ones, we can straightforwardly respond that science will view moral judgements as desires, and so if we insist on the priority of science that verdict inherits such priority.

---

expressivism is that instead of seeing only a sub-class of desires as being identical to evaluative beliefs, DAB analyses all desires in this way. Doing so creates technical problems (as discussed by Lewis) and seems vulnerable to counterexample (when I have a craving, that is a desire, but it is not a belief that what I crave is valuable in any way or that I have reason to pursue it). But the main problem with DAB compared to expressivism is that DAB analyses something that we already have a theoretical grip on – desire – in terms of something mysterious – evaluative belief – and therefore fails to be illuminating.

What this chapter aims to have shown is that expressivism is best placed to solve this location problem, explaining moral judgements in a naturalistic way whilst at the same time retaining its special practical character (exhibited through the Euthyphro, the OQA, and the plausibility of judgement internalism), which realism is unable to achieve. What quasi-realism adds to this, apart from helping to save many of the appearances of moral thought and discourse, is to show that there is no inconsistency in marrying the expressivist explanation of morality from the external standpoint with an internal view of morality that is not infected by any scepticism or relativism stemming from expressivism. This leaves two hostages to fortune, however. First there is the task of showing that it is possible for the quasi-realist project of explaining how a truth-apt moral discourse can be built on expressivist foundations; that is to show how there can be an internal standpoint, which presupposes that the right to talk of truth has been earned. This task is taken up in Chapter 2. Secondly it is important to understand that although quasi-realism shows that there is no obvious conflict between expressivism and moral objectivity, it is still possible that there are internal grounds for scepticism about moral objectivity, meaning that it has not been established that moral objectivity is really possible on an expressivist (or any other) foundation. The task of confronting this real challenge of moral objectivity and securing genuine moral justification is pursued in Chapter 3.[67]

The problem with the internalism vs. externalism debate over moral judgements as a way of resolving the cognitivism vs. non-cognitivism debate is that it relies on intuitions which are really intuitions about which of the two contents are primary; and that issue should not be settled by direct appeal to intuition. Rather we should ask ourselves about

[67] This is Korsgaard's (2003) view: expressivism is the correct answer to the explanatory question (of locating moral judgement within a scientific world-view), but it does not in itself provide an answer to the normative question (of showing that some moral norms are actually valid for us).

the consequences of taking one or the other content to be primary. And when we do this the importance of the arguments about realism above becomes clear. If we say that the 'Torture is wrong' kind of content is primary, then we are committed to explaining the nature of moral judgements via an understanding of that content. But such an independent understanding of moralised contents could only be forthcoming if realism were true. The only form of (alleged) realism left standing – Moorean realism – survived only by foreswearing any positive account of moral properties and facts, and cannot thus play the role which realism needs to in a defence of cognitivism. The defeat of realism means that cognitivism does not function as an explanation of moral judgement, and is thus disqualified (of course, this disqualification would only be final if expressivism/non-cognitivism could provide an explanation, but we shall see that it can).

CHAPTER 2: THE FREGE-GEACH PROBLEM

2.0 INTRODUCTION

Peter Geach (1960, 1965) thought that, borrowing from Frege, he had a decisive objection to expressivism: embedded contexts.[68] The allegation is that expressivists cannot explain the validity of moral *modus ponens* and other simple arguments.[69] But after much wrangling it is unclear what the force of this objection really is, and whether it has been answered. In this chapter I argue that Simon Blackburn's (1984, 1988) various quasi-realist replies to Geach have been largely successful, and that a combination of Blackburn's commitment semantics, suitably augmented, and a two-concept view of truth and belief can answer explain how a logic of the attitudes is possible. In §2.1 below, I explain the problem, and why it poses a challenge to expressivists despite Geach's misleading characterisation of the force of his point. In §2.2 I explain why quasi-realists should accept a two-concept view of truth and belief, and discuss how far this goes towards addressing Geach's concerns. In §2.3 I show how Blackburn's approach needs to be extended to avoid objections. In §2.4 I explore a different kind of objection to commitment semantics, and explain both how to answer the objection and why it shows that the quasi-realist needs both the two-concept view and commitment semantics in order to solve the Frege-Geach Problem. In §2.5 I explain and reply to an objection from Mark Schroeder (2008) which aims to show that even if there can be a logic of the attitudes the expressivist lacks the resources to account for the logic of moral terms.

[68] John Searle (1962) independently arrived at roughly the same objection. Hare (1970) himself gave a response to Geach and Searle which anticipated many of the later moves in the literature, and is still convincing in large part. I will not, however, attempt the task of giving a detailed reconstruction of how Hare's own ideas relate to the literature or to my own views about the Frege-Geach Problem; that would introduce unnecessary complications and distractions (such as Frege exegesis).

[69] Following the literature I will talk as if the Frege-Geach problem applies specifically to expressivist accounts of moral thought and discourse. But it is normativity more broadly that is at stake here, so almost anything said here about e.g. moral claims will apply equally to all normative claims.

2.1 THE CHARGE OF EQUIVOCATION

A moral *modus ponens* argument looks like this:

(1) If X is wrong then Y is wrong.

(2) X is wrong.

---

(3) Y is wrong.

Why are expressivists subject to the charge that on their view such arguments equivocate? Expressivists think that moral utterances express attitudes. Supposing that (2) expresses an attitude, that attitude is not expressed by the antecedent of (1). So the string 'X is wrong' is not used in the same way in (1) and (2), and thus there is an equivocation.

Thus stated, however, the anti-expressivist argument is far from compelling. For the alternative view to expressivism is that moral utterances express beliefs, not attitudes.[70] And supposing that (2) expresses a belief, that belief is not expressed by the antecedent of (1). If expressivism tells us that moral *modus ponens* equivocates, then so apparently does the alternative. Since it is obvious that (2) expresses either an attitude or a belief, any argument to the effect that it cannot is obviously incorrect, though it may not be so obvious just what is wrong with the argument.

Suppose we try to explain why Geach's objection is not fatal to the view that moral utterances express beliefs. One obvious move is this: moral *modus ponens* is valid because when the premises are true the conclusion must be true as well. So the alleged equivocation cannot be present. The problem is that this response is equally available to

[70] Note that the term 'expressivism' is thus potentially confusing. It is hardly controversial that a sentence such as 'Lying is wrong' is conventionally used to express *some* mental state; expressivists disagree with their opponents about what kind of state is expressed, not whether any state at all is expressed. And expressivists also hold that the meaning of moral claims is to be ultimately explained in terms of the mental states they express (their sincerity conditions) rather than their truth-conditions. The reason such a misleading term has becomes standard is that the term it replaced – 'non-cognitivism' – was even more misleading.

the expressivist. Expressivists agree that (1) and (2) can be true, and that their truth entails that (3) is true as well. The expressivist's opponent (henceforth 'descriptivist') perhaps relies on a suppressed premise: any truth-apt sentence is standardly used to express the belief that that sentence is true (or perhaps that the proposition which that sentence represents is true). This premise would do what the descriptivist wants, but it is not accepted by (most) expressivists. They claim that moral sentences are truth-apt, but primarily express attitudes.[71] This claim is of course contentious, and it may well appear incoherent. I will discuss the expressivist view of truth in more detail in §2.2 below. Let us note for now that the problem concerning truth is not the Frege-Geach Problem, though it may be the kind of problem that some descriptivists have in mind when they urge that the Frege-Geach Problem is insoluble. Expressivists are obliged to say something about truth for quite independent reasons, for if they deny that moral sentences are truth apt, they are committed to saying that, for instance, it is not true that murder is wrong. Though the Frege-Geach Problem is thus meant to be distinct from the problem of truth, they are not unconnected, because the expressivist will have to give a unified account that can cater for both problems – hence the relevance of discussing it here. Moreover we will get a better idea of what the real challenge to expressivism is by seeing what the expressivist's account of truth in moral discourse leaves out.

For now, however, we want to know whether there is indeed a distinct Frege-Geach Problem. So is there any way for the decriptivist to explain, without mentioning truth, why there is a problem for the view that moral sentences express attitudes that

[71] It is not going too far from standard usage to say that expressivists who accept that moral utterances are truth-apt are quasi-realists; that is how I will mean 'quasi-realism' henceforth, and even if it is a simplification the adjustments required to make the treatment here more precise will be merely terminological. For various reasons, some of which are discussed above, quasi-realism in this sense is the only plausible form of expressivism for ethics, and perhaps in general.

does not also defeat the view that moral sentences express beliefs? The argument perhaps goes like this: expressivists think that the meaning of moral sentences lies in the attitudes which they express, whereas descriptivists think that their meaning can be given quite independently of the beliefs which they express, in terms of truth-conditions for example. If that were the case, then the descriptivist would have a way of characterising meaning that did not seem to introduce an equivocation between (1) and (2), whereas the expressivist's only way of characterising meaning would introduce that equivocation. Unfortunately, this does not seem to rule out an extension of the previous response: if we grant that moral sentences are truth-apt, then it will be possible to characterise their meanings in terms of truth-conditions. For instance, if we think that there is some connection between

(T) *S* is true iff *p*.

and

(M) *S* means that *p*.

then we will have no problem applying this to moral sentences. So expressivists should not say that the only way of giving the meaning of a moral sentence is to say what attitude it expresses. One can also give the meaning by presenting another sentence with the same truth-condition (and with whatever other features the semanticist requires for sameness of meaning). Expressivists are committed to truth-conditions for moral sentences by being committed to their truth aptness: once it is admitted that 'Murder is wrong' can be true, it is clearly true iff murder is wrong.

Perhaps it will seem that the expressivist is now conceding far too much. For going in this Davidsonian direction will force the expressivist to admit that there are moral properties, and this is just to admit that the descriptivist was right after all. But this is to

think that what marks out expressivism is the denial of realist metaphysics, which (as discussed in Chapter 1) is a mistake: the expressivist can happily take on board the (naturalist) realist view of moral properties, and so happily agree that a full truth-conditional semantics is possible. What the expressivist denies is that this kind of extensional account of the meaning of moral claims gives any real explanation of the distinctiveness of morality. The expressivist is not denying the realist metaphysics, or even the realist semantics, but rather insisting on further explanation (and so disputing various kinds of realist meta-semantics, e.g. Cornell Realism, which preclude such explanation).

So it looks as though expressivists could simply respond to the Frege-Geach Problem as follows. There is no equivocation between (1) and (2) because the meaning of (1) is a function of the meaning of (2) in whatever way the correct truth-conditional theory of meaning says. Whatever resources the descriptivist has for explaining why there is no equivocation can be appropriated by the expressivist. Not of course that it is obvious that such appropriation is legitimate. Perhaps there is some descriptivist argument to show why it is impossible, or why it leads to collapse (though I doubt it). What is clear is that Geach's charge of equivocation is little more than a clue to where the real action is. On the face of it, there is nothing about expressivism that makes the charge of equivocation especially pressing, and no real burden on expressivists to apply themselves to any problem raised by Geach.

And indeed more recent accounts of the Frege-Geach Problem do not emphasise the charge of equivocation, recognising that doing so places the burden of proof squarely with the descriptivist. The more up-to-date claim is that the expressivist owes us an account of the meaning of (1) and of the mental state which (1) expresses. But putting things that way hardly amounts to a compelling problem for expressivists, unless some

reason is giving for thinking that expressivism makes it harder to explain (1). Otherwise the expressivist could turn the problem around: until the descriptivist gives a compelling account of the meaning of (1), we should assume that descriptivism is false. And this would be a serious matter for descriptivism, since in fact there is no consensus on the meaning of (1) in the descriptivist camp; all the standard proposals for the meaning of the indicative conditional are beset with well-known objections. Yet of course it would be silly to take this line: the lack of an agreed descriptivist account of the indicative conditional is no reason to doubt that at least some sentences figuring in *modus ponens* arguments express beliefs. But then why exactly should the lack of an agreed expressivist account of the indicative conditional be a reason to doubt that at least some sentences figuring in *modus ponens* arguments express attitudes?

The descriptivist attack can be improved by emphasising that the expressivist can only have access to truth-conditional theories of validity and meaning at the internal level of the discourse. Suppose that the quasi-realist story is correct, and claims about the truth and falsity of ethical utterances get interpreted as ethical claims themselves. This will allow for a truth-conditional semantics and so enforce the right logical relationships, but without explaining why that logic is compulsory. As we shall see, we can distinguish syntactic constraints that allow for (minimal) truth-aptness. Once we have done this, the descriptivist should admit that the brute fact that ethical discourse satisfies those syntactic constraints is sufficient to make *modus ponens* and other such argument-forms valid. But the descriptivist can ask what it is about ethical discourse that makes such a syntactic regimentation suitable. The idea is that in order for a discourse to have a right to syntactic

discipline, it must have a corresponding semantic discipline.[72] And the descriptivist should say that the only kind of semantic discipline that can ground syntactic discipline is maximal truth-aptness (e.g. a correspondence theory of truth). This is what expressivists, quasi-realists included, must deny: their claim is that the view of ethical properties as potentially being satisfied by actions and states of affairs arises out of the syntactic discipline and minimal truth-aptness of ethical discourse. Thus expressivists cannot say that the syntactic discipline of ethical discourse is grounded in a semantics based on objects satisfying properties; for them the direction of explanation is reversed.[73]

So expressivists *do* have a burden of explanation that descriptivists do not share. They must give a semantics for moral discourse that is not based on truth, and which can explain its right to syntactic discipline without being grounded in truth. Note that here there is no symmetry between expressivists and descriptivists; it is not the case that descriptivists' struggle to give a truth-conditional semantics (for example) for all discourse is equivalent to the expressivists' struggle to give the semantics we are discussing here. For expressivists are committed to there being a semantics for all the discourses which they are not expressivist about (and there had better be some such discourses, for global expressivism or quasi-realism is not an attractive doctrine). Thus expressivists share the theoretical burden of doing general semantics, and have an additional burden as well: they are committed to there being (at least) two distinct kinds of semantics that are capable of

[72] By 'a right to syntactic discipline' I mean roughly the same as Blackburn when he speaks of 'earning the right to talk of truth'. That is, the general grammaticality of embedding for a discourse must be explained by there being a way of giving a meaning to those embeddings that does not circularly rely on the truth-aptitude which the grammaticality of the embeddings itself grounds.

[73] Again, this is not at all to deny that there are (really!) moral properties, or that something like a Tarskian theory of moral truth could be given in terms of them. The expressivist agrees with the realist about the availability of all this semantic machinery, but disagrees that by using it we explain the meanings of moral claims; the extensional characterisation of meaning is in this case too shallow.

grounding syntactic discipline.[74] But all this is not to say that Geach was right in his original charge of equivocation. For suppose that the expressivist project of finding a new semantics for ethical discourse fails. Then it is not that expressivists will have assigned equivocating meanings to (1) and (2); rather, they will have failed to explain the meaning of (1) altogether. The way Geach sets up the argument actually grants expressivists the most important point at issue, namely that they have a right to the use of ethical conditionals. Once this is granted, there is no further issue of the validity of moral arguments. The challenge for expressivists is to show why if ethical conditionals did not exist they would have to be (or at least could be) invented.

Much of the last few paragraphs, however, stands in need of explanation. In particular, we need to know what is meant by talk of syntactic discipline, and by 'minimal' and 'maximal' truth-aptness. The following section (§2.2) aims to make all this clear by giving a quasi-realist account of truth and truth-aptness. In doing so, I aim to justify my earlier assertions that there is no serious danger of expressivism entailing that moral *modus ponens* is invalid – syntactic discipline guarantees its validity. Rather the problem is how to show how that validity is to be explained without relying on syntactic discipline. That task will have to wait until §2.3.

2.2 THE PROBLEM OF TRUTH

Quasi-realists accept minimalism about truth and truth-aptness. That is, they accept that instances of the homophonic T-schema or Disquotation Schema:

(DS) '*S*' is true iff *S.*

[74] Note though that the two types of semantics might have much in common. For instance, they might share an account of the indicative conditional. So the expressivist may have to do rather less than twice the work of the descriptivist.

(where '*S*' stands for a sentence) are valid just in case they are grammatical, and they accept that '*S*' is truth-apt just so long as it embeds into the consequent position of the biconditional. Thus they accept cases like

(4) 'Lying is wrong' is true iff lying is wrong.

but they reject cases like

(5) 'Do not lie' is true iff do not lie.

simply because (4) is grammatical and (5) is not. So quasi-realists accept that moral sentences are typically truth-apt. They do so since they hold the T-schema to be platitudinous, because they say that no more is involved in asserting e.g. '"Lying is wrong" is true' than in asserting 'Lying is wrong'.

This form of minimalism is approximately the disciplined syntacticism of Crispin Wright (1992).[75] The 'syntacticism' indicates that truth-aptness is held to depend on whether the relevant instance of the T-schema is syntactically well-formed (i.e. grammatical). The 'disciplined' refers to the way that syntax is connected to behaviour in embeddings: '"Lying is wrong" is true' is grammatical because it makes sense to embed 'Lying is wrong' into conditionals. But as we shall see, quasi-realists must disagree with strict minimalists on a crucial point: whilst minimalism holds that the concept of truth is characterised by syntactic discipline, quasi-realists should say that *one* concept of truth is characterised in that way, but that there is another concept of truth, for which truth-aptness is harder to come by.

But first a brief word on why expressivists should be quasi-realists. I mentioned earlier the straightforward point that it is simply advantageous to be able to say that some

[75] It is decidedly not the same as Paul Horwich's (2001) version of minimalism. Neither Horwich nor Wright has an established right to the term 'minimalism', so I will use the term to refer to Wright's views without further qualification.

moral claims are true, and that some moral arguments are valid in the sense that the truth of the premises is incompatible with the falsity of the conclusion. Insofar as expressivists deny these claims, they are in conflict with common sense. So unless quasi-realism is somehow incoherent, expressivists should adopt it; quasi-realism should be the first resort for expressivists, and a more hard-line expressivism a second resort. It's worth noting one thing that expressivists cannot say. They cannot follow the suggestion of Jackson, Oppy & Smith (1994: 289) to accept claims like (4) without accepting them as true. The reason is that to accept a biconditional is to commit oneself to accepting one side whenever one accepts the other. Since expressivists may well accept the right-hand-side of (4) ('Lying is wrong') they thus incur a commitment to accepting the left-hand-side if they accept (4). But the left-hand-side is '"Lying is wrong" is true', and if expressivists are prepared to accept that then they thereby concede that moral claims are truth-apt. So hard-line (i.e. non-quasi-realist) expressivists must reject (4) outright, and that means rejecting grammatical instances of the T-schema.[76]

When the expressivist does take the more attractive quasi-realist route of accepting (most of) minimalism, a problem soon emerges. The quasi-realist wants to deny that moral judgements are beliefs. But conceding that those judgements are truth-apt is tantamount to conceding that they are beliefs. Here then is a simple argument against quasi-realism:

(6) Moral sentences are truth-apt.

(7) If moral sentences are truth-apt, then moral judgements are truth-apt.

[76] It follows that Jackson, Oppy and Smith are wrong to think that there is an interesting distinction between minimalism about truth and minimalism about truth-aptness. Minimalism about truth is (or entails) the view that all grammatical instances of the T-schema are to be accepted, and thus it entails by the above argument that all sentences which embed in the way required by the T-schema are truth-apt, which is minimalism about truth-aptness.

(8) If moral judgements are truth-apt, then they aim at truth.

(9) If moral judgements aim at truth, then they are beliefs.

---

(10) Moral judgements are beliefs.

Since quasi-realists accept (6), they apparently have to reject one of (7), (8) and (9). Rejecting (7) seems hopeless, because biconditionals like the following seem uncontroversial whatever one's theory of truth:

(12) 'Lying is wrong' is true iff the judgement that lying is wrong is true.

Similarly, it is very hard to deny (9), because quasi-realists will want to characterise the difference between beliefs and desires in terms of direction of fit.[77] Beliefs aim at word-to-world fit: when a mismatch between our beliefs and the world comes to light, our beliefs change. Desires aim at world-to-word fit: when our desires are unsatisfied, we ordinarily try to satisfy them. But this is all just to say that beliefs aim at truth whereas desires aim at satisfaction. If moral judgements have both truth-conditions and satisfaction conditions (where we mean 'satisfaction' as it relates to desires), these are not the same: what satisfies the judgement that lying is wrong is an absence of lying, whereas what makes it true is lying being wrong. One desperate escape from (9) would be to posit states with both directions of fit – 'besires' – but that is not an option for expressivists, since besires are inconsistent with the Humean Theory of Motivation which expressivists take to be fundamental.

It might seem more promising for quasi-realists to reject (8), since it is not in general true that truth-apt mental states aim at truth. For instance, it is plausible that imaginings are truth-apt but do not aim at truth. But whilst it is not a mistake to imagine what is not the case, it must be a mistake to make a false moral judgement. One cannot

[77] See Anscombe (1957: 56).

deliberately make a false moral judgement, in the same way that one cannot deliberately maintain a false belief. And that is surely enough for it to be clear that moral judgements aim at truth.

The quasi-realist thus has no option but to accept that the argument (6-10) is sound. But now the quasi-realist seems to have conceded too much: once you agree not only that moral utterances are truth apt, but also that moral judgements aim at truth and are beliefs, it is hard to convince anyone that you are an expressivist. But this is the path which the quasi-realist must tread. Since expressivism involves holding that moral judgements are attitudes (understood as desire-like), quasi-realists must hold that moral judgements are both beliefs and attitudes, and somehow say this whilst avoiding espousing besires. The key is to say that whether they are beliefs or attitudes depends on the perspective from which one looks. It is only internally to moral discourse that moral judgements look like beliefs. From the external perspective, those judgements are still attitudes. So we have a result which perhaps should not have come as a surprise: quasi-realism is also quasi-descriptivism. Not only will quasi-realist go along with the various realist-sounding claims that are internal to ethical discourse, such as claims about the mind independence of moral truth, but they will also concede that insofar as descriptivism and makes claims from that internal point of view, those too are correct. Insofar as the quasi-realist is committed to 'saving the appearances' – vindicating our everyday ways of speaking in ethics – talk of 'moral beliefs' will be saved whenever talk of 'moral truth' is.

All this, however, raises a challenge to quasi-realists: how can the distinction between quasi-realism and plain old realism be maintained? It is worth distinguishing this important challenge from a less significant worry. That is that many realists and descriptivists will want to declare victory at this point. It may well be that all they were

arguing for has already been conceded, and there is little point in quibbling over the name of the resultant position – 'quasi-realism' or 'sophisticated realism'.[78] But it is one thing to hold that quasi-realism is best classified as realism, and another to say that no difference at all can be maintained between quasi-realism and realism. That is the real threat: that quasi-realism collapses because the external perspective cannot be maintained.

Quasi-realists can stave off collapse by distinguishing between two concepts of truth, and two concepts of belief: minimal and maximal.[79] Minimal truth-aptness is what syntactic discipline gives us. But maximal truth-aptness is not come by so easily. It requires a substantial connection between the truth-bearer and the world, so that the truth-bearer represents the world in a way susceptible to causal explanation, or so that acceptance of the truth-bearer will make the agent more successful in causally predictable ways. Similarly, minimal beliefs are mental states which aim at minimal truth, whereas maximal beliefs are representational and relevant for success, and aim at maximal truth. Recognising these two concepts of truth and belief allows the quasi-realist to admit the soundness of the (6-10) argument. Each of (7-9) is only true if all occurrences of 'truth' and 'belief' are read as uniformly indicating the minimal concepts, or uniformly indicating the maximal concepts. Thus if (6) is interpreted as involving minimal truth, the version of (10) that can be derived will involve minimal belief. The expressivist only accepts (6) in the

[78] Gibbard (2003) prefers the latter, but this is a terminological rather than a substantive disagreement with Blackburn.

[79] It is important to keep the view defended here, that there are multiple *concepts* of truth, distinct from Wright's (1992) view that there are different truth properties which all realise the same concept. In fact, although the two-concept view is a more radical form of pluralism about truth than Wright's, it is not vulnerable to a crucial objection to Wright's view. The problem for Wright's property pluralism is that it cannot deal with mixed sentences, like 'Her action was wrong, but she thought it was right, because she did not anticipate the consequences'. If one truth property attaches to moral discourse, and another to psychological discourse, for instance, there is no single truth property appropriate to the whole sentence. Concept pluralism has no such difficulty, however, because such mixed sentences are minimally truth-apt in virtue of their syntactic discipline, regardless of whether any of their sub-clauses are maximally truth-apt. It is an interesting point that strong pluralism about truth should be so much more defensible than weak pluralism, and it counts strongly in favour of Blackburn's construal of the realism debate over Wright's.

minimal sense, so only has to admit that moral judgements are minimal beliefs.

It is worth noting that strictly speaking it is unnecessary to have two concepts of both truth and belief in order to dodge the (6-10) argument. The quasi-realist could say that the only concept of belief is the maximal one, or that the only concept of truth is the minimal one. Indeed, Blackburn (1998b and 1998a: 318) himself prefers the latter alternative. On the other hand, James Lenman (2003a) has argued in favour of two concepts of truth, but not two concepts of belief. As a terminological point, it seems best to agree that anyone who accepts either two-concept view can be a quasi-realist; indeed quasi-realism may amount to little more than making such a distinction, and saying that in some discourses only the minimal concept(s) applies. But in my view there is clear reason for favouring the double two-concept view. Put simply, the point is just that the link between truth and belief is so platitudinous: belief aims at truth. If you have a concept of minimal truth with no concept of minimal belief, then there is intuitive pressure to say that those mental states which do aim at minimal truth count as beliefs in some sense. If you have a concept of maximal belief with no concept of maximal belief, there is intuitive pressure to count what maximal belief aims at as truth in some sense. Some reason must be given for resisting such pressure, and I do not see that Blackburn has given any. It seems most likely that Blackburn resists maximal truth because he does not want to admit that there is any sense in which moral claims cannot be true. But he does not need to deny maximal truth in order to avoid this; he can instead say that when we consider the truth of a claim which is only minimally truth-apt, it is only minimal truth which is at issue. So although moral claims are not maximally truth-apt, we must still say that they are true when we endorse them.

The two-concept view is independently plausible once we reflect on the fact that

the whole point of Wright's syntactic discipline approach is that there is only a syntactic constraint on truth-aptness. Yet we can see the possibility of a semantic constraint, based on representation, or success, or the distinctive role of belief in motivating action. And there is no reason to assume that these constraints would be bound to coincide. The quasi-realist proposal is that the semantic constraint is in fact much tighter than the syntactic constraint; and no wonder, since the semantic constraint is given by the world, whereas the syntactic constraint is weakened by our natural eagerness to employ the useful device of a truth predicate as widely as possible. It would have been a fluke if the two constraints had coincided precisely; and any mismatch between the constraints almost demands a two-concept view.

I do not wish here to take a strong stand on exactly which way of explicating the double two-concept view is best, but I will make some suggestions. First, if I am right in thinking that having any commitments about the extensions of moral predicates (i.e. having any moral views at all) commits us to genuine, natural moral properties (see Gibbard 2006 and Chapter 1 above), then it is difficult to explain why moral truth is not maximal truth. If we thought that moral predicates could not be construed as referring to properties then it might be best to start with moral language, like Hare, and found everything on a distinction between descriptive and non-descriptive predicates. Since I am sceptical about this approach, I therefore suspect that a better route is to emphasise that from the external standpoint moral judgements have the functional role of conative states. That is what allows us to say that moral judgements are only beliefs in a minimal sense, and that in turn they are only truth-apt in a minimal sense. I am not confident that this is the only viable approach, however.

One might think that the two-concept view is all that quasi-realists need in order to

solve the Frege-Geach Problem. After all, if a minimal concept of truth applies to moral claims, then arguments like (1-3) will come out as valid. So the two-concept view appears to answer Geach's challenge: it gives an explanation of the validity of the relevant arguments which is compatible with expressivism. More, however, does need to be said by way of explanation. What the quasi-realist needs to do is to explain why syntactic discipline itself is appropriate to moral discourse, without relying on truth.[80] At a minimum, this means giving an account of inferential rules for moral argument that could be justified at the external level, and which will in turn justify the syntactic discipline of moral discourse. In giving such rules the quasi-realist will elucidate the meaning of claims involving moral expressions in unasserted contexts. The attempt to do this occupies the next section (§2.3).

2.3 COMMITMENT SEMANTICS[81]

Simon Blackburn has spearheaded the attempt to show that there can be a 'logic of the attitudes'. In doing so, he has tried our several different approaches (Blackburn 1984, 1988, 1993b), and it would be tedious to survey the details of all such attempts.[82] The basic idea that has emerged of such a logic is the idea of 'commitment', and thus the approach is called 'commitment semantics'. The central point is to interpret the acceptance of a conditional as a commitment to accepting its consequent when one accepts its antecedent: i.e. *modus ponens*. The reason for choosing such an interpretation is that it does not rely on truth, but it does allow for a disciplined syntax that mimics a

[80] This point is made by, for example, Walter Sinnott-Armstrong (2000).
[81] This section is an updated version of material published as Elstein (2007).
[82] Note that Gibbard (2003) asserts that his proposed solution (in terms of plans and hyper-plans), although superficially rather different from commitment semantics, is in fact equivalent to it. For this reason I will not engage in a detailed discussion of Gibbard's view.

truth-conditional semantics.

Blackburn bases his account around the rule that $[\phi \rightarrow \psi], \phi \vdash \psi$. Basic wffs in commitment semantics are of the form A(*p*), meaning acceptance of *p*. Negation can occur both internally, as in A(¬*p*), and externally, as in ¬A(*p*). We can flesh out the details of the account by considering how Blackburn ought to respond to an objection from Jorn Sonderholm (2005). Sonderholm points out that as it stands Blackburn's account is unable to validate certain simple arguments involving disjunction, the most basic of which is the inference:

$(p \vee p) \vdash p$

Blackburn interprets the disjunction p v q as two conditional commitments: if ¬p is accepted, accept q, and if ¬q is accepted, accept p. This is symbolised as [A(¬p) → A(q)] & [A(¬q) → A(p)]. Sonderholm explains that Blackburn cannot deduce an inconsistent set of commitments from accepting the premise of the inference in question whist not accepting its conclusion. An attempted proof will proceed as follows, stopping before an inconsistency is reached:

1	(1) [A(¬p) → A(p)] & [A(¬p) → A(p)]	1 ASS
2	(2) ¬A(p)	2 ASS
1	(3) A(¬p) → A(p)	1 &E
1,2	(4) ¬A(¬p)	2,3 MTT

Sonderholm rightly says that this argument could be completed if we were allowed as a rule $\neg A(\neg A) \vdash A(A)$, and that this rule is unacceptable: refusing to reject p is not the same as accepting p. There is, however, a different rule that would allow us to complete the argument: a reduction rule which says that if one is conditionally committed to absurdity on accepting A, then one is committed to accepting ¬A. We can symbolise this as

[A(A) → A(⊥)] ├ A(¬A). We want to generalise this rule to cases where it is only on certain assumptions that accepting A commits one to accepting absurdity, in which case it is only given those assumptions that one is committed to accepting ¬A. Thus the full reduction rule is:

(R) ϕ, [[ϕ & A(A)] → A(⊥)] ├ A(¬A)[83]

This licenses running the argument as follows, now deducing A(p) directly, rather than showing that the premise is inconsistent with ¬A(p):

1	(1) [A(¬p) → A(p)] & [A(¬p) → A(p)]	1 ASS
1	(2) A(¬p) → A(p)	1 &E
3	(3) A(¬p)	SUPP
1,3	(4) A(p)	2,3 MP
1,3	(5) A(⊥)	3,4 ⊥
1	(6) A(¬¬p)	2-5 R
1	(7) A(p)	6 N[84]

The strategy also works for Sonderholm's less simple case: the inference [(p&q) v (p&r)] ├ p:

1	(1) [A¬(p&q) → A(p&r)] & [A¬(p&r) → A(p&q)][85]	1 ASS
1	(2) A¬(p&q) → A(p&r)	1 &E
3	(3) A(¬p)	SUPP
3	(4) A¬(p&q)	3 J[86]

[83] Here, and in what follows, 'ϕ' is a schematic letter for any wff, whereas 'A' is a schematic letter for any string (not containing any occurrences of the A( ) acceptance operator) such that 'A(A)' is a wff.

[84] N is the rule that A(¬¬A) ├ A(A). We are entitled to assume a commitment to classicism.

[85] Note that there is something dubious about Sonderholm's translation here, because we should be unwilling to allow 'A¬(p&q)' as a wff, since it clearly means the same as 'A(¬p v ¬q)', which Blackburn disallows. It would be better to paraphrase the former in the same way as the latter. In the text I give Sonderholm the benefit of the doubt arguendo in framing his alleged counterexample.

3	(5)	A¬(p&r)	3 J
1,3	(6)	A(p&r)	2,4 MP
1,3	(7)	A(⊥)	5,6 ⊥
1	(8)	A(¬¬p)	2-7 R
1	(9)	A(p)	8 N

It is worth mentioning four possible worries. The first is that rule R and rule K are equivalent; if this were so then my reply would be in no better shape than the one which Sonderholm rightly rejects. But the rules are not equivalent: R is weaker than K. There are circumstances where one does not accept ¬p, but accepting ¬p would not involve absurdity. According to rule K one is committed to accepting p, but rule R does not apply.

The next worry is that I play fast and loose with absurdity: there is a difference between the internal contradiction involved in the combination A(p) and A(¬p), and the external contradiction in A(p) and ¬A(p). I concede the point, and that is why I use the expression 'A(⊥)' rather than '⊥' when internal contradiction is in play (but if my symbolism offends a different symbol could be substituted). The crucial point is that we have a standing commitment to avoiding contradiction, and this is reflected in our use of reductio reasoning. Anyone who accepts reductio reasoning must accept rule R, since what rule R records is simply our commitment to accepting the results of reductio reasoning. Since the expressivist is trying to justify all classically valid inferences, not just intuitionistically valid ones, that commitment can legitimately be assumed. Those points are the ones on which my reply stands, and they are unaffected by quibbles about what 'absurdity' means.

A third worry is that rule R commits us to accepting a contradiction. If we let L be

[86] J is the rule that A(¬A) ⊢ A¬(A&B). This rule eliminable, because both A(¬A) → [A(A) → A(B)] and A(¬A) → [A(B) → A(¬A)] are theorems, and so A(¬A) → {[A(A) → A(B)] & [A(B) → A(¬A)]} is also a theorem. The latter is equivalent to A(¬A) → A¬(A&B), so J is a shorthand, rather than an addition to the system.

the liar sentence ('This sentence is false'), then we get both A(L) → A(⊥) and A(¬L) → A(⊥) as theorems. But then R allows to deduce both A(¬L) and A(L), so it turns out that A(⊥) is a theorem! But it is if anything an advantage for commitment semantics that it takes the Liar to be paradoxical. It is not as if standard reduction rules can cope consistently with the Liar; truth-conditional semantics thus has companions in guilt. So this worry is only worth taking seriously if the expressivist's opponent has her own solution to the Liar, which cannot be adapted to commitment semantics. If commitment semantics has the same problem with the Liar that everyone else does, that counts as a success (albeit an odd one) for expressivists in stealing the clothes of realists.

The most serious worry is that my reply has the advantage of theft over honest toil. After all, the point of Blackburn's project is to explain why we are inclined to accept certain inferences. But I am just taking it for granted that reductio inferences are acceptable. Clearly this would not be acceptable as a general strategy: whenever a valid argument is presented that commitment semantics in its present form cannot deal with, invent a rule licensing the argument. On the other hand, the expressivist must be allowed some materials to work with. Blackburn takes as basic rules that (are tailored to) validate modus ponens and modus tollens. Sonderholm has demonstrated that Blackburn's rules are insufficient for all the arguments we want to validate. I conjecture that the following rules are sufficient[87]:

(→I) $[\phi \vdash \psi] \vdash [\phi \rightarrow \psi]$

(MP) $[\phi \rightarrow \psi], \phi \vdash \psi$

(&I) $\phi, \psi \vdash [\phi \,\&\, \psi]$

(&E1) $[\phi \,\&\, \psi] \vdash \phi$

[87] Assuming that we do not have internal conjunction – see previous two footnotes.

(&E2) $[\phi \,\&\, \psi] \vdash \psi$

(¬I) $\phi, [[\phi \,\&\, \psi] \rightarrow \bot] \vdash \neg\psi$

(¬E) $\neg\neg\phi \vdash \phi$

(R) $\phi, [[\phi \,\&\, A(A) \rightarrow A(\bot)] \vdash A(\neg A)$

(N) $A(\neg\neg A) \vdash A(A)$

The need for (R) and (N), and the belief that their addition is sufficient, follow from the obvious point that once there is a distinction between internal and external negation, there need to be rules for internal negation introduction and elimination, and (R) and (N) are natural parallels of the classical rules for negation. This simplicity in the set of rules required should allow my proposal to count as a reply in the spirit of Blackburn's project. That project is to reconstruct the validity of all valid inferences from the acceptance of a set of rules designed to validate a small set of very basic inferences. Reductio arguments can legitimately be seen as part of this basic set, which gives expressivists a right to rule R. Thus, *pace* Sonderholm, quasi-realism has not yet broken its promise to make expressivism non-revisionist.

Before concluding that commitment semantics appears capable of doing all we need it to (though that is what I conclude), it is worth mentioning an objection made by Bob Hale (1993: 362 fn. 27).[88] Hale claims that commitment semantics needs to distinguish the expressive form of the conditional from the ordinary truth-functional conditional, and that Blackburn's proposal cannot do this because he gives the expressive conditional the very same inferential powers. But Hale is mistaken about what commitment semantics

[88] I concentrate on the argument in this footnote because the rest of the paper is concerned with showing that Blackburn's account of a model theory for commitment semantics fails. I consider those arguments to be irrelevant to the matter at hand here, since there is no need for such a model theory to allow commitment semantics to do the work it needs to. That is the contention which Hale attacks in this footnote (he reports Christopher Peacocke as putting it forward). Indeed, such a model theory might be seen as positively undermining the quasi-realist project by bringing in a form of descriptivism at the external level.

needs to do. It does not need to stave off quasi-realism from collapse; that it taken care of by the double two-concept view.[89] All that is demanded of commitment semantics is that it show how there can be logic without truth; in other words how moral arguments can make sense at the external level, before truth has been earned. To do this it is unnecessary for the account to mark out moral commitments from non-moral ones at the external level, since **all** claims express mental states of one kind or another. Hale is falling victim to the confusion mentioned in §2.1: thinking that just because the view that moral utterances express attitudes is called 'expressivism', descriptive utterances fail to express beliefs. Once it is clear that at the external level the expressive conditional applies to all discourse, in a way that does not need to be distinguished by different inferential powers, Hale's objection falls away.

It is worth explaining why commitment semantics does not fall victim to a problem which commonly faces attempts to give the meanings of logical constants in terms of their inferential roles – the possibility that this allows in Prior's (1960) 'tonk'. The potential objection is that we could add a new connective with introduction and elimination rules that trivialise the system. If the only way of seeing what is wrong with such a connective requires minimal truth, then the project of reconstructing logical consequence without presupposing minimal truth will have failed. But it is in fact possible to rule out trivialising connectives using resources that are available to commitment semantics. The way to do this is with the idea of satisfaction. A wff is satisfiable if the beliefs which the atomic wffs derivable from it (with other connectives) express could all be (maximally) true together, and the attitudes which the atomic wffs derivable from it express could all be satisfied

[89] Hale thinks that if quasi-realism succeeds in earning the right to talk of truth, then realism will result and quasi-realism will have collapsed. This point does not succeed so long as it is only minimal truth that has been earned. This point is made by Lenman 2003a.

together. We simply say that a connective is banned if its addition to the system permits the derivation of an unsatisfiable wff from a satisfiable one.[90]

For the sake of completeness it is worth explaining why a few other objections to commitment semantics which seem initially plausible ultimately fail. Unwin (1999: 341) has objected that Blackburn seems to conflate refusing to accept a sentence with the acceptance of its negation: 'To accept the negation of a sentence S is to accept something, whereas to refuse to accept S is consistent with accepting nothing at all'. At least on the extension of commitment semantics which I have outlined, this objection is groundless, since I explicitly allow for both internal and external negation, thus providing exactly the resources required to make Unwin's distinction.[91] Van Roojen (1996) has suggested that Blackburn's analysis conflates logical inconsistency with pragmatic inconsistency. The allegation is that the way in which two attitudes may be inconsistent is merely the way that Moore's paradox is inconsistent.[92] We shall see in the next section (§2.4) that the division between merely Moorean inconsistency and the kind of inconsistency used in commitment semantics creates serious problems for Blackburn's view. So van Roojen is at least quite mistaken about how commitment semantics is meant to work. One problem in evaluating the objection is that van Roojen may simply be assuming that our only grasp on semantic (i.e. non-pragmatic) inconsistency is via truth; but this would simply beg the question against commitment semantics. On the other hand, the expressivist might make the point

[90] Note that to do this we need operators indicating expressions of belief to be distinct from ones indicating expressions of attitude, because we do not want satisfaction to range over beliefs and desires together – there will be no problem of having a belief that p and a desire that $\neg p$.

[91] In §2.5 below I deal with an objection from Schroeder (2008a) also concerning negation. Readers may wonder what difference there is between the two objections, and why Schroeder's needs to be taken more seriously that Unwin's. The difference is that Unwin alleges that Blackburn can only account for one kind of negation not two, which is certainly a mistake; whereas Schroeder alleges that Blackburn can only account for two kinds of negation, not three, and indeed Blackburn's explanation of the third kind of negation is not satisfactory as it stands.

[92] Moore's paradox is saying both 'P' and 'I do not believe P', which seems inconsistent despite the claims having consistent truth-conditions.

that two attitudes are inconsistent if they are not co-satisfiable; but van Roojen might simply reply that this looks like pragmatic inconsistency to him, and then there is a mere clash of linguistic intuitions concerning 'pragmatic'. A better line for the expressivist is to say that if van Roojen were right, Blackburn would count all Moorean inconsistencies in attitude as semantic inconsistencies. In §2.4 we will see that there are cases of Moorean inconsistency within ethical discourse, and these are by quasi-realist lights pragmatic inconsistencies in attitude. So within the category of inconsistencies in attitude there is just the same contrast between Moorean and non-Moorean inconsistency that there is in the category of inconsistencies in belief (though commitment semantics cannot make the contrast). In the latter category we label the Moorean inconsistencies 'pragmatic', and the non-Moorean ones 'semantic', and we should do the same in the former category. So the expressivist can plausibly claim that non-Moorean inconsistencies in attitude are semantic, and thus that commitment semantics is justified in building its logic on that concept of inconsistency.

2.4 FALLIBILITY AND DOUBT

A problem for commitment semantics lies in what I shall call 'Moorean conditionals'. These are conditionals which it does not make sense to interpret as involving a commitment to accepting the consequent when one accepts the antecedent, since to accept both the antecedent and the consequent is Moore-paradoxical.[93] A well-known example is

(12) If Thatcher is a spy, I'll never believe it.

If I accept that Thatcher is a spy, it is inconsistent for me also to accept that I'll never

[93] This objection to commitment semantics I owe to Timothy Williamson (personal communication).

believe that she is, for by assumption I already believe that she is. So commitment semantics seems to entail that accepting (12) conditionally commits one to a Moorean inconsistency: one is committed to accepting 'Thatcher is a spy and I do not believe that she is' conditionally on accepting that Thatcher is a spy. In order to avoid this inconsistency I must reject the antecedent of (12) when I accept (12) itself; and this means that accepting (12) carries an unconditional commitment to holding that Thatcher is not a spy, or so commitment semantics has it. But that is no good, for (12) has a perfectly sensible interpretation which does not involve that kind of commitment: it expresses my expectation that my belief that Thatcher is a spy will not be sensitive to whether she is a spy, presumably because the evidence against her being a spy will always be compelling, even if she is a spy. (12) does imply an outright commitment, but it is only to Thatcher **probably** not being a spy. In short, commitment semantics says that (12) means something that it does not; it represents (12) as paradoxical when it is not.

Now this does not pose an immediate threat to commitment semantics unless there are cases like (12) within the discourses we want to be expressivist about. But it is not too hard to come up with such examples. For instance,

(13) If I ought not to support the minimum wage, I'll never believe it.[94]

Here the idea is that even if the minimum wage policy is wrong, I will always find the evidence in its favour compelling. Of course, that must be because I currently find the evidence in its favour compelling, but that evidence does not rule out the possibility that the policy is wrong. It is consistent to acknowledge the possibility that the policy is wrong, whilst knowing that one will never be able to see this (perhaps because one will never have the necessary knowledge of economics). But commitment semantics cannot interpret

[94] 'Believe' here is used loosely. If we wanted to be strict, it could be replaced with 'accept'.

(13) as this kind of admission of moral fallibility.

There is a reply available to the expressivist, but it is only partially successful. There is something odd about Moorean conditionals, so that the interpretation of them as verging on Moore-paradoxical is no mirage thrown up by commitment semantics. Moorean conditionals are not *ponenable*: *modus ponens* arguments involving them can never be cogent.

(13) If I ought not to support the minimum wage, I'll never believe it.

(14) I ought not to support the minimum wage.

---

(15) I'll never believe that I ought not to support the minimum wage.

Here the combination of (14) and (15) is Moore-paradoxical, so this argument can never give one reason to accept (15), though of course the argument is valid.[95] The expressivist will continue by noting the tight conceptual connection between indicative conditionals and *modus ponens*. To accept a conditional just is to be prepared to follow the relevant *modus ponens* argument; so if following the argument is inconsistent, accepting the conditional commits one to rejecting the minor premise of that argument, i.e. to rejecting (14). So the descriptivist is presented with a dilemma: either admit that commitment semantics interprets (13) correctly, or else admit that (13) isn't a genuine conditional.

The reason that this reply is inadequate is that the descriptivist does not need (in the current debate) to insist that (13) is a genuine conditional.[96] That is because commitment semantics interprets disjunctions as (two) conditionals. So we can forget about (13) and consider instead

[95] Just as Moore-paradoxical sentences are unassertible and unacceptable though they may be true.

[96] If we were arguing about the correct interpretation of indicative conditionals, then someone pushing (12) and (13) as counterexamples to the Ramsey/Adams interpretation would not concede defeat so easily. My own view, however, is that the argument given in the text does show that (12) and (13) are not normal indicative conditionals, and so they do not count against the Ramsey/Adams interpretation.

(16) Either it is not the case that I ought not to support the minimum wage, or I'll never believe that I ought not to support it.[97]

Now commitment semantics holds that disjunctions like (16) involve the conjunction of two conditional commitments: the commitment to accepting the second disjunct when one rejects the first, and the commitment to accepting the first disjunct when one rejects the second. In this case the first of these two commitments is just the one that got us into trouble earlier; it is the same commitment that was allegedly expressed by (13). That in turn implies the outright commitment to the minimum wage not being wrong. And of course (16) carries no such outright commitment. As before, it merely implies that the minimum wage is probably right. Whilst earlier we could plausibly argue that (13) was no ordinary conditional, it would be silly to contend that (16) is not a disjunction. Although (16) could be substituted for (13) in the above argument, there is no tight connection between accepting disjunctions and accepting such arguments. So the defender of commitment semantics is caught; (16) is a counterexample to the proposed interpretation of the disjunction.

Navigating our way out requires attending to the dialectical position. The expressivist has tried to do more than is necessary in claiming that commitment semantics can explain all disjunctions. There are certain sentences which only make sense on an internal reading; trying to give an external reading of such sentences is bound to be fruitless. Sentences like (13) and (16) only make sense in the way they do because of our fallibility. But for the quasi-realist, fallibility is an internal feature of the discourse.[98] The point is that if I am fallible I must have some defect in an admirable quality for moral

[97] I.e. Either it is permissible for me to support the minimum wage, or I'll never believe that it is not permissible for me to support it.

[98] It is not that from the external perspective we are infallible. Rather, the idea of fallibility is impossible even to formulate from that perspective. See Blackburn 1998a: 318.

epistemology. So to imagine myself as fallible I must have in mind such defects *as defects*. That is why (13) and (16) resist treatment via commitment semantics. The underlying phenomenon is just the same as with counterfactuals such as

(17) Even if no-one thought that lying was wrong, it still would be.

The quasi-realist will say that we automatically give such sentences an internal reading, and thus that it is perfectly compatible with quasi-realism to endorse them.[99] If that account is satisfactory for (17), it should be allowed for (13) and (16) as well.

So what the descriptivist has succeeded in showing is what the quasi-realist was prepared to admit anyway: that some sentences cannot be correctly interpreted by any external semantics like commitment semantics. Two questions emerge. Firstly, how exactly should the quasi-realist restrict commitment semantics so that it does not overreach? Secondly, does the failure of commitment semantics to give the meaning of all sentences of the discourse invalidate the quasi-realist's response to the Frege-Geach challenge? With respect to the former question, the quasi-realist should say that commitment semantics gives the meanings of sentences to the extent that it fully explains the validity and cogency of arguments involving them.[100] Since the creation of an internal perspective allows for additional elements of meaning, commitment semantics does not account for those elements directly. On the latter, more important question, the quasi-realist is in a strong position. The key point is that the demand placed on commitment semantics to fully explicate the meanings of all embeddings was too strong.[101] What

[99] See for instance Blackburn 1998a: 296.

[100] Commitment semantics correctly has it that the (12-15) argument is valid but not cogent; what it seems to get wrong is the meaning of (13). If the descriptivist now insists that validity is not enough, that is an interesting reversal of the original Frege-Geach point, which was that expressivism had a problem with explaining validity. It now seems as if validity is one thing which expressivists do not have a problem with.

[101] It was always part of the quasi-realist position that the view from the external position is incomplete. Thus any insistence that the quasi-realist explain the meanings of all moral sentences without use of the internal perspective (which includes truth) begs the question against quasi-realism.

commitment semantics does do is provide an account of a sufficiently large fragment of ethical discourse (excluding the part which assumes fallibility) to provide discipline which justifies embeddings. All that commitment semantics needs to do is to provide enough discipline to justify the syntax, which in turn earns us the right to speak of truth for the discourse. We can then use truth (once earned) to explain the sentences that were previously recalcitrant. Thus quasi-realists have not given up on explaining those sentences; rather they distinguish between those sentences which can be explicated directly through commitment semantics, and those like Moorean conditionals, which can be explicated indirectly, by going via truth.

This way of dealing with Moorean conditionals may well be applicable to a different problem urged against expressivists. Michael Smith (2002) has argued that expressivists have difficulty in explaining how moral judgements can display different levels of certitude. Insofar as moral judgements are really a kind of desire, they can vary in strength, but it would be a mistake to see that variation as mapping onto certitude: if the strength of desire maps onto anything, it will be the importance of the judgement, which is different from how certain it is.[102] This means that the expressivist cannot hope that degrees of desire will do service for degrees of belief. Another feature of desires is their robustness – how subject they are to change with the circumstances (e.g. new information) – but again this is clearly different from certitude, because I can be certain even when (I do not realise that?) my judgement is not robust, or uncertain even though in fact nothing will make me change my mind. So Smith's challenge, which he thinks cannot be answered satisfactorily, is to find a feature of desires which could correspond to

[102] One can be certain about an unimportant judgement, and uncertain about an important one: it is surely more plausible that there is a strong desire in the latter case than in the former.

certitude (of moral judgements), as expressivism appears to require.[103]

The way that Smith's certitude problem relates to that of Moorean conditionals is that both appear to involve irreducibly internal judgements. Certitude involves a judgement concerning how easily one could be wrong. Moral uncertainty (when it is a matter of not being sure of what the correct moral principles are, rather than of whether the facts are such that those principles apply) implies doubt about one's own ability to distinguish correct from incorrect moral principles. One thing we might say then, as above, is that the question of certitude only gets going once the discourse has already attained the level of discipline necessary for truth, which allows the formulation of the internal question, "Is my judgement true?" and thereby allows us to think in terms of certainty and degrees of belief. So it is tempting to say that once the core Frege-Geach Problem is solved, earning expressivists the right to talk of moral truth, the right to talk of moral certitude must follow.[104] Smith might plausibly reply, however, that the expressivist is still committed to moral judgements being capable of being seen as desires (from the external standpoint), so that they must still in that guise have some feature which corresponds to certitude. If we just say that a successful resolution of the Frege-Geach problem is bound to allow for moral certitude, we still leave a mystery concerning what that amounts to from the external perspective.

The first step towards dissolving this mystery lies in avoiding a certain misconception about degrees of belief and degrees of desire. We might think that a belief of degree <1 is just like a belief of degree 1, except somehow attenuated, perhaps by analogy to a sound with the volume turned down. I think that once such a picture is made

[103] There have been various attempts to solve this problem which I do not intend to compare or discuss in detail, but see Lenman 2003b, Ridge 2003, Bykvist & Olson 2009 and Sepielli 2012.
[104] Sepielli (2012) makes the same claim, though gives a somewhat different argument.

explicit, it is immediately implausible. We do not need to think of beliefs as items with a kind of volume setting in order to make sense of degrees of belief. Rather beliefs are functional states: we understand what it is for beliefs to come in degrees in terms of what they do (e.g. rationalising action such as betting). What the intrinsic nature of beliefs is which allows them to perform this function is not a question that philosophers (as opposed to e.g. neuroscientists) can be expected to have much insight into. Similarly, the functional role of degrees of desire is that when there are trade-offs between the things we desire we have to choose between them, and a desire being stronger is manifested in its winning out in this choice. What the expressivist needs to do then is show that the functional role of moral certitude can still be understood from the external standpoint (with moral judgements thought of as conative states); but this does not mean finding some feature of desires which lines up perfectly with the certitude of beliefs.

The best idea to pursue is Blackburn's (1993a: 20, 1998a: 318 & 1999) thought that for an agent to recognise the possibility of error is to acknowledge the possibility of an improvement to her epistemic position. Since we commonly recognise that we might be more knowledgeable, experienced, sensitive etc., and that we might have different attitudes if we were (or advise our actual, present selves to have different attitudes[105]), we thereby accept the potential of improvement to our attitudes. For any moral attitude and any change I might undergo, I make two judgements: a partial belief concerning the likelihood of undergoing this change resulting in a change to my attitude, and an attitude concerning the change itself (whether it counts as an improving change). When I have a positive attitude towards a potential change (e.g. gaining more experience of a certain kind), and I think there is a chance that this change would induce a change to some moral

[105] See Lenman 2003b, which argues for a position I think equivalent to mine, though the presentation differs.

judgment (attitude) I now have, this amounts to uncertainty concerning the moral judgement itself. This means that the expressivist can explain how the functional role of moral certitude is filled even from the external standpoint where moral judgements are viewed as attitudes rather than beliefs.

I will address two complaints about this kind of solution to Smith's problem. First, Sepielli (2012) worries that Lenman's (2003b) solution is a form of ecumenical expressivism, which finds moral certitude in a cognitive state rather than a conative one – he would doubtless say the same about my approach, which is similar to Lenman's.[106] The response to this is that we do not need to think of the certitude of a state as something intrinsic to it, but should instead think in terms of its functional role. The question of whether a certain mental state can be improved on (i.e. is in error) is a matter of whether a change in epistemic situation would result in a change in that mental state, and whether it would count as an improvement if it did; this applies whether the mental state is cognitive or conative. So at the very least the conative attitude to the change is involved rather than just the belief about the likelihood of the change altering the moral judgement. Certitude is always a matter of a certain mental state having a certain functional role within the whole mental economy, and there is no need to worry about the degree of belief depending on an aspect of the belief itself rather than its context.

Second, Bykvist & Olson (2009) complain that any solution along the lines of Blackburn and Lenman (and so mine as well) only tells us about empirical uncertainty, not genuine moral uncertainty. I find this complaint somewhat obtuse: the whole point is that it does deal with uncertainty about the moral principles themselves rather than in the

[106] Sepielli's own solution builds on the solution to the Frege-Geach problem of Schroeder (2008a); I explain in §2.5 below why we should reject Schroeder's view, and so *a fortiori* why Sepielli's solution to Smith's certitude problem is unavailable.

empirical circumstances (which would determine which principles apply). But I think the root of Bykvist & Olson's concern is that according to the expressivist all I am really uncertain about when I lack moral certitude is the empirical question of whether my moral judgement would be altered if my epistemic situation changed in some way. Now I do not necessarily think there is a genuine objection here, because the expressivist can say that it applies across the board, so there need be nothing especially moral about the uncertainty itself: whenever I am uncertain about my judgement the question is whether I would judge the same way if my epistemic situation improved. A different way of developing Bykvist & Olson's worry is that the solution assumes certainty about what counts as an improving change, when uncertainty about that is part of what needs to be accounted for. Again, however, this seems uncharitable: the account can simply be iterated with respect to judgements concerning what changes are improving (since these judgements are also attitudes). When performing such iteration some restrictions are bound to be in play: in order to consider my judgements about one kind of change I must hold fixed at least some of my other judgements about which kind of changes are improvements, rather than questioning them all at once (not all the planks of Neurath's boat can be repaired at the same time). Still, there is clearly room to be uncertain about any particular judgement that a change would count as an improvement. One way of taking this kind of issue further is to worry about whether the expressivist can make sense of fundamental moral uncertainty and fundamental moral error. But since I think that there are reasons to doubt whether anyone can make sense of fundamental moral uncertainty, I do not believe that this kind of challenge is one expressivists are required to meet. I discuss Andy Egan's (2007) objection to quasi-realism from fundamental moral error at length in Chapter 3 where it is

more appropriate to do so.[107]

Let us return briefly to van Roojen's criticism of Blackburn, which was that Blackburn conflates logical inconsistency with pragmatic inconsistency. There is a grain of truth to this accusation: **commitment semantics** cannot fully make the distinction: it counts at least some pragmatic inconsistencies as logical contradictions. But this does not mean that **quasi-realism** conflates the two kinds of inconsistency. We need the (minimal) concept of truth in order to make the distinction, but quasi-realism insists that it gives us a right to this concept in ethical discourse. Once we have that concept we can see that some of the inconsistencies captured by commitment semantics were genuinely logical, and some were not. Van Roojen's point would only have been decisive against the claim that commitment semantics alone was sufficient to give an account of the logic of moral discourse directly. But we can use commitment semantics to earn us the right to talk of truth in ethics, and so to give us the logic indirectly, via truth. All this allows us to see what is right in Blackburn's suggestion that there might be a high road and a low road to solving the Frege-Geach Problem. The suggested quick and easy route is to hijack truth, and so do without a semantics; the arduous path is commitment semantics (or some substitute). We now see that neither road is sufficient alone. We need commitment semantics in order to earn the right to speak of truth (as we say in §2.2), and we need to speak of truth in order to take in the cases which elude commitment semantics (as was argued above). They are thus not alternative paths, but rather essential components in a solution including both.

[107] It may be that the nub of Bykvist & Olson's objection is ultimately the same as Egan's: if so, I deal with it there, having dealt with Smith's original objection here.

2.5 THE NEGATION PROBLEM

Mark Schroeder (2008a, 2008b) takes the view that the really hard problem for expressivists is not showing that a logic of the attitudes is possible (he concedes that it is), but showing that it can account for the complex logic of morality (i.e. deontic logic). Schroeder (2008a) presents a problem regarding negation; he also suggests a solution to that problem; but the solution he offers is a poisoned chalice for expressivists, because Schroeder (2008b) goes on to argue that it breaks down in the end. I argue that in fact the expressivist can solve the problem more straightforwardly and without leaving these hostages to fortune by borrowing the clothes of the realist in a new way: adapting the standard semantics for deontic logic by giving an expressivist construal of acceptability.

Schroeder (2008a) observes that for a typical sentence describing a moral judgement such as:

(18) Jon thinks that murdering is wrong.

there are three negations:

(19) Jon does not think that murdering is wrong.

(20) Jon thinks that murdering is not wrong.

(21) Jon thinks that not murdering is wrong.

Suppose that expressivists give a standard account of Jon's judgement that something is wrong in terms of his disapproving of it. So what is going on in (18) is that Jon disapproves of murdering; in (19) Jon does not disapprove of murdering; in (21) Jon disapproves of not murdering; but what is his attitude in (20)?

The problem is that there does not seem to be a spare place for the third negation to go. What expressivists seem to need is some mental state which is inconsistent with

disapproval of murdering but which is distinct from all of the following: merely not disapproving of murdering; disapproving of not murdering; approving of murdering.

A standard expressivist line (starting with Blackburn 1988) is that what is going on in (20) is that Jon *tolerates* murdering, where the attitude of toleration is such that it is inconsistent to both tolerate and disapprove of the same thing. This will then explain why Jon is inconsistent if both (18) and (20) are true. But Schroeder holds that this is not a good explanation of the inconsistency, because we do not have an explanation of **why** toleration and disapproval are inconsistent. Thus the invocation of the attitude of toleration seems to have only the advantages of theft over honest toil.

To see what constraints Schroeder thinks the expressivist lies under in explaining (20) and its inconsistency with (18), and why he thinks that only his kind of solution can meet those constraints, we need to see what resources expressivists have available. Schroeder grants that expressivists are entitled to some inconsistencies between attitudes. It is inconsistent both to plan to stay home and to plan not to stay home. The inconsistency of those states derives from the inconsistency of their contents: the attitudes are inconsistent because staying home is incompatible with not staying home. This explanation of the inconsistency of planning attitudes is just the same as the explanation of inconsistency between beliefs: it is inconsistent to both believe that p and believe that not-p, because p and not-p are inconsistent.

So the kind of explanation of inconsistency which expressivists have available to them derives the inconsistency of attitudes from the inconsistency of the contents of those attitudes. Some attitudes, such as planning and intending, transmit inconsistency (like belief), in the sense that when two contents are inconsistent with each other, then it is inconsistent to take the same, inconsistency-transmitting attitude to both contents.

Schroeder calls this kind of inconsistency A-type inconsistency. Expressivists are entitled to to A-type inconsistency. But the alleged inconsistency between tolerating and disapproving is not A-type inconsistency, because they are different attitudes to the same content, rather than the same attitude to inconsistent contents. So expressivists who adopt the standard solution in terms of toleration seem to be invoking a second, unexplained kind of inconsistency, B-type inconsistency, and this appears unsatisfactory.

Schroeder offers the expressivist a solution which does everything in terms of A-type inconsistency. He suggests the following analyses (marked with *s below):

(18) Jon thinks that murdering is wrong.

(18*) Jon is for blaming for murdering.

(19) Jon does not think that murdering is wrong.

(19*) Jon is not for blaming for murdering.

(20) Jon thinks that murdering is not wrong.

(20*) Jon is for not blaming for murdering.

(21) Jon thinks that not murdering is wrong.

(21*) Jon is for blaming for not murdering.

We observe that Schroeder invokes a single attitude, *being for*, but makes the content that this attitude attaches to more complex: instead of contents such as *murdering*, we have contents such as *blaming for murdering*. Doing things this way makes an extra space for negation, so that there are now three spaces for negation in the analysis of (18), just as in (18) itself. Moreover, the inconsistency between (18) and (20) emerges as an A-type inconsistency on this account: *being for* is an inconsistency-transmitting attitude; being for blaming for murdering is inconsistent with being for not blaming for murdering, because blaming for murdering is inconsistent with not blaming for murdering.

Schroeder's solution has a certain elegance, but it is not entirely intuitive. One problem is that on some views in normative ethics it must be possible to think that murdering is wrong without being for blaming for murdering. The kind of view I have in mind is a consequentialist view according to which the decision concerning whether to blame for murdering depends on the consequences of blaming for murdering, whereas whether murdering is wrong depends on the consequences of murdering. It is possible that the consequences of murdering are bad, but the consequences of blaming for murdering are also bad, in which case murdering is wrong but one should not be for blaming for murdering. Such a view may be implausible, but it seems coherent; and yet Schroeder's account appears to make it impossible to hold such a view. It is orthodox (though perhaps incorrect) to hold that positions such as expressivism should be neutral with respect to normative ethics, and so advocates (and even opponents) of expressivism would thus be reluctant to endorse a reconstruction of expressivism that makes it incompatible with such a form of consequentialism. There may be some reformulation of Schroeder's solution that avoids this solution (involving some alternative to *blaming* – Schroeder mentions *avoiding*); or it may be that Schroeder can bite the bullet on ruling out the possibility of the view just described in normative ethics. I do not intend to argue that Schroeder's solution does not work, merely to point out that we need not find it immediately compelling in a way that would lead us to think that he must have got hold of the right solution.

It is worth considering Schroeder's alternative suggestion of replacing *blaming* with *avoiding*, both because it may seem the best version of his view (immune to the above objection) and because doing so allows us to see that there is something implausible in

itself in thinking that the inconsistency between (18) and (20) is A-type. The following analyses would apply on that suggestion:

(18) Jon thinks that murdering is wrong.

(18') Jon is for avoiding murdering.

(19) Jon does not think that murdering is wrong.

(19') Jon is not for avoiding murdering.

(20) Jon thinks that murdering is not wrong.

(20') Jon is for not avoiding murdering.

(21) Jon thinks that not murdering is wrong.

(21') Jon is for avoiding not murdering.

The whole point of Schroeder's analysis here is that the inconsistency between (18') and (20') is a consequence simply of the inconsistency between avoiding murdering and not avoiding murdering, which is held to be transmitted by *being for*. But is it plausible that the difference between someone who thinks that murdering is wrong, and someone who thinks that murdering is not wrong, is that their attitudes have different contents, though they have the same attitude? Consider the standard expressivist idea that the motivational role of moral judgements is to be explained by thinking of them as attitudes. According to the expressivist, someone who thinks that murdering is wrong will be appropriately motivated, and this is explained by their (desire-like) attitude. What then of someone who thinks that murdering is not wrong? It has not been thought that there is anything motivational to explain here: thinking that murdering is not wrong does not seem to have the same kind of motivational import as thinking that murdering is wrong. Before Schroeder, we would explain this by the difference in attitudes: the motivational import of disapproval is different from that of toleration. But Schroeder holds that we have the same

attitude in both cases: someone who thinks that murdering is not wrong will have just the same kind of motivation for not avoiding murdering as someone who thinks that murdering is wrong has for avoiding murdering. Is this plausible?

Perhaps the response will simply be that there is only an apparent difference between the motivational powers of the different moral judgements here because there is a difference in whether or how frequently these powers are manifested. *Being for avoiding murdering* will have its motivational power manifested whenever the agent is deciding whether to murder: it will motivate her not to murder in those cases. When, in contrast, is the motivational power of *being for not avoiding murdering* manifested? Not in cases where the agent is deciding whether to murder, because not every decision not to murder will count as a decision to avoid murdering (if it did, then *being for not avoiding murdering* would be equivalent to *being for avoiding not murdering*, which is obviously not what Schroeder, or the expressivist, wants). What we seem to have to say is that *being for not avoiding murdering* will motivate the agent to murder in cases where not murdering would count as avoiding murdering; but which cases are those? One thought is that the agent would be avoiding murdering if she were motivated by *being for avoiding murdering*; but we cannot understand avoiding murdering in that way because that content must be understood prior to *being for avoiding murdering*, on pain of circularity. So perhaps what is intended is that to avoid murdering is to choose not to murder (partly) on the basis that the contemplated act is murder (that is to choose non-murder qua non-murder). This captures the idea that to think that murder is not wrong is to reject being constrained from murdering by the thought that what one would be doing is murdering. But still it is hard to avoid the thought that in the end the motivational role of *being for not avoiding*

murdering is just that it involves rejecting being motivated in the way one would in *being for avoiding murdering*.

So if we classify attitudes by their motivational role it seems that the orthodox expressivist is right to understand the distinction between disapproval and toleration as a difference in attitudes rather than in contents. Schroeder may be right that a form of words like *being for not avoiding murdering* can stand in for *tolerating murdering*, but only at the cost of classifying attitudes in a way that is at odds with their grouping in terms of psychological role. Of course, this is not at all to defend those orthodox expressivists, who employed the attitudes of disapproval and toleration, against Schroeder's criticism that they give no proper explanation for the inconsistency between the two attitudes. But it does remind us that the idea of **different** inconsistent attitudes is attractive here, if it can be adequately explained. What we want is a story which explains why those two attitudes are inconsistent, still classifying attitudes in accordance with their psychological roles. We need a good way of understanding what toleration is that both explains the inconsistency and respects our classification of attitudes in accordance with psychology. Schroeder's kind of solution should immediately make us suspicious that it distorts the psychology, simply in virtue of trying to do everything in terms of A-type inconsistency.

What kind of alternative solution is possible? It may seem that there is no room for an alternative solution, because in order both to respect the constraint of making the relevant inconsistency A-type, and to have three places for negation, only a structure like the one offered by Schroeder will do. I will not rehearse the details of Schroeder's arguments against Horgan and Timmons (2006), Gibbard (2003) and Dreier (2006). The bottom line is that they all rely on B-type inconsistency in an objectionable way, because they do not have any way of explaining the state of tolerating murdering which goes

beyond the stipulation that it is inconsistent with disapproving of murdering. Moreover, this is not just a commitment to one unexplained kind of inconsistency; the unexplained attitudes proliferate indefinitely. Schroeder gives the example of conjunction:

(22) Jon thinks that murdering is wrong and Jon thinks that stealing is wrong.

(23) Jon thinks that murdering is wrong and stealing is wrong.

Once we see that (23) is not the same as (22), Jon's attitude in (23) cannot simply be disapproval, and so the question arises of which attitude is inconsistent with that attitude. It will not simply be toleration, since the attitude we need it to be inconsistent with isn't disapproval; so we seem to be stuck with another pair of B-type inconsistent attitudes without an explanation. Clearly we can obtain more such pairs by starting with more complex sentences.

The alternative I will outline is that the relevant moral judgements are to be understood in terms of conditions on the agent's preferences, which I call *preference conditions*. I will explain this concept by giving one way of describing an agent's preferences and then illustrate using examples of different preference conditions. An agent's preferences are defined over a set of worlds representing potential total states of affairs. The agent's overall preference function, ***v***, assigns rankings, numerical values, to such worlds (but perhaps not every world is ranked). A content such as *murdering* partitions the worlds into those that satisfy the content and those that do not: M worlds (where there is murdering) and non-M worlds (where there is no murdering). A preference condition is a condition on the rankings of the worlds that satisfy a particular content. The agent has an acceptability threshold, which is some value *t*, such that the agent takes every world valued at *t* or higher to be acceptable and every world valued at less than *t* to

be unacceptable. (Note that by not specifying the acceptability threshold we leave it open whether the agent satisfices or maximises.)

One further clarification: the preferences involved here are not to be thought of as mere personal preferences, but rather as moral preferences. We assume that the expressivist has the resources to distinguish between these two kinds of preference, since this is something which the expressivist has to do anyway, and the challenge of doing so is quite distinct from the Frege-Geach Problem. With this background, we can start to give examples.

We start with (18). When Jon thinks that murdering is wrong, he finds every M world unacceptable, meaning that his preferences satisfy

(18#) $\forall w[Mw \supset v(w) < t]$.

How about (19)? When Jon does not think that murdering is wrong, he does not find every M world unacceptable, which is to say that his preferences do not satisfy $\forall w[Mw \supset v(w) < t]$ and so they do satisfy

(19#) $\neg\forall w[Mw \supset v(w) < t]$.

Note that one way this could occur is by *v(w)* being undefined for some M world, because Jon has not ranked that world.

And now (20)? When Jon thinks that murdering is not wrong, he finds some M world acceptable, which is to say that his preferences satisfy

(20#) $\exists w[Mw \,\&\, v(w) \geq t]$.

And (21)? When Jon thinks that not murdering is wrong, he finds every non-M world unacceptable, so his preferences satisfy

(21#) $\forall w[\neg Mw \supset v(w) < t]$.

To explain why this counts as a solution we first note that the analyses of (19), (20) and (21) are distinct; in particular (19#) $\neg\forall w[Mw \supset v(w) < t]$, does not entail (20#) $\exists w[Mw \ \& \ v(w) \geq t]$, because the former is consistent with Jon not assigning a value to some situations, which is just the kind of agnosticism we think is involved when he does not think that murdering is wrong but does not think that murdering is not wrong either.

What about the inconsistency between Jon's thoughts in (18) and (20)? There is no preference function such that every M world is unacceptable (18), and some M world is acceptable (20). So there is no problem explaining inconsistency. The problem is that this analysis appears to make it **impossible** to think both things, rather than merely inconsistent. The difficulty is the assumption that Jon's preference function is a genuine function, i.e. it assigns at most one value to each argument (world). Insofar as Jon's preferences really do work like this, it is not possible for him to both think that murdering is wrong and that it is not wrong. But we can still model the situation where he does think both these things, if we allow that he can assign inconsistent values to an M world so that it is both acceptable and unacceptable to him. Another way of putting this inconsistent assignment of values is to say that Jon would prefer that M world to itself, which violates a basic constraint on the coherence of preferences.

We have then an explanation of how Jon is inconsistent when both (18) and (20) are true: he finds every M world unacceptable and also finds some M world acceptable, which means that he has an incoherent preference regarding some M world, preferring it to itself. Is this an A-type inconsistency? Apparently not, since it is the inconsistency between two attitudes to the same content, rather than the same attitude to two inconsistent contents. Perhaps there is some devious way of showing that incoherent preferences involve an A-type inconsistency, but as I have indicated earlier I do not agree

with Schroeder that this is even desirable. Let us proceed on the assumption that on this account all we have is a B-type inconsistency. Does this mean that Schroeder's objections still apply?

Schroeder's main worry about B-type inconsistency is that it is unexplanatory. But the explanation I have given above seems to meet this worry. The kind of inconsistency at stake is the kind involved in failing to have a coherent preference function. We believe in this kind of inconsistency quite independently of expressivism, and so it is not ad hoc for the expressivist to rely on it. To see why this kind of solution makes sense we can usefully think in terms of deontic logic, since it is that logic, with its inconsistency between *obligation not to* and *permission to*, which the expressivist needs to recapture. But the standard way of doing semantics for a deontic logic is as a modal logic, with a possible world semantics (see, just as an example of how this is orthodox, McNamara 2010). The basic idea is to have an acceptability relation between worlds (like the accessibility relation in alethic modal logic), and define the deontic operators in these terms, for example saying that *p* is impermissible at world *i* (in the simplest case, the actual world) when *p* does not occur at any world acceptable from *i*. The relevant point is that the thematic thing for the expressivist to do is to try to imitate this semantics, and the preference function provides a way to do so. By defining acceptability from the point of view of a preference function, we recapture the structure used to define the deontic operators. Effectively we embed the structure of deontic logic into the structure of the preference function. Seen in this light, the solution put forward here is very natural: since we can give the semantics for deontic logic in terms of the acceptability of worlds, and given that an agent's preference function delivers an account of the acceptability of worlds from that agent's perspective, we can

put these moves together to obtain a semantics for an agent's deontic judgements in terms of her preference function.

One disanalogy between this account and the standard semantics for deontic logic is worth emphasising to avoid confusion. The standard semantics gives truth conditions for moral claims, e.g. murder is permissible if there are some acceptable M-worlds. To extend this to an analysis of moral belief, we would just say that insofar as Jon thinks that murdering is not wrong (i.e. permissible), he thinks that there are some acceptable M-worlds. Now the expressivist can in turn analyse this in terms of Jon's preferences: what it is for Jon to think that there are some acceptable M-worlds is for his preferences to be such that some M-worlds meet his acceptability threshold. But this does not mean that the expressivist relativizes the acceptability of worlds to the preference functions of individuals. If Jon says, "Murder is not wrong", his claim is not true just in case some M-worlds meet his acceptability threshold given his preference function – that would be subjectivism rather than expressivism. We have a truth-functional semantics for claims about what Jon thinks about murder, but not for moral claims themselves. Jon's moral claim expresses his preference function (to the extent of it satisfying the relevant preference condition), rather than reporting it; that he has such a preference function is a sincerity condition rather than a truth-condition for his claim. This is just a reminder that the distinction between expressivism and subjectivism remains in force and is being respected.

It is worth noting here that it is in some ways surprising that the discussion of the negation problem for expressivism so often proceeds without mentioning deontic logic. After all, the general shape of the debate is to show that realists have logical resources which expressivists do not; and so one would expect such a debate to consider how it is

that realists get to solve the relevant problem themselves. Of course the omission is not entirely surprising: the realist will think that their right to an inconsistency between thinking that murdering is wrong and thinking that murdering is not wrong follows straightforwardly from understanding wrongness as a genuine property. And yet given that thinking that murdering is wrong is the same as thinking that murdering is impermissible, and thinking that murdering is not wrong is the same as thinking that murdering is permissible, and impermissibility and permissibility are deontic operators for which there is a standard semantics to explain the inconsistency, one might think that the question of how much of that semantics expressivists might be able to capture should be on the radar.

What of Schroeder's second worry about B-type inconsistency: that the expressivist ends up committed to indefinitely many pairs of irreducibly inconsistent attitudes? Because the solution under consideration embeds the semantics of deontic logic, this problem will not arise: once we get a single notion of B-type inconsistency in terms of the incoherence of the agent's preferences, this can be used to explain all the inconsistencies we need in the obvious, systematic way. To see this, we will consider Schroeder's example from earlier:

(23) Jon thinks that murdering is wrong and stealing is wrong.

This is just to say that Jon finds every M world unacceptable and every S world unacceptable, which is to say that his preferences satisfy $\forall w[(Mw \vee Sw) \supset v(w) < t]$.

The attitude inconsistent with this is where Jon finds some world acceptable that is either M or S, which is to say that his preferences satisfy $\exists w[(Mw \vee Sw) \,\&\, v(w) \geq t]$.

The inconsistency between these two attitudes is just the same B-type inconsistency as before: to hold both attitudes Jon must have incoherent preferences. So it

is a mistake of Schroeder to assume that B-type inconsistencies cannot have the kind of systematicity required. It is plausible to think that the requirement of coherent preferences will suffice to explain all the relevant inconsistencies.

Next consider how well my proposed solution fares with the requirement (urged above against Schroeder's A-type inconsistency solution) that what we say about the attitudes conforms to moral psychology. There is a well worked-out theory – decision theory – which at least to a first approximation explicates the role of preferences in explaining action. So it is hardly as if doing everything in terms of preferences is going to lead to any problems with explaining moral motivation. Disapproval and toleration are here understood as different kinds of preference conditions, and so as different attitudes in that sense with different motivational roles; but the fact that they are both conditions on preferences means that we still have a unity to the explanation of moral motivation (in terms of preferences).

One worry about B-type inconsistency that Schroeder does not emphasise so much is that that it is a less genuine kind of inconsistency: that if disagreement in attitude is all there is here then the expressivist is still conceding ground to the realist. I think the best way to combat this kind of suspicion is to say more about how expressivists should understand moral thinking in view of the account given here. Expressivists hold that moral thought is fundamentally a kind of practical thought: moral judgement is a special kind of judgement about what to do. So it can hardly be surprising that what moral consistency comes down to is having coherent preferences (of that special kind): to have coherent preferences just is to fail to have a consistent view about what to do. We can understand moral disagreement in the same way: we disagree morally when there is no coherent way

of combining our moral judgements into an overall consistent view about what to do; that is to say when the conditions on our preferences cannot be coherently satisfied together.

Before concluding, a brief word on how this solution relates to ones in the literature. As Schroeder highlights, solutions such as that of Gibbard (2003) which focus on planning states will not have the same resources; this is because the preference function carries richer information on which to build a semantics than any hyperplan. A preference function is capable of representing the acceptability or unacceptability of a world, whereas a hyperplan can represent only acceptability and not unacceptability. Closer to this solution is the treatment of the Frege-Geach Problem found in Weintraub (2011), though her focus is not on the negation problem. Weintraub's approach is to do the whole semantics in terms of the agent's preference function, which means that it is straightforward to extend her approach to the deontic case in exactly the way indicated here.

The solution to the negation problem offered here involves B-type inconsistency, but neither leaves the nature of that inconsistency unexplained or ad hoc, nor requires there to be more than one type of B-type inconsistency. This shows Schroeder's insistence that an adequate solution must rely only on A-type inconsistency to be unwarranted. Moreover, because the strategy of recreating the semantics of deontic logic relative to the agent's preference function is less revisionary to the standard (descriptivist) semantics for deontic logic, it has some claim to being a more natural solution than the one offered by Schroeder.

What I hope to have done in this chapter is show that no version of the Frege-Geach Problem provides a good reason for avoiding expressivism. This task is important because it has proved one of the most compelling obstacles to the adoption of

expressivism, and doubtless contributed strongly to Hare's own views falling into disfavour during the 1980s. Because the Frege-Geach Problem is apparently a technical one, there is a lamentable tendency towards unwarranted arguments from authority ("As a logician I can tell you that this problem is insoluble" etc.) which may serve to intimidate meta-ethicists who are not technically minded. Perhaps some of the foregoing explanatory work provides some reassurance that expressivism does not have to be abandoned for obscure technical reasons, and that meta-ethicists are capable of assessing its merits themselves. In particular, that portion of neglect for Hare's views that stemmed from this worry should be abated.

CHAPTER 3: NORMATIVE JUSTIFICATION

3.0 INTRODUCTION

In this chapter I will consider the prospects for obtaining an acceptable view of moral objectivity given quasi-realism. Taking it that quasi-realism has been satisfactorily established in Chapters 1 & 2 (at least bracketing the worries considered in this chapter), I turn to some difficulties which Blackburn's position faces which are not (perhaps surprisingly) consequences of his quasi-realism, but rather of his other commitments, and in so doing I hope to sketch out a distinct position that is more in line with Hare's. I will argue that Blackburn's Humean approach to normative justification is inadequate, and that a Kantian approach is superior. I will then show how to reconstruct Kant's (1996/1785) arguments in *Groundwork* §II-III within the quasi-realist framework, and thus how his Categorical Imperative and Formula of Universal Law can be established.

There is a strategic reason for pursuing this line, in my overall project of rehabilitating Hare's version of expressivism. The central distinction between Hare and other expressivists, including Blackburn, is Hare's insistence on universalizability, not as one normative principle amongst many, but as an indispensable part of anything comprehensible as morality. Unfortunately, Hare's arguments for this position tend to be linguistic: he tries to elicit linguistic intuitions to show that moral terms had to be used in a universalizable way.[108] The problem with this strategy is that even if we do use moral terms in the way that Hare claims, that does not in itself provide us with a reason not to change our usage; thus the danger is that even if Hare succeeds in showing that utilitarianism can be derived from universalizability, he will not have shown that we should be utilitarians, because we will still be able to ask why we should not start thinking in non-

[108] See e.g. Hare 1952: 129-30 and 1963: 10-3.

universalizing terms. And if Hare insists that the use of the word 'should' commits us to universalizing, we can ask why we cannot use it differently.[109] He could reply that when we are doing moral philosophy we are in the business of answering questions of the form, 'What ought I to do?', and since this question makes use of the word 'ought', it is appropriate for us to consider the meaning of that word when answering it. But this seems only to push the worry further up the line: we could be interested in different questions, so what Hare's story leaves out is an explanation of the importance and inescapability of *that* question. Why are the questions we ask when moralising better or more important questions than those we would ask if we were shmoralising instead (where by 'shmoralising' we intend a practice similar to moralising but not implying universalizability)?[110] Indeed, there is a risk that Hare's theory will end up being as implausible as the analytical descriptivism that was dismissed in Chapter 1, in that they both seem to make the ultimate normative questions into questions about meaning.[111] What we need is an argument that we **must** engage in a form of thinking that involves the requirement of universalizability; an argument that normativity itself presupposes

[109] This objection is considered by Singer (1990: 156-9) and Blackburn (1998: 227). Hare (1989b: 87) does mention the problem explicitly (noting that Blackburn worried about this objection when reading a draft of *Moral Thinking*): he imagines someone saying that if the ordinary uses of moral words commit them to utilitarianism they will simply abandon those ordinary uses.

[110] The only place I know of where Hare says something explicitly directed to this problem is at 1989b: 88, where he says that the explanation of why we should keep asking the moralising questions in given in *Moral Thinking* ch. 11 (1981: 188-205). But there the argument is that there are prudential grounds for moralising rather than abandon morality altogether and falling into amoralism. Even if we were satisfied with an argument based on prudence, ruling out complete amoralism falls a long way short of justifying moralising, because shmoralising is not envisaged as throwing off the institution of morality altogether, but rather as replacing it with a new institution with at least some of the same features, though without universalizability. The kinds of reasons Hare gives for thinking that there is something imprudent about amoralism do not obviously apply to that kind of shmoralism.

[111] Hare (1981: 218-28) worries about this problem, but whilst I do not disagree with what he says in response, it does not seem to answer the objection; Hare seems more concerned about the classificatory question of whether he counts as a descriptivist than in the more pressing problem of avoiding the implausibility of descriptivism.

universalizability.[112] And that is what this chapter aims to provide. But enough with motivating what follows; whatever my particular reasons for pursuing it, a defence of the Kantian project in meta-ethics is surely of independent interest.

In §3.1 I discuss Egan's objection to Blackburn's version of quasi-realism which alleges that it is unable to make sense of our being in fundamental moral error. Having defended Egan's argument against Blackburn's response, I suggest that the best course is to bite the bullet. In §3.2 I expand on themes from Chapter 1 to develop the normative regress argument which threatens us with normative scepticism. In §3.3, §3.4, §3.5 and §3.6 I propose a Kantian solution to this problem, explaining and defending different parts of the solution in those each of those sections; overall it takes the form of a transcendental argument that establishes universalizability. In §3.7 I consider whether, as such a transcendental argument, the Kantian solution can really be successful.

3.1 FUNDAMENTAL ERROR

We begin our consideration of Blackburn with the defence of quasi-realism which he offers in his 1999 article 'Is Objective Moral Justification Possible on a Quasi-realist Foundation?'. Here Blackburn rightly insists that the quasi-realist is not committed to relativism simply in virtue of not being a realist. It is true that the main difference between us and those we think in grave moral error (his example is the Taliban) is that we and they

[112] Although Hare acknowledged a large debt to Kant, and thought that his own use of universalizability was largely equivalent to Kant's (see e.g. Hare 1993), he also tended to assimilate Kant's stance on universalizability to his own. For example, he asserts (1999: 65) that Kant's solution to the problem of objectivity in ethics is found in the logical properties of moral language. On the other hand, he admits (1999: 123) that although in *Groundwork* I-II Kant is giving an analysis of morality and its structure, he is doing something more in *Groundwork* III, which goes beyond the metaphysics of morals (i.e. of ethical theory), and Hare agrees with Kant that 'there is more to be done' that just to give the analysis. So it may be that Hare in the end agrees that his theory needs some equivalent of *Groundwork* III; and this chapter is my attempt to provide it.

have different attitudes. But there is no need to hold that there is nothing to choose between our attitudes and theirs. The following passage is crucial:

> And why does that [i.e. expressivism] not imply that divergent moral opinions are on all fours? Well, all I can hear that as meaning is that they are all *equally good*. And that is just not true. The Taliban's opinion on the education of women is not as good as mine. In fact, it is diametrically wrong, wrong root and branch. And notice that this would be true even if we were less minimalist than I have been about facts. Suppose a substantive or robust theory of truth were developed, giving us some notion of correspondence. Suppose it proceeds by isolating some metaphysical category of Facts (note the upper-case). And suppose finally that for the kinds of reason I have outlined, there are no normative or ethical Facts (all these doctrines belong to the *Tractatus Logico-Philosophicus*). This would be a metaphysical result. So it clearly could not imply that all moral opinions are on all fours. It could not imply, for instance, that it is permissible to hold that women should not be educated. It could at best imply that, in holding this, you do not trespass against the upper-case Facts. But that is all right. It was not *that* (or, not simply that) that is wrong with the Taliban view. The main thing that is wrong with the view is that it is inhumane, cruel, arbitrary, and so on. The metaphysics cannot imply that it is all right to be like that! [Blackburn 1999: 217]

Blackburn rightly insists that quasi-realists are perfectly entitled to talk of the correctness and incorrectness of moral opinions, and of facts as justifying some opinions rather than others. In saying all of these things they are (on the quasi-realist view) expressing further attitudes (higher-order ones). The only reason to doubt that having such attitudes amounts to holding that those opinions are **really** correct or justified is the assumption that normative judgement could not be a matter of attitude, which would beg the question against quasi-realism. So there is nothing relativist about quasi-realism. Indeed

there is a hint towards the end of the passage that it is realism which is confused on this point: if realism takes the justification of a moral opinion to consist ultimately in correspondence with facts, then this risks a disconnect with the things of genuine moral importance which figure in genuine moral justification. This relates back to the criticism of realism which I made in Chapter 1: that to think of a fact as being relevant to moral justification presupposes a normative judgement concerning the justificatory power of facts of that kind.

What Blackburn establishes is the right of expressivists (or at least quasi-realist expressivists) to talk of moral justification; there is nothing in expressivism which entails that moral judgements cannot be justified. But this still falls short of an account of how moral judgements are justified. There are two powerful arguments which can be levelled against Blackburn's position: the argument from fundamental moral error, and the normative regress argument. The upshot of these arguments (and a common worry lies behind both) is that expressivism is compatible with objective moral justification, but only as Kantian expressivism; what gets Blackburn into trouble is not expressivism but Humeanism about practical reason. There is an immediate reply that may be tempting to Humeans like Blackburn even before hearing the arguments, which is that quasi-realism is not in the business of saying what justifies moral judgements, since that is a normative question rather than a metaethical one. For instance, opposition to torture is justified *inter alia* by the pain which torture causes, but this is no part of quasi-realism. The reason that the reply is unsatisfactory is that the objections we will consider challenge the right of quasi-realists to say anything of that form at all.

The objection from fundamental moral error is formulated by Egan (2007), and I follow his account of the challenge below. In order to explain the problem with

fundamental moral error, Egan first recounts Blackburn's explanation of how quasi-realists can talk about the possibility of ordinary moral error.[113] It initially seems difficult for quasi-realists to account for claims like 'I believe that stealing is wrong, but I could be mistaken.' If the first part expresses disapproval of stealing, what does the second part express (it had better not simply be the opposite attitude)? Blackburn has a response to this worry:

> The problem comes with thinking of myself… that I may be mistaken. How can I make sense of my own fears of fallibility? Well, there are a number of things that I admire: for instance, information, sensitivity, maturity, imagination, coherence. I know that other people show defects in these respects, and that these defects lead to bad opinions. But can I exempt myself from the same possibility? Of course not (that would be unpardonably smug). So I can think that perhaps some of my opinions are due to defects of information, sensitivity, maturity, imagination, and coherence. If I really set out to investigate whether this is true, I stand on one part of the (Neurath) boat and inspect the others. [Blackburn 1998a: 318]

The thought is that I might in future change my mind about the wrongness of stealing, and this change might be a result of applying my other norms to my moral reasoning, in which case I would count the change as an improvement. So the 'I could be mistaken,' expresses indecision about whether to count a putative change in my opinion as an improvement; in other words (applying an expressivist account of improvement) indecision about whether to approve of such a putative change.

So, as Egan concedes, the quasi-realist has a plausible account of moral error in such cases. But is not it also possible to worry that I am more fundamentally mistaken: that although there is no change in my moral judgement of a case that I would recognise

[113] This discussion will be familiar from §2.4 above. Note that Smith's certitude problem discussed there was distinct from Egan's worry here, because the latter concerns **fundamental** moral error. Egan effectively concedes that Smith's problem is soluble, so might agree with what is said in §2.4.

as an improvement, this is only because I am also wrong about what counts as an improvement? The quasi-realist might be tempted to simply deny the coherence of this thought. To guard against this, however, Egan strengthens the objection by pointing out that it seems possible for moral disagreement to be fundamental: two people may hold conflicting moral opinions which are stable, in the sense that they hold background norms such that neither would count a change as an improvement.[114] If we hold, as a quasi-realist must, that when there is moral disagreement at least one side is wrong, we must also acknowledge that a case of stable disagreement is a case of fundamental error on at least one side. So if I am party to such a stable disagreement, it seems that it would be 'unpardonably smug' (Blackburn's phrase above) to discount the possibility that the fundamental error is on my side. And yet the quasi-realist has no way of interpreting the thought that I am fundamentally morally mistaken, because Blackburn's proposal for dealing with moral error is that I am morally mistaken when I could recognise a change in the putatively mistaken belief as an improvement; and this is precisely what I cannot do in the case of fundamental error.

Blackburn (2009: 205-7) responds by arguing that Egan confuses the following two principles:

(M) If something is entrenched in my outlook, in such a way that nothing I could recognize as an improvement would undermine it, then it is true.

(I) If something is entrenched in my outlook, in such a way that nothing that is an improvement would undermine it, then it is true.

[114] Note that in this sense of 'fundamental', the disagreement may not be over what either side thinks of as most important. There might be fundamental disagreements between two social democrats, as well as between a social democrat and a Taliban. And similarly, fundamental error might not be error that has wide-ranging practical consequences, but simply error that is stable in light of the agent's norms of improvement.

According to Blackburn, he is only committed to (I), whereas Egan's argument depends on attributing (M) to him. What Blackburn is here emphasising is that the account of error extends to the second-order question of what counts as an improvement (as I emphasised in §2.4): my judgements about what counts as an improvement are each themselves susceptible to change, and such changes may be ones that I count as improving (in light of my other judgements). Since agents are thus open to the possibility that they are in error about what counts as an improvement, they can make the conceptual distinction between their recognising something as an improvement and it being an improvement, and hence the commitment to (I) does not entail commitment to (M).

We can bring out Blackburn's point with respect to stability as follows: to say that a judgement of mine *J* is stable is just to say that it would survive anything that I **already** recognise as an improving change. But that leaves it open that my judgements about what counts as an improving change are themselves subject to a change I already recognise as improving, and after that change there is some change which *J* would not survive and which is **then** recognised as improving (though it is not **now** so recognised). Now we can define some terms: let us say that a judgement *J* is unstable$_1$ if it is not stable in Egan's sense, i.e. if there is some change already recognised as improving that *J* would not survive. And let us say that *J* is unstable$_{n+1}$ if there is some change already recognised as improving after which *J* would be unstable$_n$. What Blackburn points out is that I can see a way in which my judgements might be in error even if they are not unstable$_1$: they might be unstable$_2$. And indeed there is no reason to stop there: even my judgements that are not unstable$_1$ or unstable$_2$ or … or unstable$_k$ (where *k* is large) might still be unstable$_{k+1}$, meaning that I can still imagine being wrong. Admittedly the higher *k* is the harder the error will be to detect, but that is of no concern to the quasi-realist because all that is

required to refute Egan's argument is that the quasi-realist can **conceive** that even her stable beliefs are wrong.

What Blackburn misses here is that we can conceptualise a more radical kind of stability: let us say that a judgement is super-stable when it is not unstable$_n$ for any *n*. Admittedly it is considerably less likely that any particular judgement of mine is super-stable than that it is stable, and perhaps it is unlikely that any of my (or anyone's) judgements are super-stable. Even so, as it is at least conceivable that some of my judgements are super-stable, Egan's objection can still be formulated in terms of super-stability. Since the quasi-realist (at least in Blackburn's presentation) imagines being in moral error by imagining her moral judgement being unstable$_n$ for some *n*, she has no way of conceiving that any of her (*de dicto*) super-stable judgements are in error. The moral of this discussion is that Egan only went wrong in defining fundamental error as error in one's **stable** beliefs; that kind of error is not as fundamental as error in one's **super-stable** beliefs, and it the latter, genuinely fundamental error which the quasi-realist cannot conceptualise in the first-person case. It seems that if the quasi-realist imagines a moral judgement of hers to be super-stable, and a conflicting moral judgement made by someone else also to be super-stable, she cannot conceive that the error lies on her side, and this does indeed look unpardonably smug. If it really is fatal to quasi-realism if the quasi-realist cannot imagine herself to be in fundamental moral error, then the patched-up version of Egan's argument poses a serious problem. But as I shall argue, the problem is not for quasi-realism itself, but for Blackburn's Humean version, which cannot bite the bullet of fundamental *a priori* knowledge.

3.2 NORMATIVE REGRESS

Egan takes the foregoing argument to be decisive against quasi-realism, because he cannot see an alternative quasi-realist account of moral error that copes with fundamental error (and neither can I). I want to draw a more subtle moral: Blackburn's position stands refuted, but not quasi-realism, because Kantian quasi-realists do not have to count the rejection of fundamental error in one's own case as unpardonably smug. To understand the difference between the Kantian and Humean views we need to investigate the second argument against Blackburn: the normative regress argument. This is suggested by Cudworth's argument against voluntarism that we met in Chapter 1, and its generalisation as an argument against realism and for expressivism. There the point was to show that there cannot be a metaphysical explanation of morality; but the same style of argument threatens to show that no norms can be justified at all. It is not clear exactly who first came up with the regress argument; as I tried to show earlier, its roots lie with Socrates and Cudworth, though it is given more explicitly by Copp (1995: 37-49) and Cohen (2003); it is likely one of those arguments that periodically gets re-invented. I believe that Kant had something like this argument in mind when he claimed that morality depended on there being a categorical imperative. The sceptical form of the argument goes as follows: suppose that some norm *N* is justified; we can ask what justifies it; but whatever the answer, it will presuppose that the fact cited justifies *N*, which is a further normative claim (since claims about justification are normative); so every justified norm depends on some further norm for its justification; thus there is a vicious regress of normative justification, and so no norms are justified.[115]

[115] A point about terminology: where I use the term 'norm', Copp uses 'standard' and Cohen uses 'principle'. I am confident that we are all talking about the same thing, and that the difference is merely verbal.

Let us proceed more slowly: before considering the regress, I will first establish that every norm depends on another norm for its explanation, by an argument I call the normative relevance argument.[116]

(1) Some normative fact *N* is explained by a set of non-normative facts *F*.[117] (Supposition)

(2) Only if *F* is normatively relevant in the right way can *F* explain *N*.

(3) That *F* is normatively relevant (in the given way) is itself a normative fact.

(4) The normative fact that *F* is normatively relevant cannot itself be explained by *F*.

(5) So the explanation of *N* by *F* depends on a further normative fact.

(6) Since *N* and *F* are schematic, **any** explanation of a normative fact depends on a further normative fact.

For example, (1) suppose that the (normative) fact that an action A is right is explained by the (non-normative) fact that it maximises utility. (2) That explanation can only succeed if it is normatively relevant in the right way that A maximises utility. Now consider the fact that it is normatively relevant (in that way) that A maximises utility. Facts about normative relevance are normative facts, in that they are facts about what matters normatively[118], so (3) the fact that it is normatively relevant that A maximises utility is a normative fact. (4) That A maximises utility is not normatively relevant **because** A maximises utility.[119] (5) That A is right can only be explained by the fact that it maximises utility if that fact is

[116] Väyrynen (2013) discusses almost precisely this argument, crediting my formulation.

[117] To avoid begging the question by definition against the view that some normative facts are identical to non-normative, let us specify that a normative fact is simply a fact with a normative mode of presentation, and a non-normative fact is a fact with a non-normative mode of presentation.

[118] Someone who seriously doubts that facts about normative relevance are normative may be hard to convince, though it seems platitudinous to me. Here is one attempt at persuasion: in saying that a fact is normatively relevant (in some way), we are saying that it ought to be taken account of (in that way), which is a paradigmatic normative claim. Or in saying that it is normatively relevant we are saying that it is a normative reason, and a claim that something is a normative reason is a normative claim.

[119] Again, it is simply platitudinous that if there's a reason why a fact *F* matters, that reason cannot be *F* itself; an explanation of why *F* matters would have to be something **about** *F*.

normatively relevant in the right way, which is a further normative fact, meaning that the original explanation is not self-contained: that A is right is only explained if a further normative fact holds.

Väyrynen (2013: 160) worries that even if *F* only explains *N* under the further condition that *F* is normatively relevant in the right way, that need not mean that the further condition is **part** of the explanation of *N*. We could say that *F* is a complete explanation of *N*, but that a further explanatory question arises of why *F* can explain *N*, i.e. why *F* is normatively relevant in the right way. As Väyrynen (2013: 160) says, 'closing one explanatory question and opening a further explanatory question is not the same as regressing one and the same explanatory question.' But in fact the questions of what makes for a complete explanation, and what count as parts of an explanation, though interesting and controversial, are not crucial to the normative relevance argument. What matters is that insofar as an explanation only functions under a further condition, unless the condition holds the *explanandum* is left unexplained. Even if Väyrynen is right that there is no regress of a single explanatory question, there can still be a normative regress. If normative facts need to be explained, then there will be a regress, because *N* lacks a (successful) explanation unless *F* is normatively relevant, which requires that normative fact in turn to have an explanation. The appearance of regress is that unless and until the chain of normative facts bottoms out somewhere it looks as though it is still unresolved whether the original normative fact to be explained actually has a successful explanation. It is important to note here that the case of **normative** explanation is special, in that an explanation of why a normative fact holds is a justification of the norm specified by that fact. 'Why?' questions in ethics (and other normative domains) are asking for justification (e.g. 'Why shouldn't I lie?'), so since 'Why?' questions always ask for explanation, we can

see that normative explanation coincides with justification. An unexplained normative fact means an unjustified norm, so a regress of normative explanation is also a regress of normative justification: it threatens the conclusion that no norms are justified.

The challenge for Blackburn and other Humeans is to resist this regress. There are a number of options to consider. We might reject the idea that our norms need to be justified, perhaps by equating it with the Kantian idea (which Blackburn 1999: 226 rejects) that in order to be justified in applying a norm we must be able to justify it to those against whom it is applied. But this would be a conflation, because it is not justification **to others** which is at issue here. The regress challenges our ability to justify norms to ourselves. Alternatively we might hold out for some kind of coherentist solution to the regress, as is suggested by Blackburn's use of the Neurath boat analogy. The thought would be that we justify each of our norms by relying on some of the others, but that it is acceptable for chains of justification to be circular, so that at least one norm depends for its justification on a norm that it is ultimately used to justify. The problem now is very similar to the earlier difficulty with fundamental error: it is possible for there to be two conflicting sets of norms, each of which is coherent (i.e. it is possible within each set to justify each of the norms by relying on the others). If there are two such conflicting but coherent sets, then only one of them can be justified, which means that coherence is not sufficient for justification. Moreover, it is then necessary to say what makes one of the sets justified and the other not. Blackburn entertains the possibility that the Taliban have such a coherent set of norms, but that our norms are better than theirs.[120] He implicitly holds that the

[120] It is tempting to try to avoid the problem by denying the coherence of the Taliban's norms. But nothing here hangs on whether the Taliban's norms in particular are coherent; all that is required for the argument to go through is that there be at least two coherent, but mutually inconsistent, sets of norms. And it can easily be seen that this is possible: let Set 1 consist of the two norms A (Drainpipes should be green and norm B is justified) and B (Norm A is justified), and let Set 2 consist of the two norms C (Drainpipes should be

Taliban's norms are worse than ours because they discriminate against women. This is doubtless correct, but it cannot be an ultimate explanation of the superiority of our norms, because it presupposes that norms ought to be neutral between the sexes, which is one of our norms. It would only be sufficient as a justification if coherence were enough; but if coherence were enough the Taliban's norms would also be justified, which is impossible. It seems that Blackburn's position only makes sense if one switches back and forth between accepting and rejecting a coherentist theory of justification.[121]

One alternative worth considering for Blackburn is the one that Copp endorses. Copp (1995: 42-4) holds that there is a regress, but that it is not vicious. His view is that all the regress argument shows is that if there are any justified norms, then there are infinitely many justified norms (so that when we go up the chain of justification we never reach a final norm). It is only if we assume that there cannot be infinitely many justified norms that the argument can lead to scepticism. Copp notes that he does not mean that

---

red and norm D is justified) and D (Norm C is justified). Sets 1 and 2 are both coherent but that does not make it plausible that either is justified.

[121] See also Copp's (1995: 41-2) discussion of coherentism. Talk of coherence here may seem to strengthen an analogy between the normative regress argument and the famous regress argument concerning epistemic justification which is used to justify foundationalism. That argument is roughly that in order to be justified in believing a proposition on the basis of some evidence one has also to be justified in believing that that evidence adequately supports that proposition; so either some beliefs are foundationally (rather than inferentially) justified, or else there is a vicious regress and no beliefs are inferentially justified. Although this analogy is strong, with the Kantian position I defend being an unorthodox kind of foundationalism, and the solutions to the infinite regress I reject being similar to epistemic coherentism and Klein's (1998) infinitism, there are two ways in which the strength of the analogy might be overestimated. The first would be to simply assimilate justification as a status of a norm with justification as a status of a belief (that the norm holds). These are not at all the same: it is (at least conceptually) possible that a norm is justified though no-one is justified in believing that, or that someone is justified in believing (e.g. by testimony) that a norm holds though it does not. The question of whether a norm is justified is not an epistemic question, and thus the normative regress argument is not itself an epistemic regress argument. The second mistake would be to think that the solutions to the two regress arguments must be the same, or that if the normative regress argument demands a Kantian solution the epistemic regress must do too. I do not rule out a Kantian approach to the epistemic regress, but it may be unnecessary; that is because the epistemic regress is really a restricted version of the normative regress, in that it focuses on epistemic norms, which are a subset of norms in general. Because the question of epistemic justification is in this sense local, it might be settled by reference to other norms. Another reason why the epistemic regress might be more tractable is that coherentism is more of an option, because it is not obviously wrong to say that two conflicting (coherent) sets of beliefs are both justified, whereas conflicting sets of norms cannot both be (all things considered) justified.

all infinite chains of norms (with a hierarchy of justification) are justified; that would be objectionable for the same reason that coherentism is: 'But the point I am insisting on is that the fact that a standard has a place in an infinite hierarchy of standards is *not* sufficient to show it *not* to be justified' (Copp 1995: 43). It is worth noting that such a view has uncomfortable epistemic consequences: we surely do not grasp an infinite hierarchy of norms, and thus we never grasp the full justification for any of our norms. But such epistemological worries are secondary here, for there are more basic problems. Suppose that we are in possession of norms which do indeed form part of an infinite hierarchy of justified norms. There will presumably be other infinite hierarchies of unjustified norms; suppose that the norms of the Taliban form part of such an unjustified hierarchy. In virtue of what could our hierarchy be the justified one (and the Taliban one unjustified)? What reason could we have for thinking that our norms are justified? If some infinite hierarchies have a special ingredient that others lack which makes them justified, as it seems they must, then this ingredient must enter into the story at some point, and it seems needlessly obtuse and of doubtful coherence to defer this point to infinity.[122]

The regress can only be defeated if there is some way for norms to get justified which does not presuppose prior norms. Copp's solution needs this to distinguish between justified and unjustified infinite hierarchies; but once it is in play there is no need to grant the regress at all. The point against Copp can be restated slightly differently: suppose that there are two infinite hierarchies of norms, of which one is justified and the other not; now there must be a reason why the justified one is justified (something which

[122] Again, it matters not at all whether there could be an infinite hierarchy including the Taliban norms; what matters is that there can be two infinite hierarchies which conflict. Let Set 1 consist of the norms A (Drainpipes should be green), B (Norm A is justified), C (Norm B is justified) and so on; and Set 2 be structurally similar but with the first norm stating that drainpipes should be green. Sets 1 and 2 are both infinite hierarchies, but not plausibly justified.

differentiates the two hierarchies); and so there must be a further norm to explain why that reason is a reason; that norm cannot be justified simply by its place in the justified hierarchy, because the unjustified hierarchy may contain a contrary norm, according to which it is justified, and thus justificatory symmetry between the hierarchies would be preserved; and so the further norm must require further justification, and the regress continues. The upshot is that even if Copp is right that there is nothing special to prevent infinite hierarchies from being justified, they are of no use for stopping the regress.[123]

Radzik (2000) suggests a way of solving the regress that is vulnerable to a similar criticism: she holds that although some norms cannot justify themselves, some can, and thus stop the regress. There are indeed many norms which **purport** to justify themselves, in that they are justified according to themselves; call such norms 'self-supporting'. What Radzik is after is a norm that genuinely provides itself with justification, so that it really is justified in virtue of being self-supporting. But just as we ask what the difference is between the infinite justified hierarchy and the infinite unjustified hierarchy, so we must ask what differentiates the justified self-supporting norm from the unjustified self-supporting norm (because, as before, there could be two self-supporting norms which conflict with each other), and there will be no satisfactory answer to this question. Radzik

[123] Radzik (1999) agrees that Copp's infinitary solution to the regress does not work, but for rather different reasons, which I disagree with. She holds, against Copp, that the regress argument relies on a first-person conception of justification, according to which agents must have access to their reasons in order to be justified. She claims that an infinite chain of justified norms would flout this requirement, because agents cannot use infinite chains of reasoning when they deliberate. I agree with Copp, against Radzik, that the regress does not depend on the first-person conception. The argument is that in order for a norm *N* to be justified, there must be some reason why it is justified; but whatever that reason is, it seems that there must be further norm to explain why the cited reason is a justification for *N*. That argument concerns what reasons there are, not what reasons are accessible to an agent. My objection to Copp is thus dialectically stronger than Radzik's, because I do not rely on her dubious assumption of a first-person conception of justification, and so grant more to Copp. My version of the regress argument is effective whether or not we accept this assumption.

could hardly say that it is just a basic fact that particular self-supporting norms are self-justifying; that would be simply dodging the normative question, rather than answering it.

It is tempting to follow Schroeder's (2005) response to Cudworth's original argument (against voluntarism) in thinking that the regress can be stopped via constitutive explanations, as discussed in §1.1. As Schroeder rightly insists, some explanations terminate in identity facts. 'Why are Clark Kent and Superman never seen together (as two people)?' – 'Because Clark Kent **is** Superman!' And Schroeder is also right to say that such identity facts do not themselves require explanation. Although *a=b* is more informative than *a=a*, as far as the world is concerned the former requires no more than the latter. To ask for an explanation of why *a=b* is therefore effectively to demand an explanation of why *a=a*, which is absurd. Schroeder's strategy then is the naturalist one: some normative properties are natural properties (and so non-normative in the sense of having a non-normative mode of presentation), and so some normative facts are identical to non-normative facts (they are identical facts with both modes of presentation).

We have seen in §1.2 that naturalist versions of realism are untenable because of the Open Question Argument. So the question of whether Schroeder's solution to the regress would work for the naturalist is moot. But it may seem as though even if we are expressivists it is still possible to run Schroeder's line: after all, it was argued in §1.3 that expressivism is at least compatible with the metaphysics of naturalism, so that the expressivist can accept identities between normative and non-normative facts. But, perhaps surprisingly, Schroeder is wrong that the relevant identities do not require explanation. Suppose, for example, that the fact that *A* is right is identical to the fact that *A* maximises utility. Presumably fact-identities are to be explained in terms of the facts having the same constituents and structure (assuming that facts are made of individuals

and properties/relations), so what is more fundamental here is that rightness is identical to maximising utility. But what it is for rightness to maximise utility is no more than for 'is right' and 'maximises utility' to be necessarily co-extensive[124]; and that in turn is no more than for utilitarianism to be (necessarily) true.[125] It turns out then that to follow Schroeder's line is to think that the normative regress is stopped by, and normative explanation/justification bottoms out in, the truth of utilitarianism or some other such ethical theory. But if anything stands in need of justification, it is an ethical theory such as utilitarianism![126]

There are three further possible responses to the normative regress argument which should be ruled out. One is to say that some norms are simply obviously justified, and so need no further justification. But it is unclear how this avoids the central problem avoiding all other accounts: it seems quite possible that two conflicting norms will both seem obvious, and thus seeming obviousness cannot be sufficient for justification. If the reply to this is that we are not talking of apparent obviousness but rather **genuine** obviousness, then the account becomes question-begging, because a norm cannot be genuinely obvious without being justified, and we are entitled to ask for an account of genuine obviousness, which will be no easier to come by than an account of justification itself.

A second account seems related: it is a kind of primitivism about justification, according to which it is simply a brute fact whether a norm is justified or not. Heathwood

[124] At least that is all it could mean for expressivists, see §1.3. Again, the question of whether realists could mean something else by property identity here is not worth pursuing, not only because it is very difficult to fathom what it would be, but also because realism has already been ruled out.

[125] This is just another example of the point discussed in §1.6 that many apparently metaphysical claims about morality are really internal and so themselves normative. See e.g. Dworkin 1996: 100-1.

[126] Of course we could at this point add that it is a brute fact that e.g. utilitarianism is true. See below for a reply to this line.

(2012) argues that, as a consequence of an argument similar to the normative regress, morality cannot be grounded in any non-moral facts, and so that it must be brute which fundamental norms hold: morality cannot have a source. My view is that Heathwood is wrong to assume that the only kind of source for morality would lie in non-moral facts. He thinks this because he incorrectly assumes that constructivism tries to ground morality in non-moral facts; I will discuss Heathwood's concerns about constructivism later when I have explained what I take constructivism to involve (§3.4-3.6). Regardless of whether Heathwood misdescribes the options, his overall point is fairly clear: brute justification is what remains when every other possible source of justification has been eliminated.

The crucial point concerning such a view is that it is not compatible with expressivism. The business of this chapter is to show that Blackburn's version of expressivism is inadequate, and that a different, more Kantian, version of expressivism is needed. It is only to be expected that some will take the argument as showing what is wrong with expressivism itself, and so retreat to some variety of realism. I have tried to explain in Chapter 1 why realism is so unfruitful as an answer to the explanatory question. What we have here is the attempt to use realism to answer the normative question; but note that a realist answer to the normative question cannot make sense unless realism can also answer the explanatory question.[127] It is no good to say that the fact that some norms are justified is simply bedrock in answering the normative question unless you also have some account of what you are talking and thinking about when saying and judging that norms are justified. So unless realism can be defended against the objections of Chapter 1 (which I shall not repeat), it is out of order to attempt to stop the normative regress by

[127] The idea that there are explanatory and normative questions in meta-ethics is from Korsgaard (1996). The idea is that the explanatory question concerns the nature of normative judgement and its place within a causal/explanatory scientific world-view, whereas the normative question concerns the justification of norms. For more on this see §3.4.

realist foot-stamping. So we should not see the elimination of the other options as leading to brute justification; it is instead more natural to think that such a surrender in the face of the normative regress argument leads to scepticism.

A final attempt is to claim that 'the normative question presupposes, incorrectly, that it makes sense to morally evaluate moral values' (Curry 2005: 169).[128] Certainly the normative question does presuppose this, but why think it incorrect? One reason would be despair at the regress. Another would be the thought that since expressivists have given a full explanation of human moral judgement, there is no further question to ask about what really is good or right. According to Curry, a separate argument is required to show that there is anything to get right in the realm of normative judgement. One point against this which can be made immediately is that we actually do make moral evaluations of moral values, so to say that this does not make sense is to put forward an error theory of our moral thinking, and to give a sceptical answer to at least second-order normative questions. And it seems that such scepticism is bound to seep downwards: if it makes no sense for me to think that there is anything wrong with moral values which permit torture, then it is unclear how I can think that there is anything wrong with torture itself. Part of Kant's argument, which I will explain in more detail in §3.5, is that this kind of normative scepticism is seriously unattractive, indeed untenable. Finally, the point of Kant's transcendental argument is precisely to show how there can be an objective answer to normative questions in a way that fits in with expressivism; and that is a task for the remainder of this chapter.

[128] Note that Blackburn does not take this line – insofar as Blackburn and Curry are both broadly Humean, there is a split in the Humean camp on whether answering explanatory questions makes it unnecessary or senseless to try to answer normative questions. As we have seen, what Blackburn objects to is not normative questions, not even normative questions about moral values, but Korsgaard's idea that there is a single normative question.

At this point it is worth noting a difference between Copp's version of the regress argument and Cohen's (2005). For Cohen, the argument is meant to show that fundamental norms cannot be fact-sensitive.[129] The structure of Cohen's regress is an alternation between norms and facts: facts justify norms, and higher-level norms justify the ability of facts to justify lower-level norms, but these higher-level norms depend on further facts etc. That is why Cohen can say that if norms can be justified in some other way then the regress ends, and normative scepticism does not result. But for Copp the regress is meant to cover all possible ways in which norms might be justified. His thought is that simply by considering whether a norm *N* is justified we invoke a further norm to settle the question; that further norm might make whether *N* is justified depend on the facts, or it might not; that is not Copp's concern. There is a simple way to extend Cohen's argument so that it is closer to Copp's: substitute 'reason' for 'fact'. Sometimes the reason why a norm *N* is justified is some fact, but the next step in the argument does not depend on that; because whatever the reason is, there is always a further question of why it is a reason for *N*, and that presupposes a further norm. So if Cohen's argument establishes that fundamental norms must be fact-insensitive, it also seems to show that they are reason-insensitive. But it is harder to make sense of the idea of reason-insensitive norms, and thus normative scepticism again looms large. We seem to need a reason-insensitive norm to stop the regress, and that is of dubious intelligibility.

[129] Cohen's agenda is to show that Rawls (1999: 398) is wrong to think that principles are ultimately grounded in facts. This focus on Rawls means that Cohen is less interested in pursuing the implications of his argument for meta-ethics more generally.

3.3 A KANTIAN SOLUTION?

As I have already intimated, my view is that the normative regress can only be stopped in a Kantian fashion: the thought is that the idea of a reason-insensitive norm is the same as Kant's idea of a categorical imperative. Before I go further with my exegesis of Kant, it is worth briefly stating what I aim to achieve therein. The primary goal here is to argue for Kant's Formula of Universal Law (FUL)[130]; insofar as I present arguments for it which I attribute to Kant, it is more important that those arguments are successful than that they are really Kant's. I do not have space here to present the detailed textual evidence which would be necessary to demonstrate the superiority of my interpretation to that of contemporary Kantians. The main evidence that I present here in favour of my interpretation is simply the reasoning for the success of the arguments I attribute to Kant; it is better, other things being equal, to interpret Kant as giving good arguments rather than bad ones (as we shall see, leading Kant scholars admit that Kant's arguments are fallacious on their interpretations). Why then, if scholarship is not my primary goal, do I bother to say that these are Kant's arguments, rather than presenting them as my own? The reason is that these are the arguments which I find in the *Groundwork*; it would be dishonest to claim to have invented them myself.

With these disclaimers, I will begin by sketching the stages of Kant's main argument in the *Groundwork*. I take it that in Section I Kant is trying to show how his theory fits in with common moral concepts, and that the argument only really gets started in Section II (first a description of what we implicitly think, and then an explanation of why it makes sense to think that). There are then four important parts to the argument. Firstly, Kant

[130] *Groundwork* 4: 421 (Kant 1996/1785: 73). Henceforth I shall omit author/date and page number references to this text, and simply refer to it as *'Groundwork'*, giving the standard Akademie pagination.

argues that there must be a categorical imperative (4:406-417). Secondly, he argues that the content of a categorical imperative must be deducible from its form, and thus that there is a single categorical imperative which can be stated as FUL (4:417-421). Thirdly, he explains how to apply FUL to derive directly action-guiding principles (4:422-424). Fourthly (this is now Section III), he argues that the Categorical Imperative (CI) is binding on all rational beings because rational beings have to take themselves to be free (4:446-463). There is, of course, substantial material at the end of Section II, dealing with the Formula of Humanity (FH) and autonomy; much of what is said there is relevant, but these passages are largely inessential to the central thrust of the argument. The current fashion[131] for emphasising FH over FUL seems to me to make little sense of the text, and is perhaps motivated by the mistaken view that Kant's argument for FUL is fallacious.

Of these four elements in Kant's argument, I will defend the first, second and fourth. I postpone discussion of the third until the next chapter, but I will reject it. My rejection of Kant's application of FUL is not intended to rely on the arguments or interpretation which I give in this chapter; indeed there is no reason why those who disagree with me about the argument for FUL should not agree with me about its application, and vice-versa. On the other hand, the arguments of this chapter are not independent of the arguments given in the previous chapters for expressivism. I thus perceive a kind of independence between two parts of an overall argumentative strategy to establish an ethical theory; but it is not the traditional gap between meta-ethics and normative ethics, unless FUL is counted as part of meta-ethics. Rather the distinction is between the formal and substantive parts of ethics. The standard view that meta-ethics and normative ethics are independent is embraced by Blackburn, who holds that quasi-

[131] E.g. Wood (1999).

realism does not commit one to any view in normative ethics. This contrasts with the position which I argue for, and which Hare (e.g. 1999: 107) held, that universalizability is not an optional extra for expressivists and that meta-ethics and normative ethics are not fully independent; my arguments in this chapter are meant to show that expressivism is only defensible if one also accepts FUL.[132]

I return now to the puzzle presented to us by the normative regress argument; it seems that the regress can only be stopped if there is a norm (at least one) which is not sensitive to reasons. I now wish to point out that Kant is worried by exactly the same problem. First, he accepts what looks very much like the normative regress argument, or at least its crucial step[133]:

> Nor could one give worse advice to morality than by wanting to derive it from examples. For, every example of it represented to me must itself first be appraised in accordance with principles of morality, as to whether it is also worthy to serve as an original example, that is, as a model; it can by no means authoritatively provide the concept of morality. [Kant *Groundwork* 4:408]

Notice here the point that the use of an example would presuppose a prior norm. And what he says follows is that 'there is, then, no genuine supreme basic principle of morality that does not have to rest on pure reason independently of experience.' (4:409) Later we are told that this is what is meant by a 'categorical' imperative: 'if the action is represented

[132] Note that it is the combination of expressivism and genuine objectivity which entails universalizability for ethics. Since humour and aesthetics, where expressivism is also correct, are not objective, a non-universalizable discourse for those areas might make sense. That is not to deny that actual aesthetic discourse involves universalizability; the point is that even if aesthetics is universalizable, shmaesthetics might not be, whereas even shmethics would have to be universalizable if it was still meant to be objective.

[133] It is not entirely clear from the next two passages that Kant endorses the whole of the normative regress argument. But he is committed to the two premises of that argument: that the justification of a norm by a fact or a reason presupposes a further norm, and that fundamental norms (such as 'the imperative of morality') cannot be based on further normative presuppositions. Since putting these two premises together gives us an argument to the conclusion that either there is an imperative which is somehow valid without there being any reason why (i.e. a categorical imperative), or else no norms are valid, and since this is the very conclusion which Kant wants, it is more than tempting to attribute the whole argument to him.

as *in itself* good, hence as necessary in a will in itself conforming to reason, as its principle, *then it is categorical*.' (4:414) And Kant is clearly worried by normative scepticism:

> On the other hand, the question of how the imperative of *morality* is possible is undoubtedly the only one needing a solution, since it is in no way hypothetical and the objectively represented necessity can therefore not be based on any presupposition, as in the case of hypothetical imperatives. Only we must never leave out of the account, here, that it cannot be made out *by means of any example*, and so empirically, whether there is any such imperative at all, but it is rather to be feared that all imperatives which seem to be categorical may yet in some hidden way be hypothetical. [Kant *Groundwork* 4:419]

Suppose that all imperatives were hypothetical, or in other words (going by the definition of 'categorical' which Kant gives above) that no actions were good in themselves, what would follow? As I interpret these passages, that would mean that the question of the goodness of an action could never be settled by just looking at the action; there would always have to be some reason why that action was good, which would depend on it belonging to a class of actions which are good. But that class of actions would also (*ex hypothesi*) not be good in themselves: their goodness would depend on belonging to some more general class of good actions, and so *ad infinitum*. (Why is ϕ-ing good? Because ϕ-ing is ψ-ing and ψ-ing is good. But why is ψ-ing good? Etc.) This is the very same normative regress as before. So it looks very much like Kant rejects the idea that all imperatives are hypothetical because he sees it as leading to a regress and thus normative scepticism, and this motivates him to give an account of how a categorical imperative is possible. Whether this is the correct interpretation of Kant is, however interesting, not the most important question for current purposes; for Kant seems to have a solution to the normative regress

very different from the ones canvassed and rejected above. What we need to investigate is whether Kant can come to the rescue of the expressivists, and in particular Hare.

3.4 THE CONTENT OF THE CATEGORICAL IMPERATIVE

Once we see the need for a categorical principle, two questions arise. First, what its content is (what it tells us to do), and second, whether it really exists (whether we are really bound by it, and how we are if we are). The latter task is, as I have already indicated, left until Section III of the *Groundwork*, but the former is to be dealt with immediately: 'In this task we want first to inquire whether the mere concept of a categorical imperative may not also provide its formula containing the proposition which alone can be a categorical imperative. For, how such an absolute command is possible, even if we know its tenor, will still require special and difficult toil, which however, we postpone to the last section.' (4:420) Kant's central argument over the content of the CI is very brief, and is not generally thought to be valid. The key idea is that the content of the CI must be derivable from its form alone, because there is nothing else to give it content. Here is the crucial passage:

> When I think of a *hypothetical* imperative in general I do not know beforehand what it will contain; I do not know this until I am given the condition. But when I think of a *categorical* imperative I know at once what it contains. For, since the imperative contains, beyond the law, only the necessity that the maxim be in conformity with this law, while the law contains no condition to which it would be limited, nothing is left with which the maxim of action is to conform but the universality of a law as such; and this conformity alone is what the imperative properly represents as necessary. There is, therefore, only a single categorical imperative and it is this: *act only in accordance with*

that maxim through which you can at the same time will that it become a universal law.

[Kant *Groundwork* 4:420-1]

The move which has seemed problematic here is from the requirement to conform your actions to universal law as such (CI), to the requirement to conform them to maxims which you can will as universal law (FUL).[134] Aune (1979: 29) complains that '[FUL] differs significantly from [CI] in practical import, at least, and the line of thought leading naturally to [CI] does not seem adequate to render [FUL] credible.' And Wood (1999: 81) claims that the 'fallacy in Kant's deduction' is that 'it does not follow from the mere concept of a categorical imperative that *the will of a rational being* – what a rational being wills or can consistently will – has any role to play in determining the content of universal laws.'[135] This criticism of Kant is not always taken to be decisive; Kitcher (2004) produces a convoluted explanation of how to bridge the gap, reconstructing a detailed argument from other bits of Kant. But my view is that the move is just as simple as Kant represents it as being, with no missing steps. The key is expressivism. For expressivists, normative judgement, including moral judgement, just is a kind of willing. So when I judge that my maxim of action conforms to universal law (i.e. I judge that it conforms to a norm which is universally valid), I must will that that maxim be followed universally.

Let us expand on how expressivism helps Kant out here. For expressivists, all normative judgements are attitudes (i.e. willings). But we are here considering a special kind of normative judgement: one that is not meant to depend on anything else. This

[134] The objection can be found as far back as Schopenhauer (1965/1841: §II.7), who wonders what bearing what we will could have on the CI.

[135] Orthodoxy holds that Kant's real (i.e. valid) argument for the content of the CI comes later, when he discusses autonomy; I have even heard it said that Kant does not mean to be providing a derivation of FUL from CI at all. This seems to me to be a case of wishful thinking: if you think that Kant's argument for FUL is embarrassingly bad (as most Kantians do), it seems charitable to interpret him as not making that argument at all, and to ignore his talk of whether 'the mere concept of a categorical imperative may not also provide its formula'.

means that such judgements have to be universal, since there is nothing to restrict their scope: a non-universal judgement would be one that was restricted in some way, and would thus presuppose the normative relevance of that restriction. So to judge that a maxim is categorically (i.e. non-derivatively) correct is to will it universally. Now the final step to FUL is trivial, for we can safely assume that we are obliged to act in accordance with maxims which we judge to be correct. Of course, we may make incorrect moral judgements, but this is no objection to the deliberative principle which tells us to choose actions which conform to our best judgement. The point is that to make a normative practical judgement is already to judge that one should act in accordance with it; in willing that everyone follow a maxim, *a fortiori* one wills that one follow the maxim oneself.

Wood is mystified by the idea that what we will could help to determine the content of the CI; but it is not mysterious how the content of the CI can be connected to what correct normative judgement amounts to, and it is not mysterious (given expressivism) how normative judgement can be a matter of willing. Given the simplicity of this reconstruction of Kant's argument, an argument that is usually considered to be obviously invalid or even non-existent, it is surprising that the view that Kant is an expressivist is not more mainstream. I am afraid to say that the likeliest explanation is to be found in the misunderstanding of expressivism amongst Kant scholars. It cannot be said that the expressivist interpretation is ignored because it has not been prominently suggested: Hare repeatedly insisted that Kant was a prescriptivist, and thus in contemporary terms an expressivist.[136] Indeed Hare (1963: 34) hints at the very same explanation of FUL as that given above. And indeed Wood for one is not ignorant of Hare's position, though he does not acknowledge, or does not realise, that it provides a solution

[136] Hare makes this point from the beginning of his writing on prescriptivism (e.g. 1952: 16), and insists on it even more later on (e.g. 1993: 12).

to the problem which he sees for Kant. Unfortunately, Wood is dismissive of Hare's expressivist interpretation of Kant:

> Since Kant holds that moral truth is irreducible either to what people think or to the results of any verification procedures, he is a moral *realist* in the most agreed-upon sense that term has in contemporary metaphysics and meta-ethics. Kant is a moral realist because realism is the only way of preserving the *critical* stance necessary to all moral thinking, the open-endedness of moral inquiry. To say that the moral law rests on an *idea* is to say that it is always in principle possible for us to be mistaken about what we think is right, no matter who we are, how many of us there are, or what decision procedures we may have applied in arriving at our moral beliefs. [Wood 1999: 157-8]
>
> This makes it especially wrongheaded for R. M. Hare to associate Kantian autonomy with moral antirealism and the 'is-ought gap'. [Wood 1999: 375, note to the previous passage]

The problem with what Wood says is that he does not appear to know much about contemporary meta-ethics. Expressivists do not hold that moral truth is reducible to what people think or to the results of any verification procedures, or that we are immune to moral error, or that morality is mind-dependent in any other relevant way. It is not clear that even the earliest expressivists, like Ayer, held any of these views; certainly Hare did not, and, as discussed in Chapter 1, Blackburn has demonstrated how expressivists have a right to this kind of mind-independence. Regardless of whether Wood is correct to define realism with reference to these ideas of mind-independence (I think not), it is clear that he takes expressivism to be an anti-realist view, and so implicitly holds that expressivism is committed to one or other of the sins which he contrasts to realism. So Wood relies on an

understanding of expressivism which had been debunked by the time he was writing.[137] This means that his objections to Hare's interpretation have no force at all.

It is worth tracing the issue between Hare and Wood a little further, to discuss the passage from Hare which Wood quotes in the footnote cited above:

> The reason why heteronomous principles of morality are spurious is that from a series of indicative sentences about 'the character of any of its objects' no imperative sentence about what is to be done can be derived, and therefore no moral judgement can be derived from it either. [Hare 1952: 30]

Hare is not exactly trying to put forward an expressivist interpretation of Kant here. Rather, it seems to be the normative regress argument, or at least its crucial step, which is doing the work. The idea is that a heteronomous judgement is one that reads off a normative conclusion from some fact about 'the character of any of its objects' (Hare is referring here to Kant's definition of 'heteronomy' at 4: 441). But as was argued above, drawing a normative conclusion from some fact presupposes a further norm. This is why Kant says in that passage that 'this relation [of the object to the will], whether it rests upon inclination or upon representations of reason, lets only hypothetical imperatives become possible: I ought to do something *because I will something else*.' Hare says (to Wood's consternation) that Kant is here relying on the Humean doctrine that an 'ought' cannot be derived from an 'is'; and this should make us reflect again on the connection between expressivism and the normative regress argument. What I want to emphasise is that we

[137] This is one reason why I am prepared to commit myself to unorthodox readings of Kant: although the Kant scholars with whom I disagree have a vastly greater knowledge of Kant's writings, it is unlikely that this knowledge will give them a secure grasp of Kant's meta-ethical views unless they also understand what the meta-ethical options are. Wood's inaccurate pronouncements on meta-ethics do not inspire much confidence on that score. It is not clear to me how deep these misunderstandings run within Kant scholarship, though I fear that Korsgaard is an honourable exception and that Wood is more typical.

should not see the is-ought gap as a consequence of expressivism[138]; rather it is a datum of reflection on our normative concepts that norms are not derivable from facts alone. That is because any purported derivation would assume that the relevant facts really did justify the derived norm, and that would be a normative presupposition. As was argued in Chapter 1, expressivism is the view best able to account for this datum. And yet Wood apparently simply rejects the is-ought gap; I confess that this leaves me mystified as to how he can account for the various passages, such as the one which Hare notes, where Kant seems to rely on it.

Before leaving the topic of Kant and expressivism, I wish to draw attention to an obscure passage from Korsgaard:

> Expressivism, I believe, is like realism also true after all, and also in way that makes it boring. From the descriptive and explanatory perspective that is appropriate to scientific or perhaps in this case social-scientific inquiry, those who use normative language will appear to be simply expressing their values. When you are not in the grip of practical problems that provide standards for their own solutions, the truth and falsehood of statements employing concepts that embody those problems must be elusive. The trouble with expressivism is that it describes moral language from the outside, as if we were not ourselves the creatures who face practical problems, but only someone else making anthropological observations about them. Behind that stance is the idea that so long as we are reasoning we must remain at this anthropological level, and behind that view is the same error that animates moral realism – the view that the business of cognition is describing the world. [Korsgaard 2003: 122 n49]

There are a number of points to note here. Korsgaard holds that expressivism is true but boring. Thus she need not (unlike most contemporary Kantians) dissent from my view that

[138] As it would be if we took motivational internalism to be the basic argument for internalism, and deduced the is-ought gap subsequently, as Hume is often supposed to have done.

Kant is an expressivist; but she surely would disagree with my claim that expressivism is crucial to the derivation of FUL (that would hardly be boring). She also compares her attitude to expressivism to her attitude to realism: they both end up being true, without being the right places to start (2003: 118). Notice that as an attitude to realism, this is quite consistent with Blackburn's take on the matter. He too holds that moral claims end up being true or false, but that what is important is how they end up being so: we have to earn the right to talk of truth. What Korsgaard is adding here is that realism and expressivism are both partial views: each is a correct description of the way things appear from one vantage point, whereas the correct account, constructivism, is able to explain both of them. Whilst Blackburn could agree that expressivism is only a partial view (hence quasi-realism), he does seem committed to assigning a kind of priority to expressivism over realism. And this is the basis for Korsgaard's accusation that expressivists are biased in favour of the anthropological, naturalistic stance from which expressivism is true, and thus assimilate reasoning and cognition to describing the world.

As I will argue in §3.6, there is some justice to this charge. But perhaps the focus of the disagreement can best be seen by considering the two questions which Korsgaard (1996) discusses: the explanatory question, and the normative question. Whilst we are asking the explanatory question, the anthropological standpoint is bound to come to the fore. Korsgaard thinks, and I agree, that asking the normative question forces us towards a Kantian or constructivist view which operates mainly from the practical stance. The crucial difference in Blackburn's view is that he rejects the idea of a normative question as on a par with the explanatory question (see §3.5 below); for him, there are individual normative questions, but no single big question (whereas he does seem to assume a single explanatory question). This explains why the anthropological stance looms larger for

Blackburn, even though at least officially he would not hold that reasoning can only go on at that level: it is the one where the question he is interested in answering arises. Whether there is any real asymmetry is a more difficult question. The point to take from Korsgaard here is that whilst Kant is an expressivist, he is not **only** an expressivist: he is mostly interested in the normative question, and that is how Wood can be misled into thinking him a realist. In the end, I mean to vindicate the approach to meta-ethics which lies in Kant, and which Korsgaard expresses in the passage above. But, as I shall discuss in §3.7, Korsgaard's transcendental argument fails where Kant's succeeds.

3.5 AGENCY, NORMATIVITY AND SCEPTICISM

Having given these unorthodox interpretations both of Kant's argument for the necessity of a categorical imperative (as relying on the normative regress argument), and his argument for FUL (as relying on expressivism), we now face what appears to be an even harder challenge: explaining what Kant's argument in §III of the *Groundwork* is all about. The challenge is hard because the task which Kant sets himself seems impossible on the interpretations given above. What §III purports to show is how the CI is binding on us; but this sounds very much like an attempt to show that the CI is justified, which seems impossible given the normative regress argument. Recall that if we try to give a reason in support of a norm, we presuppose some other norm; but then we can give no reasons in support of the CI, since the whole point of the CI is that it does not depend on any other norm. Once we start to think in this way it may appear that the normative regress argument has led us down a blind alley: it told us that we needed a categorical norm, but it also prevents that norm from being justified, and thus seems to entail normative scepticism. This I think is a clue to what Kant is up to in §III: he is trying to show that even

if we cannot justify the CI, we can show that it has a status, other than being justified, which can allow us to retain it in a pivotal role.

The key to Kant's argument is the rejection of normative scepticism. Expressivism leaves us with three candidate accounts of normativity: Kantian, Humean and sceptical. The Kantian demands an answer to the normative question, and insists that it compels us to accept a norm which is the source of justification, but not itself justified (the CI). The Humean denies that there is any such single question:

> But why suppose that there is then such a thing as *the* normative question? People ask why they must do some particular thing in all kinds of circumstance, and their concerns can only be addressed in appropriately different ways. There is no more a single normative question than, for instance, a single emotional question, 'What am I to feel?'. [Blackburn 1998: 258]

The sceptic on the other hand simply concedes that no norms are justified, and thus tries to do without normative judgements (it would be Moore-paradoxical to accept a norm and also deny that that norm was justified). The normative regress argument presents us with a dilemma between the CI and normative scepticism, because it rules out the Humean position. Kantians need not assume that there is a single normative question.[139] Instead, we can see the normative regress argument as showing how by asking individual, particular normative questions, we can be brought to see that there is in fact one big question lying behind them all: how does normative justification get started in the first

[139] To be fair to Blackburn, in the above passage he is responding to Korsgaard (1996), and whilst it is reasonably clear that she has the normative regress argument at least partially in view (e.g. 1996: 30), she introduces the idea of a single normative question first, and does not use the regress argument to support it. The more significant difficulty with Korsgaard's approach, which is closely related to my own, is that she seems not to be fully convinced of the Kantian project, and ends up with a kind of compromise between Hume and Kant that leaves a crucial role to reflective endorsement in the fundamental justification of moral principles. But such a position is still vulnerable to the normative regress argument, which is a further reason for thinking that she does not fully grasp its importance.

place? Given that the Humean answer is off the table, if normative scepticism can also somehow be ruled out, then we would have to accept the CI. But how could normative scepticism be ruled out without giving a **justification** for the CI (which would involve us in the regress again)?

The answer is meant to be provided by a transcendental argument: normative anti-scepticism, and thus the CI, is a necessary condition for freedom. This is connected to Kant's emphasis of the fact that we are rational beings, and that the CI is binding for all rational beings. That does not mean that some other norm is binding on non-rational beings, and that thus the status of the CI is the same as that of the pure intuitions of space and time in the Critique of Pure Reason. Such an interpretation would be misguided, because what it is to be a rational being (at least for Kant) is to be subject to norms, to be interested in acting for good reasons. Thus no norms could apply to non-rational beings.[140] This then is the core of my interpretation of §III of the *Groundwork*: Kant holds that, *qua* rational beings, we are inescapably faced with a normative question whenever we deliberate – what do I have reason to do (or believe, in the case of theoretical deliberation)? If this is right, then what immediately follows is that we cannot accept normative scepticism.

This interpretation allows us to make sense of Kant's thought that there is a distinction between freedom and autonomy, in that freedom is possible for those who are not fully autonomous, but that those who are free (i.e. all rational beings) are still committed to autonomy. If to be free is to take oneself to act on the basis of reasons, then anyone making a practical normative judgement and acting on its basis is (negatively) free. But merely making a normative judgement without accepting the CI will end up being

[140] I have not seen the interpretation I criticise here defended in print, but I have heard Thomas Pogge put it forward in a lecture entitled 'Two Queries for Cohen', given on 21/04/07 at the University of Reading.

inconsistent (as the normative regress argument is meant to show). Since being autonomous simply involves deriving the norms on which one acts from the CI, this threat of inconsistency faced by those who are merely free can be understood as a pressure towards autonomy. When Kant argues for this transition from (in his terminology) negative to positive concepts of freedom (4:446), he claims that 'freedom, although it is not a property of the will in accordance with natural laws, is not for that reason lawless but must instead be a causality in accordance with immutable laws but of a special kind; for otherwise a free will would be an absurdity.' The thought is surely that in order for action to be intelligible it must be in accordance with principles, because otherwise it could not be explained, and thus that for free action norms must do the work of laws of nature. Now Kant moves immediately to saying that the law involved with freedom must be autonomy, and thus the CI (4:446-7), but we should not be concerned at this argument seeming to move too quickly, because Kant already argued in §II that the idea of normative guidance presupposes the CI.

It may seem that the alleged impossibility of normative scepticism is a mirage. After all, it is not as if a being who did not accept norms could not act; such a being could still have beliefs and desires to determine its behaviour. Some philosophers would deny this, saying that for a judgement to count as a belief it would have to be normatively constrained, so that a normative sceptic could not have beliefs (and a similar story might apply to desires as well).[141] But we need not take that line here, because we can allow that it is possible to have beliefs and desires without norms, whilst still maintaining that this is not possible **for us**, since the relevant question is what it would be like from the inside to

[141] The idea might be that it is constitutive of belief that one ought to believe that *p* if the evidence supports *p*. For somewhat different reasons Kant also thinks that normative judgement is necessary to having genuine beliefs, see e.g. 4:452.

be like that. One would perhaps be aware of one's beliefs and desires, and aware that one's body was responsive to them. But it would not be possible to see oneself as an agent, because the idea of agency presupposes deliberation, which in turns involves normative judgement, and thus the movement of one's body could not be understood as action. So the argument that normative scepticism is impossible for us is that we are essentially agents, so that non-agency is impossible for us, and that normative scepticism is impossible for an agent.

I will now attempt to show where Kant propounds the view about freedom which I attribute to him above. One crucial passage is as follows:

> I say now: every being that cannot act otherwise than *under the idea of freedom* is just because of that really free in a practical respect, that is, all laws that are inseparably bound up with freedom hold for him just as if his will had been validly pronounced free also in itself and in theoretical philosophy. Now I assert that to every rational being having a will we must necessarily lend the idea of freedom also, under which alone he acts. For in such a being we think of a reason that is practical, that is, has causality with respect to its objects. Now, one cannot possibly think of a reason that would consciously receive direction from any other quarter with respect to its judgements, since the subject would then attribute the determination of his judgement not to his reason but to an impulse. Reason must regard itself as the author of its principles independently of alien influences; consequently, as practical reason or as the will of a rational being it must be regarded of itself as free, that is, the will of such a being cannot be a will of his own except under the idea of freedom, and such a will must in a practical respect thus be attributed to every rational being. [Kant G4:448]

Here is my paraphrase of that passage: *Every being which takes itself to be acting freely is thereby bound by the laws which are presupposed by free action (i.e. the CI). Every agent*

must take itself to be acting freely, because agents take their judgements and actions to proceed from deliberation, which is to say that they take themselves to be guided by norms, and so not determined by anything else that bypasses deliberation. And to take one's deliberation to be short-circuited by external forces is what we mean by unfreedom, so to take oneself to act freely is just to take oneself to be guiding one's actions by deliberation. What is established here is the link between agency (being rational) and freedom in the guise of apparently action-guiding normative deliberation.

3.6 THE UNITY OF REASON

Given the above understanding of what Kant says about rational beings, Kant's persistent talk of reason should sound less suspicious. There is a perception amongst those with more Humean inclinations that Kant's project is to show that morality can be derived from reason, and that he sets up acting from reason in opposition to acting from desire.[142] With regard to the second charge, I concur with O'Neill's (1989) compatibilist interpretation of what Kant says about the two standpoints.[143] For Kant, acting on the basis of desires and beliefs is the same as acting on the basis of reasons, but looked at

[142] This line is pushed persistently by Blackburn (1998: 214-224, 243-256). According to Blackburn, 'the fundamental mistake about deliberation' which Kantians make is to think of it as a way of standing back from one's desires and selecting between them. But it is a mystery to me where Blackburn finds this mistake in Kant: surely Kant's view is that when we deliberate we do not think about our desires.

[143] It is not, alas, standard to say straight out that Kant is a compatibilist. Indeed Schneewind (1998) says that Kant is an incompatibilist without explanation. Part of the problem may be terminological; compatibilism is strongly associated for many with the views of Hobbes, Hume and Ayer, and it is true that Kant's version of compatibilism is different from and more subtle than theirs. This should not blind us, however, to the fact that Kant says quite clearly that although freedom and determinism ('natural necessity') seem to be contradictory they are not: the apparent contradiction is an 'illusion' (4:456). Someone who thinks that freedom and determinism are not contradictory must think that they are compatible. The modern version of compatibilism most similar to Kant's is perhaps Strawson's (1962). Strawson reverses the traditional linkage between freedom and morality: whereas it was common to worry that if it were shown that freedom was an illusion then we would have to abandon the moral appraisal of others, Strawson maintains that the reactive attitudes (attitudes of moral appraisal) are inescapable, and thus we must see ourselves and others as free. This is, I think, the same move that Kant makes when talks about the two standpoints: we have to adopt the noumenal standpoint in order to think practically, and that is all it takes for us to be free.

from two different standpoints, external and internal respectively (or, as Kant puts it, empirical/phenomenal and practical/noumenal). Crucially, 'the concept of a world of understanding is thus only a *standpoint* that reason sees itself constrained to take outside appearances *in order to think of itself as practical*' (G4:458).[144] So there is no metaphysical problem of explaining how reason itself can motivate or anything of the kind; when we are investigating how agents are motivated we are interested only in explaining their actions in terms of beliefs and desires. Kant (1996/1783) makes this point decisively:

> In fact, the practical concept of freedom has nothing to do with the speculative concept, which is abandoned entirely to the metaphysicians. For I can be quite indifferent as to the origin of my state in which I am now to act; I ask only what I now have to do, and then freedom is a necessary practical presupposition and an idea under which alone I can regard commands of reason as valid. [Kant 8:14]

[144] Here there seems to be a point of commonality between Blackburn and Kant, since they both have a use for the idea of an internal/external distinction, as discussed in §1.6. This is not altogether surprising, since Blackburn gets the distinction from Carnap, who is pursuing a Kantian project of sorts. (The difference between Carnap's project and Kant's is, roughly, that Carnap and the other positivists wanted to replace Kant's idea of pure intuitions with that of language. Thus whilst for Kant the world of experience is structured by our pure intuitions of space and time, for Wittgenstein (1922) '*the limits of my language* mean the limits of my world' (*Tractatus* 5.6).) But there is a puzzle about the way which Kant uses the distinction, which has made it tricky to reconcile his positions in practical and theoretical philosophy: when we move between them the internal and the external seem to switch places. In the Critique of Pure Reason, the noumenal is the way the world is in itself, stepping outside our own ways of perceiving it, whereas the phenomenal is conditioned by our pure intuitions. In the *Groundwork*, the practical is the standpoint we take from within our deliberation, and the empirical is the view we have of actions when we step outside and consider how we fit into a causal order. So it looks as though the practical should be as inner as you can get. But Kant wants to identify the noumenal with the practical, and the phenomenal with the empirical. That is why he says that the practical standpoint is 'outside appearances', when it would seem more natural for him to talk of it being internal to deliberation. For Blackburn things are straightforward: the normative, practical perspective is the internal one. So what to make of Kant's view? Is it really plausible that when we deliberate we are somehow thinking about how things are in themselves? And does not Kant's humility about the noumenal mean that normative knowledge would be impossible? I do not have a firm view on this vexed subject; but here is a suggestion. What if the practical standpoint does indeed involve a projection, when seen from the empirical standpoint (as Blackburn requires), but when the practical standpoint is considered from within itself, it does seem as though we are going beyond the merely empirical, so that each standpoint takes itself to be the external one? Then there is no satisfactory answer to the question of which standpoint is **really** external, and to call the position quasi-realism is just to adopt the empirical standpoint of contemporary naturalism.

The first charge, however, seems to rest on an oversight on the part of Hume and Humeans. Hume's view is that theoretical questions are subject to reason in a way that practical questions are not. Insofar as he is an expressivist, his expressivism is confined more associated with practicality than normativity. Why else would he so obviously fail to be an expressivist concerning logic and theoretical reason in general, as he does by contrasting ethics, imbued with passion, with (theoretical) reason, which is not?[145] But as I argued in Chapter 1, it is normativity, not just practicality, which entails expressivism, so a consistent expressivism must extend to all normative judgements.[146] Hume's mistake then is either to link expressivism with the practical rather than the normative, or else to fail to notice that there are theoretical normative judgements (such as judgements concerning what is reasonable). If Hume had seen that theoretical reason is normative, and that if expressivism is true of any normative judgements then it must be the true of all normative judgements, then he would have avoided overplaying opposition between reason and passion.[147] Once a more general expressivism is accepted, one reason for suspicion towards Kant's project is removed: connecting morality with reason will not involve an attempt to derive an ought from an is, or a confusion between belief and desire. But perhaps there is a slightly different suspicion about Kant's project, that does not depend on an illicit limitation on the scope of expressivism: that Kant is attempting to derive

[145] Plausibly Hume was also an expressivist about causation and induction, but these are less obvious cases for expressivism than theoretical reason; his motivation seems to have been metaphysical scepticism rather than the thought that judgements in those areas are normative.

[146] And thus the whole framing of the debate in terms of a contrast between reason and passion is a mistake too; Kant's view, that reason and passion are two sides of the same coin, is much more in line with expressivism. I think Gibbard's (1990) expressivism about rationality supports this view.

[147] A note of caution here: part of the dispute between Hume and Kant seems to concern what mental faculties there are. We could plausibly interpret Kant as saying not just that theoretical and practical reason have a single principle (see below), but also that they involve a single faculty; whereas one implication of Hume's reason/passion dichotomy is that they involve separate faculties. It is not clear to me that the question of what faculties there are is a good question, and thus I abstain from taking sides. But if Hume meant to deduce separate faculties from a difference in the nature of the relevant judgements, then he was mistaken.

practical reason from theoretical reason.[148] The response to this kind of thought is simply that this is not what at all Kant is doing. His view is that theoretical and practical reason are not really distinct, and that neither is more fundamental; this doctrine is known as 'The Unity of Reason'.

We know that Kant believes in the Unity of Reason, because he says so explicitly – 'I require that the critique of pure practical reason, if it be carried through completely, be able at the same time to present the unity of practical with speculative reason in a common principle, since there can, in the end, be only one and the same reason, which must be distinguished merely in its application' (4:391) – and because when he talks about the principle of theoretical reason, he gives FUL:

> To make use of one's own reason [publicly] means no more than to ask oneself, whenever one is supposed to assume something, whether one could find it feasible to make the ground of or the rule on which one assumes it into a universal principle for the use of reason. This test is one that everyone can apply to himself. [Kant 2001/1786, 'What does it mean to orient oneself in thinking?', 18]

It is no surprise that Kant believes in the Unity of Reason, because the normative regress argument and his deduction of FUL entail it. Firstly recall that the normative regress argument makes no specific reference to practical norms, so if it succeeds it shows that all norms must stem from a categorical imperative. But why must it be a single categorical imperative for both practical and theoretical reason? Recall that Kant argued that the content of a categorical imperative must be deduced from its form; and since there is only

[148] This seems to be what Blackburn (1998: 223) has in mind when he criticises Kant for trying to give us something extra to say against the sensible knave – namely the accusation of irrationality. What we want to know, however, is whether the sensible knave is a normative sceptic, or just a sceptic about practical reason. If the former, then Kant would say that the sensible knave is not possible, because agents cannot be like that. But if the latter, as seems more likely, the question is how the knave can think of his norms as justified. We can think of rationality as the aspiration we have as agents towards having justified norms; but, as the normative regress argument shows, this presupposes the CI.

one form, there can be only one content, and thus the single CI, guiding both practical and theoretical reason. This does not mean that the CI has to govern all norms. Kant's (2002/1790) view of aesthetic norms is that they are not all they seem to be: because no justification can be given for aesthetic norms which links them back to the CI, no aesthetic norms are really justified, and thus aesthetic value is not objective, even though it purports to be. Note that this aesthetic error theory is not impossible in the way that general normative scepticism is.

The Unity of Reason is big news if it is correct. The idea that morality and reason are on all fours runs against the grain of a great deal of philosophy. Many moral philosophers seem to think that they can take rationality for granted when doing meta-ethics and ethical theory. For example, Scanlon (1998: 4, 153, 195-7) claims that morally acceptable principles are those which no-one can reasonably reject. But it is hardly obvious that we will be able to settle which rejections are reasonable independently of the very moral questions at issue. Notoriously, Scanlon thinks that it is reasonable to reject a principle which demands that we always save the greater number: someone who will die if such a principle is followed has a strong reason for rejecting that principle, and this rejection is made reasonable (according to Scanlon) by the fact that there are alternative principles which impose costs on others (taken individually) which are not much greater. But this depends on the idea that it is reasonable to ignore the question of how many people alternative principles impose costs on, which is surely just as contentious as the idea that it is **moral** to ignore the numbers, the conclusion that Scanlon aims to establish. This is not to say that Scanlon's criterion is incorrect, just that fails to get us anywhere.[149] The point is that Scanlon's methodology only appears plausible because we commonly

[149] See McGinn 1999.

assume that being reasonable is more basic than being moral, and if Kant's argument for the Unity of Reason goes through then that assumption is wrong. Even more startling, I suspect, are the implications which the Unity of Reason has for the study of theoretical reason. Consider, for instance, the quarrel between classicists and intuitionists over the law of excluded middle; if Kant is right we should at least contemplate resolving it by applying the CI (perhaps the CI will not cut any ice, but that is not obvious). I doubt that many logicians would be sympathetic to that kind of Kantian approach to logic, yet that is what the Unity of Reason suggests.

The Unity of Reason may not, at first sight, appear very plausible. Is the suggestion really that we can somehow derive all our epistemic norms, and even the norms of logic from the CI? Before tackling this difficult question full on, it is worth explaining what the CI does **not** have to do concerning logic, by looking at a certain kind of apparent exception to the normative regress argument. Lewis Carroll (1995/1895) famously challenges us to say why we are forced to follow *modus ponens*. Suppose I accept *p*, and *If p then q*, do I have to accept *q* as well? Can I not ask why those two premises entail the conclusion? The answer to this question must not be to supply an extra premise to the argument: *If p, and if p then q, then q*. For if the first question was in order, we can ask a similar question again, which will require a yet further premise, and we are embarked on a regress, and one seems a lot like the normative regress. There are thus two worries: first, a bad company objection to the normative regress argument, and second the thought that the CI must somehow provide the answer to the sceptical worry. The answer to the puzzle is surely to say that in accepting *If p then q* we already commit ourselves to accepting *q* when we accept *p*; indeed the conditional must be interpreted as expressing precisely this

conditional commitment.[150] Then once we accept the premises, there is no further question as to whether to accept the conclusion. The point here is that the normative question of why we should accept *q* is answered in part by a norm that we already accept: the conditional expresses a normative judgement. Now a normative regress worry can be raised concerning the justification of the conditional, but this is not a concern about the validity of the argument, but rather about its soundness; it is thus quite different from the kind of baseless doubt which Carroll warns us against. Since all deductive arguments can be put in *modus ponens* form, we need not see the normative regress as threatening logical validity, or call on the CI to justify it.

This is by no means all there is to say about the status of logic on the Kantian view; but we can get more of a handle on the more straightforward case of epistemic norms. I do not mean to attempt any derivation of those norms here; that would be beyond the scope of the current work. But perhaps by resolving some puzzles which might be thought to beset any such derivation, we can dispel some of our doubts about whether it is possible. One worry is that by being expressivists about such norms, we undermine the fact/value distinction upon which expressivism depends. For if our beliefs depend on normative judgements for their justification, are they not thereby themselves normatively loaded?[151] I am out of sympathy with this question: why ever should we think that the nature of a mental state is determined by what it depends on, rather than its subject matter? However much beliefs depend on norms, they are not about norms. So there is no immediate threat that Kantian quasi-realism collapses on itself.

[150] Note the connection to Blackburn's commitment semantics solution to the Frege-Geach problem discussed in Chapter 2.
[151] This point is pressed by Putnam (2002).

A more serious worry, however, is that we have no place to start when working from the CI. It seems hopeless to try to derive anything from the CI without relying on auxiliary beliefs as well; but if we need epistemic norms to justify beliefs in the first place, then how can the derivation ever get going? The reply we would like to make is that we start off with beliefs anyway, before we ever get into the business of making normative judgements, and that thus we can think of our practice with regard to our beliefs as more along the lines of the Neurath boat analogy that Blackburn wants to use in the case of norms.[152] But this observation may make us have second thoughts about the whole Kantian enterprise: if we have to rely on some kind of reflective equilibrium or coherentist approach anyway, even if we buy the argument for the CI, then Blackburn's suggestion that we do so from the start looks more attractive. Why should we be allowed to treat norms and beliefs so differently?

Of course, there can be attempts to provide transcendental arguments in the case of beliefs, to show that some beliefs are genuinely indispensable.[153] But even without considering whether any such arguments can succeed, there is a way of seeing an asymmetry between beliefs and normative judgements that can resolve our worries. Although expressivists hold that normative judgements are desires, this does not entail that all desires are normative judgements.[154] There is no such gap for beliefs. So when the

[152] Note that the traditional regress for epistemic justification is not the same as the normative regress, and is more easily answered. Since epistemic norms need not be the most fundamental norms, we can give reasons for coherentism or (non-transcendental) foundationalism within epistemology. But as the normative regress argument covers all norms, there is no room for this kind of justification from the outside.

[153] As was already suggested, it seems that such arguments must be limited in comparison to the kind of transcendental argument I have been considering. The argument of the *Groundwork* (as I interpret it) is meant to establish the CI for all agents, and so there is no possibility of any other fundamental norm. But the transcendentally justified beliefs will presumably be restricted to beings like us, rather than any beings capable of having beliefs.

[154] What is the difference then? I would suggest that normative judgements must be universalizable, partly because of the arguments in this chapter, but there is no need to assume this here. Blackburn (1998: 9) sees it as a matter of emotional ascent, which means that when we have pro-attitude we also have a pro-attitude

normative question arises (which happens whenever a particular normative question first arises) we start off with beliefs and desires, but no normative judgements. It is impossible to start without any beliefs, so it is pointless to try. But we do seem to be able to step outside of our normative judgements, precisely because of the availability of the transcendental argument exhibited here, and this need not involve (impossibly) doing without any desires at all. Thus, whether or not there are transcendental arguments available for fundamental beliefs, the project of starting from the CI in epistemic practice (which is an implication of the Unity of Reason) is not obviously incoherent. This is a rather tentative resting point. But remember that the normative regress argument seems to point inexorably to the CI and the Unity of Reason. So whatever doubts I feel about the derivability of epistemic norms, I am convinced that Kant's story about normativity is the best that we have.

3.7 TRANSCENDENTAL ARGUMENTS

The argument which I attribute to Kant in §III of the *Groundwork*, and which I defend here, is a transcendental argument; thus it is important to consider how it faces up to general objections which beset such arguments, and how it compares with other transcendental arguments for normative principles. What I will attempt to show is that the transcendental argument under consideration is able to meet the objections to such arguments, whereas alternatives, notably that of Korsgaard (1996), are not.

Perhaps the most influential objection against transcendental arguments is that levelled by Stroud (1968). Some statements are such that, although they are contingent,

---

to the presence of the first attitude, and so on. The problem with this is that it seems to make no sense of two-level moral theories (like Hare's – see Chapter 5) where one can think that it is better not to have the correct attitude. That is, one judges (at the critical level) that maximising utility is best, but one still thinks that the best attitudes to have (at the intuitive level) are not utilitarian attitudes.

they must be true whenever they are uttered by someone, e.g. the statement 'I am here'; Stroud calls the class of such statements the 'privileged class'.[155] Stroud suggests that the idea of transcendental arguments is to defeat scepticism by showing that the statement which the sceptic doubts is a member of the privileged class. But if this is the strategy of transcendental arguments, there seems to be a problem:

> In particular, for any candidate S, proposed as a member of the privileged class, the skeptic can always very plausibly insist that it is enough to make language possible if we *believe* that S is true, or if it looks for all the world as if it is, but that S needn't actually be true. Our having this belief would enable us to give sense to what we say, but some additional justification would still have to be given for our claim to *know* that S is true. [Stroud 1968: 255]

Whether or not Stroud is right that the sceptic can insist that all that has been shown is that we must believe S, he is clearly correct that there is a gap between showing that we have to believe S and showing that it is true. And it seems that this criticism will be applicable to the argument of this chapter, because what I have tried to show is precisely that we must accept the CI, and this does not seem in itself to dispel doubts about its truth or validity. Yet Skidmore (2002) has attempted to show that Stroud's objection is ineffective against transcendental arguments for principles of reason (i.e. norms). Skidmore's idea is that for norms, as opposed to claims about the world, there is no gap between our having to accept them and their being valid:

> The fact that we must believe external objects exist does not entail that they *do*, and this leaves room for Stroud's skeptic. However, in the case at hand regarding **NC** [a principle of non-contradiction], this distinction dissolves. To demonstrate that anyone capable of

[155] These are epistemic or primary necessities, in the terminology of two-dimensional modal semantics. See e.g. Chalmers (2005).

> belief must accept **NC** *just is* to demonstrate that **NC** is a valid principle of theoretical reasoning. [Skidmore 2002: 127]

This is a tempting line to take, but I fear that it cannot be used to defeat scepticism concerning fundamental norms. Or rather, we need a more complicated set of reflections to get us to Skidmore's position.[156] The problem is that the principle P, that if a norm *N* is such that the sceptic must accept it then *N* is valid, seems itself to be a normative principle: it is just a way of saying that the fact that the sceptic must accept a norm justifies it. It is thus in order for the sceptic to doubt P too. (This is just another instance of the normative regress that concerned us earlier.) Nor does it seem fair to stipulate that the validity of norms consists in the necessity of their acceptance; that seems no better than the Platonist move of stipulating that validity consists in correspondence with the Form of the Good. Or at least this is how things look from a fully internal perspective. Why could we not simply be mistaken even about a norm which we must accept, given that the fact that we must accept it is just that: a fact which can only have normative relevance if a further norm is presupposed? From the internal perspective normative facts are just as robust as facts about the external world, and there seems just as much room for error, even necessary error. To get the closure of the gap between appearance and reality which Skidmore wants, we have to move external: when we see normative judgements as attitudes, we must see necessity (of the kind possessed by the CI) as the limit of anything that could be thought of as validity for them. The justification possessed by the CI is thus

[156] Skidmore (2002: 127) argues that it is the limited scope of the norms we are trying to justify which 'removes the gap between appearance and reality on which Stroud's skeptic depends'. I confess that I do not understand how this is meant to help; and anyway, it will not help *me*, because the principle I wish to defend (the CI) is not limited in scope. Indeed for Kant the position with respect to limitation of scope is opposite to that suggested by Skidmore: transcendental arguments for external objects are limited in scope because they apply only to beings like us, whereas transcendental arguments for norms are unlimited in scope because they apply to all beings for whom normative questions arise. See also §3.5 above.

of a radically different kind from that of the norms it justifies (or helps to justify): it has an external justification, not an internal one.

Enoch (2006) has influentially attacked attempts (including Korsgaard's) to ground normativity in what is constitutive of agency. It is therefore important to consider here whether his objections apply to my argument, and whether they can be met. At first, the challenge is particularly threatening because it harks back to the shmoralising objection met in §3.0: Enoch asks why we should be agents rather than shmagents. In fact, however, Enoch effectively concedes a difference between the cases: being an agent is unavoidable, in contrast to moralising. Despite accepting this unavoidability, however, he thinks there is still a way for the sceptic to be an agent whilst avoiding the commitments which are constitutive of agency:

> Perhaps, "Korsgaard's skeptic may say, "I cannot opt out of the game of agency, but I can certainly play it half-heartedly, indeed under protest, without accepting the aims purportedly constitutive of it as mine." The kind of necessity the game of agency has to enjoy in order to solve the problem we are now in is *normative* necessity. Invoking other necessities here will just not do. [Enoch 2006: 188]

I confess that I do not fully understand this objection, perhaps because it depends on specific features of Korsgaard's view that are not shared with mine. What I at least identify as the aim of agency is simply to deliberate about what to do. But what could it be to accept that I cannot opt out of agency whilst refusing to accept the aim of deliberation? That would not leave anything left to being an agent. The crucial point is that to be an agent is already to be deliberating, like it or not, and in deliberating one is already involved in looking for the right answer to whatever deliberative question one faces. That is

because there is no such thing as deliberating arbitrarily, and to deliberate in a non-arbitrary way is to be committed to the existence of norms governing that deliberation.

Now Enoch (2006: 195) later suggests what he calls an alternative conclusion to draw from the unavoidability of agency: that 'the fact that our best attempts at deciding what to do and how to live our lives require normative facts (indeed, irreducibly normative ones) gives us just as good a reason to believe in normative facts.' I think I agree here, but what Enoch seems to miss is that the Kantian argument is a transcendental argument, which precisely involves the move from the unavoidability of a belief to the truth of that belief, as discussed above. So, given anti-realism, Enoch concedes everything that my argument requires. Enoch himself thinks that his alternative somehow points to normative realism – I think because he assumes that the belief in normative facts that agency necessitates is somehow committed to realism about those facts. I think I have already said enough above and in Chapter 1 to explain why the appeal to realism to ground normativity is hopeless.

Although there is much in this chapter which is in agreement with Korsgaard (1996), it is important to note the difference between the transcendental argument for morality which she offers, and the one I find in Kant and present here. Korsgaard's argument also takes normative scepticism to be impossible, in roughly the same way that I do, but rather than taking this to be a (more or less) direct grounding for the CI, she takes it to license us to value our humanity, *qua* practical identity. The problem with this is that a separate argument is now required against egoism: Korsgaard claims that because private reasons are ruled out (by a Wittgensteinian argument), all reasons must be agent-neutral. But as Skidmore (2002) argues, this seems to equivocate between different senses in which a reason might be said to be private: a reason that is comprehensible to others is in

one sense (Wittgenstein's) not private, but there is another sense in which a reason is private unless others have the same reason. What Korsgaard argues for is that reasons must be comprehensible to others, but what she needs as a conclusion is that they must be shared by others, and it is simply unclear how the two can be convincingly linked. But that is not to say that I disagree with Korsgaard's conclusion. Fundamental reasons must indeed be agent-neutral (i.e. shared by all agents), but the reason for that is the normative regress argument: to think that the facts which distinguish between people have normative relevance is to presuppose some norm concerning those facts, and ultimately such fundamental norms cannot depend on identifying facts. This is really to put in slightly different terms the argument of this chapter. Since Korsgaard comes so close to stating the normative regress argument, and connecting it to Kant's thinking, it is slightly puzzling that she does not employ it more directly to justify the rejection of fundamental agent-relative (i.e. non-agent-neutral) reasons.

I conclude by returning to Egan's argument from fundamental error. Recall that the great difficulty for Blackburn was that he seemed committed to immunity to fundamental error, but that on his account a claim to such immunity would be unpardonably smug. What the Kantian considerations above are meant to have shown is that reliance on the CI is not unpardonably smug. Indeed it may help us to get clear on exactly what the transcendental argument is meant to show if we think of it as a way of persuading ourselves that it is not unpardonably smug to take ourselves to be immune from error concerning the CI.[157] This is how it transpires that the Kantian project and the quasi-realist

[157] Note that we do not have to take ourselves to be completely immune from error, even about the CI. It may be that the argument of this chapter is flawed; but it does not seem right to see the judgement that the argument is correct as just another attitude to be traded off with other attitudes as part of Neurath's boat. The argument takes a sufficiently Archimedean stance towards the boat that we can reasonably think of it as standing outside. What it tells us is that the CI is indispensable to any boat at all; without the CI as glue, any

project are not opposed, but rather inseparable. Hare's Kantian expressivism, with its deep commitment to universalizability, stands vindicated.

boat of norms will collapse.

CHAPTER 4: FROM UNIVERSALIZABILITY TO UTILITARIANISM

4.0 INTRODUCTION

Hare was most notorious for thinking that he could give an argument from universalizability to utilitarianism. He thought that showed that the correctness of utilitarianism was a kind of conceptual truth (given our moral concepts), because he had shown that universalizability itself was analytic. The initial task of this chapter is to evaluate Hare's argument in detail, focusing on the version given in Moral Thinking and deal with the various serious objections it faces. I will then briefly consider the various Kantian and neo-Kantian attempts to ground a non-utilitarian ethics on universalizability, and show, at least in outline, why such approaches are doomed to failure. Finally I will discuss some aspects of Hare's own version of utilitarianism, with the aim of showing both how they are connected with other aspects of his theory (his expressivism and Kantianism), and what they have to offer to ethical theory.

Before we begin to analyse Hare's argument in detail, we should consider how the status of the argument changes in light of the modifications to Hare's approach that we have had to make so far. Having seen in the previous chapter how best to avoid the shmoralising objection in arguing for universalizability, we will see the resulting argument for utilitarianism in a somewhat different light. We have seen that whether or not universalizability is a conceptual truth, it is established more firmly by being based on a transcendental argument, so that it is synthetic a priori rather than analytic. This defuses the worry that we have been given no reason to moralise rather than shmoralise (where the latter concept does not entail universalizability), and in turn it will help to show that Hare's theory does not end up falling victim to the Open Question Argument. If it were really true that on Hare's view utilitarianism was a conceptual truth, that would seem to

make his view just as implausible as the analytic utilitarianism that is one of Moore's original targets. It seems that we can coherently, and whilst being competent with our moral concepts, doubt whether utilitarianism is true, and that sits far better with the synthetic *a priority* of utilitarianism than with its analyticity. So the worry that Hare's utilitarianism ends up undermining his expressivism does not get off the ground within the more Kantian framework we are now working with.

With this much understood I will go on in §4.1 to explicate and critique Hare's argument for utilitarianism at length, arguing that with some sympathetic tweaking it is still defensible. Along the way I defend Hare against various worries, both technical problems relating to the details of his argument and the charge that his methodology should be abandoned in favour of reflective equilibrium. In §4.2 I explain Hare's two-level theory of moral thinking, explain how as a version of indirect utilitarianism his theory differs from rule-utilitarianism, and defend it against Bernard Williams' objection that it is unstable. In §4.3 I consider objections from Rawls and Korsgaard to the effect, respectively, that utilitarianism does not respect the separateness of persons and, by being a teleological theory, does not recognise the relational nature of morality. I respond by insisting explaining that Hare's theory does not have these failings, and that insofar as the objections are in principle persuasive that shows the superiority of Hare's version of utilitarianism over others that are subject to those objections. I also explore the extent to which utilitarianism can capture a more substantive kind of equality by considering how it relates to Dworkin's Equality of Resources.

4.1 HARE'S ARGUMENT FOR UTILITARIANISM

Hare gave several apparently different arguments for utilitarianism, though the common thread was the idea that it followed from universalizability. The argument given in *Freedom and Reason* (1963) was criticised by Mackie (1977: ch. 4) for relying on a version of universalizability apparently stronger than what Hare was entitled to. Mackie distinguished between three 'stages' of universalizability:

(i) Ignoring numerical (i.e. non-universal) differences between individuals.

(ii) Trying to put oneself imaginatively in the place of others, so as to avoid bias in favour of oneself.

(iii) Trying to put oneself in the place of others to the extent of having their preferences, so as to avoid the bias stemming from having different preferences from other people.

Mackie took it that utilitarianism only followed from (iii), but that only (i) had any plausibility as a conceptual truth.[158] I will not consider whether the argument of *Freedom and Reason* really relied on (iii) rather than (i); at least in *Moral Thinking* (1981) Hare gave an argument relying only on (i), and since it is (i) that was justified in the previous chapter, it is that argument that we will be most concerned with.[159] Mackie's mistaken interpretation, if that is what it was, can be forgiven, because Hare certainly does make use of (iii); what becomes explicit (in Hare 1981) is that (iii) is meant to an additional ingredient of the argument, combining with universalizability rather than following from it.

[158] Fullinwider (1977) makes a similar criticism of Hare.

[159] Hare had published 'Ethical Theory and Utilitarianism' (1989a: 212-30) in 1976, giving an argument much more similar in its use of universalizability to that of *Moral Thinking* than *Freedom and Reason*, but Mackie (1977) does not refer to it, presumably because his book was already in draft by then. Hare (1988: 268) claims that he did not change his view about universalizability.

Indeed it emerges that on Hare's view (iii) derives not from the moral concepts but from the concept of the self.[160]

Of course, once we are clear that it is only the minimal kind of universalizability that we are entitled to as a premise, it is very unclear how universalizability could be thought to follow. And indeed the argument of *Moral* Thinking has universalizability (U) as one of two premises. The other, dubbed 'Conditional Reflection' by Allen Gibbard (1988), concerns the relationship between judgements concerning what one would prefer in counterfactual situations and conditional preferences concerning those situations. Hare took Conditional Reflection (CR) also to be a conceptual truth; we shall see that this verdict needs to be adjusted in a way similar to the shift in understanding the status of universalizability in the previous chapter. The two premises of the argument are then as follows:

(U) A moral judgement about a case commits one to making an identical judgement about all cases identical in their universal properties.[161]

(CR) Insofar as one fully knows what one would prefer in a hypothetical case, one must have the corresponding preference (same sign, same strength) with regard to that hypothetical case.[162] (That is, if one knows that in S one would prefer that X, then one now prefers that [if one were in S, X]. Knowledge of a conditional with one's preference as the consequent gives rise to a conditional preference. We say that the conditional preference is the conditional reflection of the preference in the hypothetical.)

And the conclusion is preference utilitarianism:

[160] See Hare 1988: 268-9.
[161] See Hare 1981: 108.
[162] See Hare 1981: 99.

(PU) Moral decision-making must assign equal weight to everyone's preferences (resolving conflict in the same way as for the intra-personal case).

I will consider four apparently serious objections to this argument:

(a) The argument does not appear to be valid: it is not explained why universalizability entails utilitarianism even given Conditional Reflection.

(b) Conditional Reflection itself seems to be false: there are apparently obvious counterexamples.

(c) It is implausible that the argument really succeeds, because the conclusion is so absurd: Hare's version of utilitarianism has absurd consequences because e.g. it cannot exclude external preferences.

(d) The argument seems to presuppose that our preferences are the raw material for morality; but this assumption is not argued for and seems to beg the question in favour of utilitarianism and views which are its close cousins.

a) To ascertain the validity of Hare's argument, I begin with his presentation:

> It follows from universalizability that if I now say that I ought to do a certain thing to a certain person, I am committed to the view that the very same thing ought to be done to me, were I in exactly his situation, including having the same personal characteristics and in particular the same motivational states. But the motivational states he actually now has may run quite counter to my own present ones. For example, he may very much want not to have done to him what I am saying I ought to do to him (which involves prescribing that I do it). But we have seen that if I fully represent to myself his situation, including his motivations, I shall myself acquire a corresponding motivation, which would be expressed in the prescription that the same thing not be done to me, were I to be forthwith in just that situation. But this prescription is inconsistent with my original 'ought'-statement, if

> that was, as we have been assuming, prescriptive. For, as we have just seen, the statement that I ought to do it to him commits me to the view that it ought to be done to me, were I in his situation. And this, since 'ought' is prescriptive, entails the prescription that the same be done to me in that situation. So, if I have this full knowledge of his situation, I am left with two inconsistent prescriptions. I can avoid this 'contradiction in the will' […] only by abandoning my original 'ought'-statement, given my present knowledge of my proposed victim's situation. [Hare 1981: 108-9]

Let us call the characters here 'Agent' and 'Victim'. The argument appears to run as follows: Agent starts by judging that he ought to do X to victim; by (U) Agent is thereby committed to judging that X ought to be done to him in the situation where he is in Victim's position (and everything else is the same); but Agent knows that Victim actually very strongly prefers that X not be done to him; and so Agent knows that if he were in Victim's position he would likewise strongly prefer that X not be done to **him**; and now by (CR) Agent has a strong conditional preference concerning the situation where he is in Victim's shoes that X not be done to him. All this means that Agent has two conflicting preferences regarding the hypothetical situation where he is in Victim's shoes: the preference, arrived at by (U) from his original moral judgement, that X should be done to him, and the stronger preference, arrived at by (CR), that X should **not** be done to him. Now Hare thinks that this **intra-personal** conflict is resolved in the normal way, with the stronger preference winning out; but if Agent ends up preferring that X not be done to him in the hypothetical situation, he cannot genuinely think that X ought to be done to Victim in the original situation, because an ought-judgement is a universal prescription. The result is that Agent must abandon his original moral judgement (and indeed reach the opposite

moral judgement); and this is a utilitarian result because the procedure aligns moral judgements with judgements of the maximisation of preference satisfaction.

Now Hare's argument has come in for various criticisms. Schueler (1984) complains that Hare is trying to compare the strength of a categorical preference with a hypothetical preference, when in fact the two will not conflict at all if the antecedent of the condition in the hypothetical preference is false. Unfortunately, Schueler has misunderstood the argument: the two preferences whose strengths are compared are **both** hypothetical preferences concerning the counterfactual situation where Agent is in Victim's shoes. Hare uses (U) to generate a hypothetical preference from the categorical preference before the comparison of strengths takes place.[163] Although Schueler's criticism is misplaced, there is a problem lurking here for Hare, which comes about when we have a case involving more than two people, as Persson (1989) points out. What Hare's procedure effectively allows for is the pairwise comparison of preference strengths concerning an action of two individuals. What we would expect from a utilitarian view is aggregation: e.g. when Agent prefers that he do X, B equally strongly prefers that Agent do X, and C has a slightly stronger preference that Agent not do X, then the preferences of Agent and B combine to outweigh C's. But Hare's procedure does not allow for this. On the contrary, when Agent considers the hypothetical situation where he is in C's shoes the result will be that his preference is outweighed and he cannot say that he ought to do X (B's preference doesn't get considered when this hypothetical situation is at stake).[164]

[163] See Hare 1981: 112-6 and 1989a 245-50.

[164] Indeed Hare's account treats this kind of case in a very puzzling way. For when Agent considers the hypothetical situation where he is in B's shoes it is C's preference that is ignored. The upshot is that Agent can neither think that he ought to do X (because of the case where he is in C's shoes) nor that he ought not to do X (because of the case where he is in B's shoes), so it seems to follow that doing X is morally optional for Agent. Such a result would also occur in a three person case where there were three options, each favoured by one person, even if one person's preference were much stronger than the other two's. So Hare's procedure is not just deficient because it does not result in utilitarianism, but also because what it does

Persson (1989) points out that if Hare were allowed to import some kind of veil of ignorance it would be possible to compare multiple preferences at once; but he also acknowledged that this is not Hare's account, nor is it in the spirit of Hare's overall view given that it involves an element of make-believe. In his reply to Persson, Hare (1989c) agrees about all this, and offers a tentative and unimpressive solution, which I shall not discuss.[165] There are three solutions which are worth considering, however. The first is due to Rabinowicz & Strömberg (1996), who start with the problem that Hare has shown using (CR) that an agent has conditional preferences regarding each of the hypothetical situations where she is in each person's shoes, but has not shown how to compare these preferences to reach an overall verdict, because they are preferences about different situations. What Rabinowicz & Strömberg suggest is that the agent should engage in the minimum necessary revision of her preferences concerning each situation to bring it about that she has uniform preferences about them (and so a universalizable moral judgement). They prove that the minimum change to achieve uniformity is produced by averaging the preferences, thus resulting in (PU) as required, and their proof extends to cases where there are multiple options.

There are several worries about the Rabinowicz & Strömberg solution, discussed by Rabinowicz (2009), of which the two most important seem to me to be as follows. First, the solution does not involve the mechanism of intra-personal conflict and resolution between preferences, and it uses universalization in a different way from Hare's (not to generate preferences about hypothetical situations but to enforce uniformity) so it is quite far from in the direction of reconstruction rather than charitable exegesis. Second, the proof of averaging relies on a particular view about what minimises preference change. If

result in is absurd.

[165] See Rabinowicz & Strömberg 1996 for discussion of Hare's (1989c) solution.

we take the agent's preferences about the various situations as a vector, and thus a point in Euclidean vector space[166], we can think of minimal change as the minimum distance between that point and the line x = y = z which represents uniformity of preference. That way of thinking about minimising change results in averaging, and it is perhaps the most intuitive. But another way is what Rabinowicz calls the 'city-block' measure, which minimises the sum of absolute distances travelled along each dimension to reach uniformity. This often produces different results to averaging. As Rabinowicz comments, the main advantage that the Euclidean distance measure has over the city-block measure is that the former is 'fairer' (gives more weight) to stronger preferences (strictly those which are most different from the preferences of others).[167] But that can hardly be a relevant consideration to the utilitarian.[168] One answer which occurs to me is that when in the **intra-personal** case the difference between city-block and averaging makes a difference to what to do, we use the averaging method (i.e. when resolving conflicts amongst our own preferences we allow particularly strong preferences to have proportionately more weight in determining what we do). So one possible rationale for the Euclidean measure is that it is an extension of intra-personal reasoning about preferences to the inter-personal case.

The problems with the Rabinowicz & Strömberg proposal have led Rabinowicz (2009) to prefer my solution.[169] What I suggest is that we modify Hare's use of (U). Hare

[166] I.e. Supposing that the is only one potential action in question, we assign each preference an integer value, with positive and negative signs indicating being in favour of or against the action respectively, and the modulus representing the strength of the preference. Thus the Cartesian coordinates (4, 4, -5) would represent equally strong preferences in favour of X for the first two situations, and a slightly stronger preference against X for the third situation.

[167] I am not entirely sure that Rabinowicz is right to call this a kind of fairness.

[168] Rabinowicz credits Christian List for pressing this point.

[169] I initially thought that my solution was simply what Hare himself had in mind, but Rabinowicz has convinced me that although it is perhaps the most faithful to Hare's intentions it is still to some extent a reconstruction.

generates an intra-personal conflict between my conditional preference concerning the situation where I am in another person's shoes, and my original preference, mirrored in that situation by universalization (but this fails to deal with multiple person cases). Instead I suggest that the intra-personal conflict concerns the actual situation, with a conflict between my actual preference, and other preferences representing each of the other people involved. These other preferences are generated first by an application of (CR), producing conditional preferences concerning the other situations as before, and then by an application of (U) mirroring these conditional preferences back to the actual situation. So effectively I make (U) work in the opposite direction from the way Hare suggests, but with the result that we get **all** the relevant preferences applying to the same situation in intra-personal conflict (rather than just a pair of preferences), meaning that we get (PU) even for multiple person cases.

To explain my solution more formally, suppose that there are agents $A_1 \ldots A_n$ who have various preferences $P_1 \ldots P_n$ respectively concerning a certain action X which A_1 is in a position to perform (for what follows you are A_1, morally deliberating). We consider different possible situations where you (A_1) occupy the position of each agent: in S_k you are in A_k's position (so S_1 is the actual situation). We have a rich sense of being in someone's position, such that this involves having their preferences (rather than yours). So S_k is a situation where your preferences are the same as A_k's, including relevantly P_k; if you were in S_k your preference regarding X would be P_k. Now we are in a position to apply (CR): since you know that if you were in S_k your preference would be P_k, you must now have a matching conditional preference for S_k. Knowing someone else's preference involves having a matching conditional preference for the situation where one is in that person's position. That is, your preferences for $\langle S_1 \ldots S_n \rangle$ are $\langle P_1 \ldots P_n \rangle$ by conditional reflection. Now

the move (where I depart from Hare) is to apply universalization to each of $P_1…P_n$ individually. That is, because I have the conditional preference P_k concerning S_k, I must have the same preference concerning S_1, since S_k and S_1 are identical in their universal properties. What I must end up with then is the full set of preferences $P_1…P_n$, but now all concerning S_1. Now the preferences are about the same situation, so they do conflict, so I do balance them in the normal intra-personal way, and (PU) results.

But why should the preferences be universalized individually? Hare thinks that moral judgements are universalizable, but these individual preferences are not themselves moral judgements. This is effectively Rabinowicz's (2009) worry about my solution. He acknowledges that we can see the preferences as *pro tanto* moral judgements (though Hare never mentions these), but he wonders why we are allowed to treat preferences as *pro tanto* moral judgements which are thus universalizable. My thought is that a better way to think about *pro tanto* moral judgements is as judgements about moral reasons. To treat P_1 as a moral reason, which is what I do by counting it as an input into the overall moral judgement, is to universalize it: to take it to be a reason in any S_i (identical in universal properties to S_1), not just S_1 itself. And the same must apply to every P_i. So once I have P_k regarding S_k (by conditional reflection) I must, in treating it as a moral reason, have matching P_k regarding S_1. It should not be controversial that (U) applies to judgements about moral reasons. Either we can simply say that such judgements **are** moral judgements, and so fall within the scope of (U) straightforwardly, or if this is denied, note that if I judge that there is a moral reason to perform X, then I am committed to judging that it is overall right to X, other thing being equal, and since moral judgements must be universalizable, so must judgements that other things being equal entail moral judgements. Assuming that this explanation is satisfactory, it seems that my solution

avoids the problems facing Rabinowicz & Strömberg, in that it is in the words of Rabinowicz (2009) 'closer to Hare and somewhat less question-begging'.

There is another possible solution on offer, however, offered by Vendler (1988), who emphasises that when Hare imagines the agent putting herself into the shoes of others, for the purposes of applying (CR), this process is very complete. It involves simply shifting to the other person's view of the worlds, with no remainder of the agent's own personality. This is as it must be in order for (U) to apply too: it must be that there is no difference in universal properties between the actual situation and the one where the agent is in another person's shoes. One thing we might ask here is what exactly non-universal properties would be if there were any. A universal principle is one that contains no individual constants; so we can assume that a non-universal property would be something like a haecceity, a trivial individual essence.[170] Now the view that there are haecceities is certainly not metaphysical orthodoxy, so it may very well be that in fact **all** properties are universal. That would mean that in fact when we imagine putting ourselves in another's shoes (in Hare's sense) we are not imagining a different situation from the actual one (though alike in universal properties), but in fact the very same situation, only imagining a different point of view on it. This is what Vendler insists upon, and Hare (1988: 284-5) indicates that he did not mean to imply otherwise. The relevance for the argument for utilitarianism is explained by Vendler (1988: 182-3) as making things easier: it seems that we can simply apply (CR) to generate preferences (matching those of the different people involved) for being in each person's shoes, but these preferences will not be about different hypothetical situations; rather, all the preferences will be about the actual

[170] For further explanation of the idea of haecceitism, see Lewis 1986: ch. 4, section 4.

situation, and so will be straightforwardly involved in intra-personal comparison, resulting in (PU).

Now there is something suspicious about this reading of the argument, particularly in that it apparently makes no use of (U). Indeed Vendler (1988: 183) says that he can make no sense of Hare's use of universalizability. But then it seems as though (at least on Hare's view), there is nothing specifically to do with morality going on in the argument. Hare does not think of (CR) as a requirement of morality, but of understanding e.g. suffering, so if the argument depends only on (CR) and not (U) then it turns out that (PU) is required simply by being able to understand what it is to have preference, which is an absurd conclusion. Rabinowicz (2009) complains that if Vendler is right then in fact (CR) is inapplicable, because on Vendler's picture to imagine the point of view of someone else is not really to imagine oneself in their shoes, and so does not generate a conditional preference. Indeed Vendler (1988: 183) is quite explicit that he is assuming that what morality brings in is indifference between the points of view of different individuals involved. What this means is that the crucial premise of Vendler's argument is neither (U) nor (CR), but a substantive principle of impartiality. Even then the argument does not really work, because it is not clear that impartiality requires conflicts between the preferences of different people are to be resolved in the same way as inter-personal conflicts, and without that we do not get (PU) or any form of utilitarianism. So Vendler's argument is, despite superficial resemblance, not at all a version of Hare's, and moreover it is less successful.

Now Rabinowicz (2009) seems to be concerned that if Vendler is right about the metaphysics then Hare's argument reduces to Vendler's, and so fails. In my view there is no inconsistency between my reconstruction of Hare's argument (or that of Rabinowicz &

Strömberg) and Vendler's metaphysics. I think that to understand the issues here we need to distinguish between (possible) worlds and centred worlds. A centred world is the ordered pair of a world and spatiotemporal location, which can do service for a point of view. Vendler is effectively saying that when I imagine being in another person's shoes (in Hare's sense) I am **not** imagining a different possible world; rather I imagine a different centred world, where both that centred world and my own (the actual world from my point of view) are centrings of the same possible world. Prudence gets a grip not on possible worlds but on centred worlds; and morality makes demands on us precisely because we have particular points of view represented by centred worlds.[171] So it is important that when Hare's argument is concerned with cases/situations, it is concerned (if Vendler's metaphysics is correct) with centred worlds. So long as we remember that (U) and (CR) concern centred worlds, there is no problem for the argument (*pace* Rabinowicz 2009).

b) So far we have been concerned only with the validity of Hare's argument. But of course it is also necessary to consider whether (CR) is an acceptable premise. Many commentators on Hare have expressed some doubt about whether it is really true that, in judging that in a hypothetical case I would prefer that X happen, I now prefer that X happen if I were in that case.[172] Hare (1981: ch. 5) builds up to (CR) by thinking about suffering. His initial thesis is that there is a very tight connection between (i) suffering; (ii) believing that one is suffering; and (iii) wanting that suffering to end. He moves from these thoughts about one's actual suffering to thoughts about hypothetical suffering: if I believe that in situation *S* I would be suffering, then I (now) desire that if I were in *S* that suffering

[171] See here what Hare (1988: 286-7) has to say in response to Vendler.
[172] For example Schueler (1984), Brandt (1988: 34), Gibbard (1988: 61), Nagel (1988: 104) and Hajdin (1990).

be ended. Next we generalise from suffering to preference in general: if I know that in *S* I would prefer that x, then I now prefer that [if I were in *S*, x]. That is, from a conditional of the form p > pref (x) one can infer a conditional preference of the form pref (p > x). And this is just (CR). What Hare says is that one cannot **know** the conditional about one's preferences without having the conditional preference (unclear whether he could say 'really believe' rather than 'know'). So any entailment between the conditional and the conditional preference is purely pragmatic and first-personal (like Moorean entailment). The point about suffering is that it is somewhat plausible that genuine understanding of what it is like to suffer involves having the relevant preferences; and this is meant to transfer over to understanding what it is like to have a preference.

The problem with (CR) is that we can doubt that what is true of suffering is true of preference in general, and there are various apparent counterexamples to the connection between knowledge of conditionals about one's preferences and one's conditional preferences, which make (CR) seem implausible. Potential counterexamples (in each case, it seems that I can know the conditional without having the relevant conditional preference):

(1) If I were (willingly) addicted to heroin, I would want to take heroin.

Or in general:

(1*) If I were irrational, I would prefer that X.

(2) If I (falsely) believed that eating crispy was healthy, I would want to eat a lot of crisps.

Or in general:

(2*) If I falsely believed that P, I would prefer that X.

What these cases suggest is that the analogy between suffering and preference breaks down in the case where the preference is clearly mistaken in some way. (Of course, suffering could be a result of a mistake, such as a painful false belief, but then the way to end the suffering is to end the false belief.)

I suggest a possible restriction to the inference to conditional preference: when I take my hypothetical preferences to be irrational, and so not to reflect my best interests, I do not defer to them.[173] I usually defer to my hypothetical preference (by conditionally reflecting them) because to think of the hypothetical individual as 'I' **is** to take that individual's interests as my own. Normally to take an individual's **interests** as one's own is to take their **preferences** as one's own too; but this breaks down precisely in the cases where the interests and preferences fail to align. There is a defeasible presumption that preferences and interests are aligned; there must be a positive reason to discount the hypothetical preferences; it is not enough merely that they differ from one's actual preferences. What happens in those cases where we do not defer to the hypothetical preferences? Is it that no conditional preferences are generated in such cases? No; what happens is that we form conditional preferences which are adjusted in light of our superior epistemic position. For example:

(3) If I liked strawberry ice-cream better than chocolate ice-cream, but mixed up tubs of the two, I would prefer to get the tub of chocolate ice-cream.

Note that in this case the preference that I would have matches my actual liking for chocolate ice-cream. But that is not the preference I conditionally reflect; rather, I correct

[173] Hare (1989a: 41-2) seems to take exactly this view in his 1979 article 'What Makes Choices Rational?', restricting Conditional Reflection to rational preferences, and even in his 1976 article 'Ethical Theory and Utilitarianism', Hare (1989a: 218) says that his theory takes into account not people's actual desires but 'what they would desire if they were fully informed and unconfused'. But for some reason he does not carry this over to *Moral Thinking*, where he gives Conditional Reflection in its unrestricted form. He also does not seem to notice that the restricted version arguably leads to a more attractive version of utilitarianism.

for the error that I hypothetically make, and conditionally prefer to have the tub of strawberry ice-cream.

Would the argument for utilitarianism be harmed if (CR) were restricted in this way? No, because normally there is no positive reason to think that the preferences and interests are misaligned. And in cases where we do not defer we will still have conditional preferences; it is just that rather than directly reflecting the preferences of others, they will reflect the interests of others (understood as their preferences adjusted in accordance with what we know). This is still a version of utilitarianism, but one that takes something like an informed preference model of well-being rather than a simple preference model.[174] But the news may be even better. A crucial objection to (at least) Hare's version of utilitarianism is that it counts external preferences (those which are not self-regarding). The revised version of conditional reflection may motivate disregarding external preferences, on the grounds that they do not conditionally reflect. If the truth in conditional reflection lies in the idea that "I"-thinking involves concern for interests, then external preferences will only conditionally reflect if they are relevant to interests, and they are not. For example:

(4) If I cared very much about Venice, I would prefer that Venice survive.[175]

But it is not plausible that in order to know what it is like to care about Venice, I have to conditionally prefer that if I did so Venice survive for as long as possible. (Suppose, to get intuitions clear, that I don't find out what happens to Venice.) Whatever 'I'-thinking involves, it is not that.

[174] This seems like an advantage because I believe that the best model of well-being is the informed advisor account, where one's informed desires at *t* are the desires a fully-informed and rational version of oneself would advise one to have at *t*. I do not think it wise to engage here in a full discussion of the different options in the theory of well-being, because that would be tangential to the issues under discussion.

[175] This example is similar to the 'Big Funeral' example discussed by Gibbard (1988). And indeed Gibbard comes to a similar view to mine, although not I think through quite the same reasoning.

If all this is right, then a plausible version of (CR) can be salvaged, allowing for a valid Harean argument for utilitarianism, but the form of preference utilitarianism which emerges involves informed preferences, and disregards external preferences:

(PU*) Moral decision-making must assign equal weight to everyone's interests, understood as their non-external preferences adjusted for limitations of knowledge and rationality (resolving conflict in the same way as for the intra-personal case).

(c) Now it should be clear even before the discussion of external preferences begins what my proposed solution is, and how that solution can be found in Hare's argument for utilitarianism. The term 'external preferences' was invented by Dworkin (1977: 234) to describe preferences which are not personal, and which therefore lead to counterintuitive consequences for preferences utilitarianism. For example, people may want their friends to flourish (and their enemies to suffer). The problem which Dworkin points to is that this seems to lead preference utilitarianism towards a failure of impartiality, because the interests of those who have many friendly supporters will end up counting more than those who do not. It seems as though by treating preferences impartially utilitarianism ends up failing to treat people's interests as mattering equally (Harsanyi 1988: 98). Hare (1981: 104-6, 182) did not initially seem to think that this was a serious worry for his view. His direct response to Harsanyi is to deny that excluding external preferences avoids all the relevant counterexamples to utilitarianism, meaning that it does not really help to exclude them (Hare 1988: 246), and to explain that his theory does not give him a good rationale

for excluding them (Hare 1988: 246-7).[176] But Hare does not respond to the charge that it is objectionable not to treat people's interests equally.

Gibbard (1988) explains however (similarly to my argument above), that external preferences are counterexamples to (CR). Insofar as Hare's utilitarianism is a product of (CR), it thus automatically excludes external preferences, meaning that, despite Hare's reply to Harsanyi, there is a rationale for excluding them. Hare (1988: 231-2) seems happy to go along with this[177], and so we can say that the problem of external preferences is solved. Hare notes, however, that this does not allow us to deal with the fanatic so straightforwardly. The fanatic is someone who has universal preferences in the service of some ideal, e.g. someone who prefers that everyone listen to Wagner whether they like it or not.[178] It is not that ruling out external preferences makes no difference to how we treat the fanatic, but that it does not help us to show why the fanatic is wrong:

> If somebody has such a great preference that homosexuals be put in prison, that he is prepared to prescribe that he himself be put in prison if he were a homosexual, even if he fully represents to himself the struggles of the homosexual in prison, we cannot argue with him. Admittedly, he cannot argue with us either, if we are not bound to form preferences similar to this for the case where we are in his position. This will be so if the Conditional Reflection Principle is qualified in the way just suggested to exclude from its operation external preferences (of which this fanatical preference is one). [Hare 1988: 233]

[176] He also thinks that the distinction between intuitive and critical thinking that will be discussed in §4.2 can help him here, in that if malevolent preferences can be excluded at the intuitive level it is not necessary to have a bar on external preferences at the critical level. My discussion in the text in the current section can be assumed to be concerned entirely with the critical level.

[177] What Hare (1988: 233-4) ends up saying is that he prefers not to take Gibbard's way with external preferences, because he wants somehow to deal with the problem of the fanatic at the same time and in the same way. But this is an odd way to respond, because it treats the question of whether external preferences are excluded from (CR) as somehow to be decided at Hare's discretion. On the contrary, if (CR) is false in its unrestricted form but true in its restricted form then external preferences must be excluded; there is no choice in the matter. Note that Hare does not defend the view that (CR) does apply to external preferences; he simply says, unsatisfactorily, that he thinks he can deal with external preferences in a different way.

[178] See Hare 1963: ch. 9 and 1981: ch. 10.

Now Hare's point here is that the fanatic himself already has the fanatical preference, and so if he is doing moral reasoning it does not matter that it is excluded from (CR).[179] So the fanatic will come to a different moral conclusion from ours, and there will be no way of deciding the question between us.

It is worth noting that the exclusion of external preferences makes another positive difference to the fanatic case that Hare has apparently not noticed. Sometimes (e.g. 1981: 180) when Hare discusses fanatics he imagines that problems arise when the preferences of different fanatics who agree with each other are combined to outweigh the preferences of the fanatics' victims. But now that cannot happen: the fanatics will not conditionally reflect each other's fanatical preferences (since they are external), and so each fanatic will be left to compare his own fanatical preferences with the preferences of all the victims. This makes the problem of fanaticism far less pressing, because it rules out the most obvious problem cases: even the fanatic will not be led to think that it is right for him to terrorise his victims unless his preference alone (disregarding the preferences of his fanatic allies) is stronger than the preferences of the victims.

Hare is well aware (1981: 181-2) that it is very unlikely that the fanatic's preference will really be stronger. To anticipate what I will discuss in more detail in §4.3, the preference utilitarian has to be careful in how she thinks about inter-personal comparisons of preference strength. The only plausible way of getting inter-personal comparison going is to take it as axiomatic that each person's preferences, considered overall, are as strong

[179] To be absolutely clear, saying that the fanatical preference does not conditionally reflect means that it does not do so insofar as it is an external preference. It may well be that the fanatical preference is in part a personal preference: e.g. the fanatic prefers that everyone listens to Wagner, and so he prefers that he listens to Wagner. That personal preference is conditionally reflected (unless it is irrational). This is a point first made by Narveson (1978: 251): 'ideals do not have zero weight against interests, but instead have some weight. How much? I suggest that it would be *the weight which they have insofar as they themselves are interests*.'

as each other person's. It follows then that we compare the strength of different people's preference by first comparing each preference with other preferences of the person whose preference it is: a preference's strength in inter-personal comparison is determined by its relative strength in intra-personal comparison. Since the preference of a victim not to be victimised by the fanatic will be one of her strongest preferences (in the more dramatic examples), the fanatical preference would have to be more or less the fanatic's strongest preference in order to outweigh it, which is implausible, and even then it could not justify victimising more than one person. So no matter how strong the fanatic's preference is, he will not be able to justify to himself the policy of locking up homosexuals. I do not suggest that these points defuse the problem entirely. It is still the case that the fanatic reaches a different moral verdict from the non-fanatic, because he gives some weight to his own fanatic preference, whereas the non-fanatic gives fanatical preferences no weight. This difference in moral verdict is itself puzzling, because it suggests a lack of moral objectivity, even if what the fanatic is able to justify to himself is not dramatically worrying.

Now in order to understand just how bad this problem is and whether it can be dealt with we need first to consider the different varieties of fanatic. Hare (1981: 170-1) distinguishes between three varieties of fanatics. The first are impure fanatics, those who have not engaged in critical thinking. What Hare is imagining here is someone who holds a moral principle without considering whether and how it can be justified. This kind of fanatic presents no problem, because we can argue with him precisely on the basis that he holds a principle arbitrarily. There are two kinds of pure fanatic who do engage in critical thinking. One kind rejects the argument for utilitarianism discussed above, and so we can argue with him by defending that argument as above. The final kind of pure fanatic

accepts the argument for utilitarianism, yet holds that his fanaticism is compatible with it. It is presumably this kind of fanatic that Hare is worried cannot be argued with.

What is strange is that Hare seems to assume throughout (1981: ch. 10) that the fanatic views his fanatical preference as a moral principle. But in that case it is easy to argue with the fanatic, because we have seen (in Chapter 3) that the only source of moral principles is universalizability. Unless the fanatic has an argument for his fanatical principle we can simply say that his belief that this moral principle is justified is false, and that is more than enough reason to criticise him. So the really challenging kind of fanatic is one who does not see his fanatical preference as a moral principle, but rather says, 'This is a preference I just happen to have, in the same way that I just happen to have various personal preferences in a way that does not require justification.' It is when the fanatic says this that his preferences seem most on a par with personal preferences, and so he seems on the firmest ground on counting them in his moral deliberations.

I do have a proposal for how the fanatic can nevertheless be made to see that it is incompatible with Hare's argument that he should take even his own fanatical preference into account when morally deliberating. The basic idea is that the fanatic reflects on the fact that others will not have reason to take his fanatical preference into account, because it does not conditionally reflect. Now the fanatic should ask himself, 'How can it be that I am justified in thinking that this preference is morally relevant, when I know that everyone else is justified in thinking it not morally relevant (and not because they are at any epistemic disadvantage)?' We can put matters like this: the fanatic's first-order moral judgement that his fanatical preferences matter morally appears to be universalizable; but the second-order moral judgement that his fanatical preferences ought to be judged to be morally relevant is not universalizable. But (U) applies just as much to second-order

judgements as first-order ones. The fanatic cannot prescribe his treating his fanatical preferences as a reason in the situation where he is in someone else's shoes (because he knows that when in someone else's shoes he will and should follow Hare's reasoning using (CR), which omits his fanatical preferences), meaning that he does not universally prescribe treating those preferences as reasons. What I suggest then is that in order for a preference to be a moral reason there has to be a moral justification for treating it as a reason, meaning that (U) has to be satisfied at the higher-order level as well as the lower. If this proposal is followed it seems that even the fanatic must admit that his fanatical preferences do not matter morally: the fact that external preferences do not conditionally reflect ultimately allows us to fully deal with the fanatic, *pace* Hare (1988: 233). Another way of reaching this point is suggested by some remarks of Hare (1989a: 216): 'my own actual situation is one of those I have to suppose myself occupying'. What is interesting about that way of putting things is that it suggests that the agent has to treat her own preferences like those of everyone else, meaning that they do not get counted unless (CR) applies to them. This would be a quicker route to the desired solution.[180]

It is not only external and fanatical preferences that may be thought to cause trouble for Hare. There are anti-social preferences that are not fanatical, and which it is counter-intuitive for utilitarianism to give weight to, e.g. sadistic preferences. Hare (1989a: 220-1) gives several responses on this point: (i) it is very unlikely that the sadistic preferences would be strong enough to outweigh the preferences of the victims (see my discussion above); and (ii) there are significant side-effects to the sadist getting what he wants (I am not entirely sure what side-effects Hare has in mind here). (i) and (ii) may serve to show that counting the sadistic preferences does not lead the utilitarian to absurd

[180] See also Hare 1989a: 219-20. This is another case where the presentation of the earlier 'Ethical Theory and Utilitarianism' may be superior to that of *Moral Thinking*.

conclusions in practice, but we might still think it an embarrassment if sadistic preferences make any difference at all to the moral calculus. Hare has two other responses though: (iii) the sadist can be given substitute pleasures or be cured; and (iv) the sadist might lose those preferences if he were more rational, fully-informed etc. Now here we get very close to the idea that the sadist's preferences do not conditionally reflect. I conjecture that a large part of our resistance to counting the sadist's desires is precisely that we do not think it even in the sadist's interests that those desires are satisfied: we do not prefer that, were we in the sadist's shoes, we have those preferences satisfied. Rather, as in the case of addiction, we prefer that we receive treatment. Of course it is at least possible to imagine an incurable sadist (though that is presumably not the usual case), and it does seem that there are not good grounds to exclude his preferences. But equally we are likely to be sympathetic to the plight of the incurable sadist: we will not think it absurd that society should take his interests into account.

Hajdin (1990) has suggested that dealing with external preferences in the way suggested by Gibbard – holding that they are ruled out by correct application of (CR) – is not only the correct way to go but also leads Hare's theory towards hedonistic rather than preference utilitarianism.[181] The point is meant to be that the reason why we conditionally reflect preferences is that we realise what it would be like to have those preferences frustrated, which is an experiential question, and so linked to pleasure and pain rather than preference satisfaction in the broad sense. Hare (1999: 126-7) seems to accept Hajdin's reasoning. But I think that Hajdin is quite wrong about this. First, the case of the sadist is a nice illustration that we do not always conditionally reflect even experiential preferences. Second, we can sympathise with the predicament of someone whose

[181] Or strictly, a version of preference utilitarianism that is restricted enough to be equivalent to hedonistic utilitarianism.

strongest preferences are unsatisfied even if she never finds out that they are unsatisfied – this is just to adapt the most common counterexample against hedonistic theories of well-being (which shows that desire-based theories are superior). So it is neither necessary nor sufficient for a preference to be included in the scope of (CR) that it be experiential, and this is because we do not take a hedonistic approach to our own interests. Although Hajdin is right to think that Gibbard's approach allows Hare to deal with external preferences, it does not lead to hedonistic utilitarianism, unless we already have a hedonistic theory of well-being (and we should not).

d) Hare's argument treats preferences as the only inputs into moral decision-making.[182] We can ask both, 'Why are preferences sensible inputs at all?' and 'Why are preferences the only inputs?' The first question is easy to deal with: we have preferences and we cannot seriously contemplate that morality should disregard our **own** preferences – that our own preferences are irrelevant to what there is reason to do.[183] But Scanlon (1988: 146) asks whether preferences 'are the *only* considerations on which a rational decision about which universal imperatives to accept can be based and whether a rational and impartial decision can be made only by combining them in the way described'. Scanlon thinks that Hare is wrong to say that preferences are the only relevant considerations. Much of his case against Hare is based on giving examples where he thinks Hare's theory conflicts with our considered moral judgements. I will come to discuss such criticisms when I discuss the details of Hare's two-level account of moral thinking in §4.2; Hare (1988: 260-8) makes clear in his reply that this is how he would deal with them, showing that the considered judgements can (usually) be accommodated by his theory.

[182] That is apart from the formal considerations (U) and (CR), as Hare (1988: 268) notes.
[183] See also Hare 1981: 89-93.

Regardless of this, Scanlon's complaint can hardly be that he does not like the results of Hare's argument: he needs to explain why there should be some other inputs, and he fails to do this. There are, I think, two possibilities for what Scanlon has in mind. The first is that Scanlon takes considerations other than preferences to be moral reasons. The way to understand Hare's argument here is as a construction of reasons from preferences; Hare does not help himself to reasons at the start. And sometimes it seems as though Scanlon too is a constructivist[184]; but in fact as made clear by Scanlon's (2009) Locke Lectures, *Being Realistic about Reasons*, he is only interested in constructing moral rightness out of reasons, not in constructing reasons themselves. Given this disagreement between Scanlon and Hare concerning this fundamental question in meta-ethics, it is hardly surprising that they come to different conclusions about where to start in normative ethics. It seems that Scanlon's critique of Hare here is external rather than internal: it presupposes Scanlon's meta-ethics rather than Hare's. Scanlon does not provide any argument that Hare, given his expressivism, should admit anything other than preferences in the construction. To completely rebut Scanlon's critique here would require engaging with his (2009) meta-ethical stance; but having spent Chapters 1 and 3 above laying out the case for expressivism and Kantian constructivism against realism I shall rely on that rather than trying to determine exactly where Scanlon's meta-ethics goes wrong.

The other possible interpretation of Scanlon (1988) is that he is simply suggesting the use of Rawls' (e.g. 1999: 18-9) reflective equilibrium methodology in normative ethics. Hare's (1989a: 145-74) intemperate review of Rawls' (1999) *A Theory of Justice* criticises Rawls precisely for using this methodology, claiming that it makes Rawls an intuitionist. The idea of reflective equilibrium is that we take all of our relevant judgements and,

[184] Readers of Scanlon 1982 and 1998 might easily have got this impression.

working back and forth between the more abstract, theoretical judgements and our considered judgements about particular cases, revise our judgements until they are in equilibrium.[185] From this point of view the intuitive appeal of particular normative principles (having been corrected for whatever distortions or biases can be detected in light of the rest of our total theory) has to be balanced against whatever considerations lead us towards utilitarianism. It is perhaps not worth debating whether Rawls was in any interesting sense an intuitionist, as Hare alleged: considered judgements are not precisely intuitions (because they are considered and because the judgements stemming from identifiable sources of bias etc. have been weeded out), and they are only defeasible considerations. Hare's complaint, however, can be recast in terms of the epistemic weight which is granted to the considered judgements. One notable feature of the reflective equilibrium view, which is evident in *A Theory of Justice*, is that where we end up depends on the considered judgements we start with – I will call this 'Contingency'.[186] Contingency is problematic for two reasons. The first is that it leads to a loss of moral objectivity: what is a reflective equilibrium for one person (or group) may not be for another, because they start with different considered judgements (and neither party's judgements are explained by bias, ignorance etc.). This means that the reflective equilibrium procedure cannot serve as a genuine construction of normative truth (unless normative truth is relative).[187] So we

[185] My discussion of reflective equilibrium below draws not only on Hare and Rawls, but also on Daniels (1979), Hooker (2000) and Kelly & McGrath (2010). I do not, however, attempt a comprehensive survey of the large related literature.

[186] As Harsanyi (1975) argues, the obvious way of setting up Rawls' Original Position and veil of ignorance leads to utilitarianism; it is only because Rawls takes anti-utilitarian considered judgements as inputs that he tunes the Original Position so as to produce a non-utilitarian outcome (his Difference Principle). Rawls is quite open that this is his procedure, though this means that his construction of the principle of justice does not itself provide an argument against utilitarianism. The crucial point here is that Rawls assumes that he and (enough of) his readers agree in having the relevant anti-utilitarian considered judgements, and it is only because they are there as inputs that the eventual theory is not utilitarian.

[187] See Hare 1981: 12. It is not at all clear whether it is intended to do so. The one attempt Rawls made to state his meta-ethical views, 'Kantian Constructivism in Moral Theory' (1980), does not sufficiently engage

have a dilemma for reflective equilibrium: the first, constructivist, horn leads to relativism; the second horn makes epistemological demands which the theory cannot meet, as I now demonstrate.

If the reflective equilibrium view does not construct (non-relative) normative truth, we can ask the question of how it is epistemologically justified: why we should expect it to lead us to true moral judgements.[188] This leads to the second problem with Contingency, which is that considered moral judgements are not properly treated as evidence by the reflective equilibrium view. In general I should count my perceptions of the world as no better evidence concerning its actual state than the perceptions of others who are my epistemic peers (people who I have no reason to think are worse at perceiving than I am). For an illustration of this, imagine you are waiting with a friend at a bus stop; a bus approaches, but it is far enough away that reading the bus's number is difficult; if you and your friend disagree on what the number is, then it is irrational to set greater store in your own perception unless you (justifiably) believe that your vision is better. And yet the proponents of the reflective equilibrium method do not treat their own considered judgements as being on a par with those of their epistemic peers: they hold that one's goal is simply to reach reflective equilibrium between one's ethical theory and **one's own** considered judgements. Of course, I do not advocate that the reflective equilibrium theory be adjusted to take into account everyone's considered judgements, because we have in fact been given no reason to think that considered judgements carry epistemic weight in

---

with the meta-ethical literature to make it clear whether Rawls is indeed a constructivist or what he thinks constructivism amounts to. For this reason both Scanlon and Korsgaard can both think of themselves as drawing their meta-ethics from Rawls, though the former is a realist and the latter an expressivist/constructivist (as discussed in §3.4).

[188] Hooker (2000: 13) says that the considered judgements (he calls them 'convictions') must have independent credibility. What I say below shows that they do not have this credibility in the way that the reflective equilibrium view assumes.

the first place. My point is that even advocates of reflective equilibrium cannot really believe that considered judgements have the required epistemic weight.

Up until now, I have omitted discussion of the most crucial point thought to be in favour of reflective equilibrium. Hooker puts the point thus:

> Note first that we cannot evaluate our evaluative beliefs, or anything else, *from a completely non-evaluative point of view*. If we take up a point of view stripped of all evaluative conviction, we have no basis for evaluation. [2000: 11]

Or in other words, as Margaret Thatcher used to say, There Is No Alternative (TINA). Even if reflective equilibrium does not secure moral objectivity, and even if it does not have a well-worked-out epistemology, TINA: reflective equilibrium is the best methodology for normative ethics because it is the **only** methodology. If Hooker is right that it is impossible to evaluate our ethical theory starting off from a completely non-evaluative point of view, then every starting point will be evaluative. But if we are prepared to countenance an evaluative starting point, what reason can we have for privileging one of our evaluative views to use as a starting point over others? Why, for example, take some abstract principle as a starting point and not allow in convictions about particular cases? The method of reflective equilibrium is precisely that which refuses to privilege any starting points without good reason. We are right to allow our convictions about cases some initial credibility, because we have to allow **some** evaluative starting points to have some credibility, and there is no reason to discriminate against convictions about cases.

How then will the advocate of reflective equilibrium explain where Hare is going wrong? A first move will be to say that Hare has an evaluative starting point, universalizability, which he privileges over convictions about cases for no reason. Hare will reply that universalizability is not an evaluative starting point, but rather a

formal/linguistic/conceptual one. At this point the shmoralising objection raises its head (see §3.0): even if, given our moral concepts, universalizability is a conceptual truth, there is a substantive question concerning whether to retain those concepts as guides to our normative practice. By assuming that we should retain those concepts, Hare smuggles in a privileged evaluative starting point. The reply that Hare (1981: ch. 11) seems tempted by is to try to give a prudential argument for retaining our moral concepts. But of course, even if it makes sense to base morality on prudence, these prudential concerns are still evaluative and so should not be privileged. I will not consider any further whether Hare might have an adequate reply to this line of attack, because the point of my attempt to derive universalizability from a Kantian transcendental argument in Chapter 3 was precisely to avoid the shmoralising objection and yet still to find a non-evaluative starting point for moral justification. Use of the TINA argument against me would assume that I had failed; so I can at least for now rely on the Chapter 3 argument as an explanation of why we do not have to resort to reflective equilibrium. In short then, the complaint that we are starting from preferences and disallowing considered judgements as inputs to moral theory is only reasonable if the overall argument I am giving for utilitarianism fails, and so criticism must fall on some point of the argument itself.

4.2 TWO LEVELS OF MORAL THINKING

Hare defends a rather unusual form of indirect utilitarianism, distinct from rule-utilitarianism, though probably overlapping with it on many issues. Hare distinguishes between two levels of moral thinking, which he calls 'critical' and 'intuitive'. Critical thinking is the kind that we are engaged in when we try to resolve conflicts between *prima facie* duties (Hare 1981: 26-7) or, most crucially, when we question our moral principles

and try to decide whether they are genuinely justified. The normative regress argument considered in Chapter 3 seems to fall in with critical thinking in this sense (though it is of course connected to meta-ethics). And Hare's argument for utilitarianism is also a piece of critical thinking. Now Hare's view is that most of our moral decision-making does not in fact involve critical thinking. The immediate question for him then, is whether we **should** try to think critically all the time. That is, should we make all of our moral decisions in accordance with (PU) (or (PU*) if my revisions to Hare's argument are successful)?

Crucially, Hare's answer to this question is strongly negative. His reason for saying that we should not always use utilitarianism to make decisions is not, in its basic form, original. The idea is that utilitarianism itself tells us not to (always) make decisions on the basis of utilitarianism. In Bentham's words:

> 'The principle of utility (I have heard it said) is a dangerous principle: it is dangerous on certain occasions to consult it.' This is as much as to say, what? that it is not consonant with utility, to consult utility: in short, that it is *not* consulting it, to consult it. [Bentham 1907/1823 ch. 1, para. 13 fn.]

Now Bentham's formulation is deliberately paradoxical, suggesting that he himself does not believe it makes sense to say that it is ever 'not consulting [utility] to consult it'. The problem is that if, faced with a particular decision, I decide not to make that decision in a utilitarian way, because that maximises expected utility, I must be confused. Either I can use the principle of utility effectively in that case, or I cannot: if I can, then I should use it; if I cannot, then how do I know that not using it will maximise expected utility? But there is nothing incoherent in thinking that, on some **future** occasion, I will maximise expected utility by not applying the principle of utility: I can think myself capable of successfully

applying the principle now, yet incapable of applying it successfully on that future occasion.

Why, however, should I ever think that my ability to successfully apply the principle of utility will be impaired in future? Concrete cases we have to make moral decisions about often involve individuals and groups we have partiality towards (most notably, ourselves), which makes it unlikely that we will be able to make impartial decisions.[189] Apart from this there is the standard point that the circumstances of decision (not enough time, lack of information) may get in the way of applying the principle of utility effectively. So we may well do better in such cases to rely on *prima facie* principles that coincide in their recommendations with the principle of utility for many cases. Moral reasoning that relies on *prima facie* principles is what Hare calls the 'intuitive' level of moral thinking. There will be exceptions, where we are able to apply the principle of utility successfully, or at least successfully enough to produce more utility than we would if we applied our *prima facie* principles. The utility lost by following the *prima facie* principles in those exception cases must be outweighed by the utility gained by applying them in cases where we would not apply the principle of utility successfully. Now we might wonder why it is not best to apply the *prima facie* principles when that it best, and apply the principle of utility when that is best. The problem is that we cannot be expected to reliably distinguish between these two types of cases, because the weaknesses that get in the way of successfully applying the principle of utility are often subtle. So we will do best if we have to be very confident in our ability to successfully employ the principle of utility in a case before we override a *prima facie* principle.

[189] See Hare 1981: 44.

There is, however, another problem which faces the utilitarian who has gone through all these thoughts. Suppose that I have formulated some *prima facie* principles for myself, and I come upon a case where these principles seem to conflict with the principle of utility (it is not immediately relevant whether I am right about this). In the face of such an apparent conflict it may seem that as a utilitarian I am duty bound to follow the principle of utility: I know that (U*) is the right way to make moral decisions, and so surely I should follow it, even if I also know that on average I do better to stick with my *prima facie* principles. But if I always resolve apparent conflicts between the principle of utility and my *prima facie* principles by following the principle of utility, I might as well not have the *prima facie* principles at all.[190] What this objection shows is that it is not enough for me to now intone that I will in future side with my *prima facie* principles against the principle of utility, because when the moment of decision comes it will make no sense, given my acceptance of utilitarianism, to follow my earlier intention. So in order to actually improve on always applying the principle of utility, I have to have a way of binding myself in advance to follow the *prima facie* principles I lay down. That is to say that the *prima facie* principles have to be internalized: I have to care about following them so that I am motivated to do so (and so I feel guilty if I do not follow them etc.).[191]

Most of the above is fairly standard amongst indirect utilitarians, including rule-utilitarians. They will agree on the fact that although morality is fundamentally impartial (impartial at the critical level), *prima facie* principles that are not impartial will often be selected e.g. ones that promote (to a moderate degree) friendship and family.[192] The

[190] This is the rule-worship objection, which aims to show that indirect utilitarianism collapses into direct utilitarianism. See Smart 1956.
[191] See Hooker 2000: 93-4.
[192] And, given Hare's distinction discussed below between different senses of 'right', Hare can agree that it is not always right to act impartially.

difference between Hare and rule-utilitarians is that on his view, at the critical level, we have to make decisions concerning which principles to (try to) internalize, and to do so by directly considering the expected utility of such acts of internalization. Since my actions typically only make a difference to which principles **I** internalize, the relevant effects for this utilitarian calculation are those of **my** having the rule internalized, not the effects of **everyone** or **most people** having the rule internalized. It may be that many people have the same rule internalized that I am considering internalizing, and whether or not this is the case may factor into my decision. In contrast, rule-utilitarians are after a criterion of the rightness/wrongness of actions, and that criterion involves the consequences of the general (rather than individual) internalization of rules. For example, here is (the beginning of) Hooker's formulation of rule-consequentialism:

> An act is wrong if it is forbidden by the code of rules whose internalization by the overwhelming majority of everyone everywhere in each new generation has maximum expected value in terms of well-being (with some priority for the worst off). [Hooker 2000: 32]

Let us ignore the non-utilitarian aspects of Hooker's formulation and focus on the consequences of Hare's individualism. It may seem that Hare will not be prepared to give a formulation of his view in terms of rightness/wrongness that is anything like Hooker's, because Hare links rightness so closely with critical thinking and so with (PU)/(PU*). But Hare (1988: 261) admits a sense in which 'right' means 'intuitively right', which is to say 'right according to the rules which are best to internalize'. So Hare could formulate his view, employing that intuitive sense as follows:

(W) An act is wrong if it is forbidden by the code of rules whose internalization by the agent has maximum expected utility.

The first important consequence of Hare's view is that it can be more demanding than a rule-utilitarianism like Hooker's. The reason rule-utilitarianism is not particularly demanding is that if the overwhelming majority of people **did** internalize the code of rules whose internalization by the overwhelming majority of had maximum expected utility, then problems such as extreme poverty would be solved without anyone having to make extraordinary sacrifices. Or in other words such a code of rules would not need to be particularly demanding in order to solve such problems. In contrast, the Harean cannot rely on the rule that **she** has to follow being selected as the one that would solve such problems if the overwhelming majority internalized it. She has to decide which rule to internalize knowing that most people do not and will not do their fair share towards alleviating extreme poverty, and it seems that the rule she is thus required to internalize and follow will thus be very demanding. She will have to do more than her fair share to make up for the fact that others do less than theirs.[193] This makes Hare's view almost as demanding as straightforward act-utilitarianism.[194]

Is this a bad result for Hare? Certainly we have strong intuitions that tell us that morality does not require us to sacrifice most of our wealth for the sake of those in extreme poverty on the other side of the world. If we were allowed to employ the reflective equilibrium method, then it might be, because a very demanding theory is

[193] Hare (1981 ch. 11) has a response which aims to alleviate this demandingness. He thinks that in practice even if it is possible to internalize a very demanding principle, it is not a good idea to do so unless one will be able to (motivate oneself to) follow it, because otherwise the main result will just be that one feels guilty all the time, making oneself miserable to no good purpose. Carson (1993) explains that if the stakes for others are high enough (as they often will be where extreme poverty is at stake and the agent is affluent), then even if internalizing the very demanding principle motivates the agent to follow it only occasionally the benefits to others will be larger than the cost to the agent in additional guilt. I agree, with the caveat that sometimes internalizing an unrealistically demanding principle will make an agent's behaviour worse rather than better, because it will induce despair at the impossibility of complying with the demands of morality. How common this latter psychological mechanism really is will determine whether Hare or Carson is closer to being right here. See here Kagan 1989: 35. In the text, however, I assume that Carson is correct and consider what follows in that case, since it is more interesting and more of a threat to Hare's overall view.

[194] Not quite as demanding, because it is not plausible that we will be able to internalize act-utilitarianism itself, or that it would be a good idea to do so.

counterintuitive. But that method has been ruled out in §1.1 above. Hare's view of intuitions is that they simply represent the *prima facie* principles that we have internalized. Since it is reasonable to assume that the *prima facie* principles which we have internalized approximate fairly well to the ones which maximise utility (since their doing so is one, though not the only, explanation of why we have internalized them), our intuitions are (weak) evidence of the moral truth. But when we have shown through critical thinking that it is better to internalize a different principle, that evidence is undermined. For Hare then, a counter-intuitive moral result is never evidence that the theory is wrong.[195]

It is worth noting, however, that there is a flip-side to the demandingness of Hare's theory that may give it an intuitive advantage over rule-utilitarianism, at least balancing out the problem of over-demandingness. There can be situations where an individual is severely suffering and I can, at trivial cost to myself, remove this suffering; and yet the rules selected by rule-utilitarianism do not require me to help this person because if those rules were generally accepted that person's suffering would be removed without my having to do anything. Rule-utilitarianism says that it is not wrong for me to ignore suffering if I have already done my fair share (according to the best rules) to alleviate that kind of suffering. Imagine that there are 10 children drowning in a pond, and you are amongst 100 onlookers who could easily save them.[196] A simple rule-utilitarian rule in such a situation would demand of the onlookers that they be prepared to save one of the

[195] See also Williams 1988: 191-2. Matters are slightly more complex in that sometimes moral intuitions reveal linguistic/conceptual intuitions, but the basic point stands – see Hare 1981: 12-3. This aspect of Hare's view also explains why he would not be persuaded to move away from utilitarianism towards prioritarianism by the kinds of intuitions concerning fairness considered by Hooker (2000: 65) which lead to the non-utilitarian aspect of his theory. I return in §4.3 to the attempt to show that (Hare's version of) utilitarianism is more egalitarian than normally assumed.

[196] This case is inspired by Kagan (1991: 924-5), and in turn by Singer (1972). The difference from Kagan's example is that he considers only two children and two onlookers. I want to make it clear that in my case if the overwhelming majority of people accepted the rules all the children would be saved, whereas in Kagan's case it could just unluckily occur that the other onlooker is one of the few who don't accept the rules even when they are generally accepted.

children. But suppose that in fact you save one of the children, and then realise that none of the other onlookers is helping to save the children at all. You could still easily go back and save another child (the remaining children haven't drowned yet). But since, by saving one child, you have done (at least) your fair share, rule-utilitarianism apparently permits you to allow the other children to drown, with the other onlookers (and not you) being to blame when they do. Common-sense morality, in contrast, says that you are absolutely obliged to carry on saving the children as if the other onlookers were not there; once you realise that they are not helping their presence is irrelevant, as are any thoughts of fair shares.

I do not by any means insist that the rule-utilitarian is committed to a permission not to save additional children in the case I describe. Some will bite the bullet[197] and others will attempt to adapt rule-consequentialism in order to avoid the counter-intuitive permission. Hooker attempts to do so by suggesting that rule-consequentialism would prescribe the following rule:

> Over time agents should help those in greater need […] even if the personal sacrifices involved in helping them add up to a significant cost to the agents. The cost to the agents is to be assessed aggregately, not iteratively. [Hooker 2000: 166]

The idea here is that there is an overall threshold for self-sacrifice (over a lifetime) that one is not obligated to go over. The rule imposes a duty to save all the drowning children in my case, unless you have already gone over the threshold (or do so after saving some but not all of the children). But common-sense morality does not recognise any such threshold, as Hooker (2000: 168) recognises. So now we have a third kind of counter-intuitiveness. To recap: (i) over-demandingness; (ii) under-demandingness; (iii) thresholds.

[197] See for example Murphy 1993, and for discussion Mulgan 1997.

It is far from obvious that (iii) is the least worrying of these. I would be inclined to say that if we have to clash with intuition here it is easiest to reject the over-demandingness intuition, because that can be dismissed as complacency. The other relevant point in this discussion is that it is now far less clear that we have a contrast between Harean indirect utilitarianism and rule-utilitarianism. What Hooker (2000: 166) claims is that the reason we should choose his suggested rule over a more demanding rule is that the more demanding rule would have higher internalization costs, but internalization costs can factor in just as much in Hare's view. Hooker also notes that the internalization costs apply many times over because the rule has to be internalized by many people. But we might reply that just as the internalization costs of lots of people internalizing the rule are higher, so proportionately are the benefits, so the Harean can simply borrow Hooker's argument (if it works). Whether or not Hooker is right about the best rule for the rule-consequentialist, it now seems that the Harean will be able to apply the same reasoning.

Once we take into account the second major consequence of the individualism of Hare's theory, however, we realise that the original verdict was correct: the Harean theory can indeed be more demanding than rule-utilitarianism, **but only for some people**. A crucial consequence of (W) is that what it is wrong for me to do depends in part on facts about my circumstances, insofar as they affect the expected utility of my internalizing various different principles. Carson (1993) exploits this fact in his argument that Hare's utilitarianism is more demanding than Hare thinks: if an agent is affluent, then her internalization of demanding principles has significantly greater benefits than internalization by one less affluent; whereas in contrast the internalization costs are not thereby greater for her. It follows that applying utilitarianism at the critical level results in her being required to internalize more demanding principles than other people, meaning

that for her Harean utilitarianism is more demanding than rule-utilitarianism. Let us also note that internal factors, such as the immutable aspects of different agents' characters, can make a difference to what principles they ought to internalize. Hare (1981: 200) says, 'I may well think it right to have personal but fairly general principles which I do not expect to be the same for everybody, but which are suited to my own capacities and condition.'

Is this Harean doctrine counter-intuitive? Not particularly, I think: it can be put as the platitude, 'with great power comes great responsibility.' Is it objectionable for any other reason? We might think that somehow universalizability is transgressed by allowing that what is wrong for A, with her immutable character, is permissible for B with his. Winch (1965) presents exactly such a case – his Billy Budd example – as a purported counterexample to universalizability. But Winch did not have Hare's (U) in mind (I shall not enter into whether he had a good counterexample to the formulation that he did have in mind): it is perfectly in line with (U) that an agent's character and condition should be morally relevant and serve to distinguish the cases, allowing different moral judgements about them. Of course there will also be some principles which only work if there is nearly full acceptance, such as rules of trust and truthfulness (Hare 1981: 200).

In comparing Harean utilitarianism with rule-utilitarianism we see that Hare's theory is indeed more demanding for some people. It is not clear to me whether this results in it being less intuitive; for Hare this question will hardly matter, though the rule-utilitarian who believes in reflective equilibrium will care about it a great deal (e.g. Hooker 2000: 101). But Hare's theory does appear to have one great advantage over rule-utilitarianism. It is never exactly clear **why** the rule-utilitarian should think that what rules it is best for society as a whole to internalize makes a difference to what she ought to do. After all, she has to sort out her own principles, not society's. In short, Hare can give a neat

explanation of how his indirect utilitarianism follows from utilitarianism at the critical level, whereas the rule-utilitarian cannot do so.

There is a problem, however, which Hare incurs and which the rule-utilitarian does not. This is an alleged psychological instability in imagining that an agent can switch between critical and intuitive thinking regarding a principle. This objection is pressed most forcefully by Bernard Williams:

> It is artificial to suppose that a thorough commitment to the values of friendship and so on can merely alternate, on a timetable prescribed by calm or activity, with an alien set of reflections. Moreover, since the reflections are indeed alien, some kind of willed forgetting is needed, an internal surrogate of those class barriers on which Sidgwick relied, to keep the committed disposition from being unnerved by instrumental reflection when they are under pressure. [Williams 1985: 109]

There is a danger that such an objection will prove too much. Effectively the aspect of Hare's view that Williams is criticising is that Hare gives an instrumental account of the virtues: what makes commitment to a principle a virtue for someone is the expected utility of her commitment to it. But Hare is hardly unique in having such a view. Consider, for example, Hume's (2000/1739: 3.2) view in the *Treatise* that justice (respect for property) is an artificial virtue. Hume specifically entertains cases where we are conscious that an action is just but also that it is not useful in the way that just actions usually are; in such cases we have less approbation for the just action. So Hume seems quite happy with the conclusion that Williams seems to think unacceptable: that reflection will unnerve the committed disposition. And why should this be a problem? If the psychological reality is that there is indeed some kind of inner conflict and doubt when we see that being virtuous is not useful on some occasion, what exactly is the problem? I think that Williams has in

mind a very neat and compartmentalised psychology that keeps critical and intuitive thinking always apart. But Hare does not have that in mind: he is aware, like Hume, that on some occasions we will engage in critical reflections that undermine our commitment to some extent, and sometimes this will even lead us to violating our commitments. Effectively Williams is objecting that the theory predicts a somewhat inelegant psychology; but given that this is the psychology we find in reality that seems like a feature rather than a bug (see Hare 1988: 289-90).

There is one other possible interpretation of Williams' complaint: that he thinks that reflection on why a virtue is a virtue always undermines it. It is hard to see why this should be so, unless to have a virtue just is to believe that the principle it embodies is fundamental and not in need of justification. But this is psychologically unrealistic, and so the burden is on Williams to explain why it should be somehow impossible to maintain a virtue whilst understanding why it is a virtue. Is the choice between the crude Lockean view that property rights are natural on the one hand, and having no respect for property on the other? Is it credible that Hume had no respect for property? Williams (1988: 190) clarifies that what he means is that the thoughts are not stable under reflection, which is not meant to be a psychological point. He means that seeing an action as just is one way of seeing it, which cannot be combined in the same thought with the instrumental view of justice. But again, what is the problem with conceding this? We have in this sense a fractured moral outlook, but Williams has neither argued that this is impossible or that it is inferior to some other, unfractured outlook. In contrast, Hare has argued that it is superior to the unfractured outlook of straightforward act-utilitarianism.

I do not want to suggest that there is nothing puzzling about the moral theory we end up with on Hare's view. The question is whether this puzzlement is something that we

live with anyway in our moral thinking, and whether, if it is, there is any way of purifying our thinking so as to do away with it. To illustrate, I take an example from William James:

> [On the hypothesis that] millions [are] kept permanently happy on the one simple condition that a certain lost soul on the far-off edge of things should lead a life of lonely torture, what except a specifical and independent sort of emotion can it be which would make us immediately feel, even though an impulse arose within us to clutch at the happiness so offered, how hideous a thing would be its enjoyment when deliberately accepted as the fruit of such a bargain? [James 1891: 333][198]

Now Hare can say about this case that, since it is such a fantastical hypothesis, our intuitions have been formed so that we do indeed consider it repugnant to enjoy the happiness of living in such a society, and it is right that they have been.[199] In other words Hare can apply (W) to explain the sense in which that enjoyment is wrong. The real challenge is this: if such a society really were on offer, and we knew this when deciding what principles to internalize, would it not be better to internalize principles which did not produce such repugnance, given that the overall gains in utility on offer are supposed to be so great? But do we not also want to say that the inhabitants of such a society do act wrongly, even if the best principles **for them** to internalize are ones which do make them feel any repugnance?[200] If so, then (W) does not reflect our only intuitive sense of the word 'wrong'. There is also the following:

(W*) An act is wrong if it is forbidden by the code of rules whose internalization **by me** has maximum expected utility. (Reading 'me' *de dicto*, not *de re*).

[198] This example was famously made vivid by Ursula Le Guin's (1973) short story, 'The Ones Who Walk Away from Omelas'.
[199] Compare Hare's (1989b: 148-66) article 'What Is Wrong with Slavery?'.
[200] This might not be so: perhaps the costs of internalizing those principles are too high; but I will assume otherwise for the sake of argument.

And perhaps what Hare (1988: 261) had in mind was more like (W*) than (W). I do not think that the profusion of senses of 'wrong' necessarily counts against Hare, in that we do seem to have a use for a contrast like that between (W) and (W*).[201] Perhaps we incline more towards giving verdicts in terms of (W*) when our moral feelings are particularly strong. And yet there is something odd about (W*): it captures a kind of judgemental attitude which cannot be endorsed on reflection. Perhaps we have no right to condemn the denizens of James's imaginary society; but we do so, and Hare can explain why even if he cannot justify it.

4.3 FAIRNESS, JUSTICE AND EQUALITY

In this section I will briefly consider two important and connected objections to utilitarianism, due to Rawls and Korsgaard, and explain how Hare's version of utilitarianism is perhaps uniquely well-placed to deal with them. In dealing with these arguments I aim to show that Hare's utilitarianism does not conflict with fairness, justice or equality in their moral senses. I then try to show that it is compatible with these values in a political sense as well, by deriving an egalitarian theory of distributive justice.

Perhaps the most notable contemporary objection to utilitarianism (and consequentialism in general) is that it does not respect the separateness of persons.[202] This criticism originates with Rawls (1999: 23-4, 26), who argues that if utilitarianism is derived from an impartial spectator view (where it is assumed that morality is social rationality and is thus a simple extension of prudence, where we imagine that a single individual has everyone's preferences) then it 'does not take seriously the distinction

[201] Note that the same kind of contrast is possible for the rule-utilitarian.
[202] Hare does not have a great deal to say about this objection, but see 1989b: 79-80 and 1988: 256-7.

between persons'. Now this is a fair criticism of the impartial spectator view: to **assume** that morality is social rationality (understood in this way) is to **assume** that the distinction between persons does not matter (in a particular way), which is indeed not to take it seriously. But as Rawls himself notes, the impartial spectator view is only one route to utilitarianism; and we should note that it is not Hare's route, since Hare derives (PU) from (U) and (CR) rather than assuming it.

How then is it that critics of utilitarianism think that Rawls' criticism of the impartial spectator view extends to utilitarianism in general? One thought is this: if the impartial spectator view does not respect the separateness of persons, then neither does any equivalent view. But this reasoning is bad: there can be valid and fallacious proofs of the same conclusion, and that conclusion itself is not imperilled by the existence of the fallacious proof (so long as the valid proof is available). The critics might reply that I have incorrectly explained how the impartial spectator view is problematic: the problem is not that it **assumes** that the distinction between persons does not matter (in a particular), but simply that it **concludes** that it does not matter (in that way). The latter is a "problem" shared by all versions of utilitarianism, since they all hold that the separateness of persons does not matter in that way: it does not get in the way of aggregation. Interpreted this way, however, the complaint simply begs the question against the utilitarian, because it points to no error on the utilitarian's part that is distinct from utilitarianism itself.

Taurek (1977) expands on Rawls' complaint in a different and less question-begging way. He claims that utilitarianism, and consequentialism in general, depends on the idea of impersonal value. That is, it assumes that we can permissibly think of states of affairs as valuable *tout court* rather than from the point of view of a particular individual. He imagines the utilitarian justification of aggregation to be that it maximises this impersonal

value. The problem is that insofar as e.g. preference satisfaction is valuable, it is valuable to the person whose preferences they are; we cannot simply infer that by adding together such personal values we obtain an impersonal value which it makes sense to maximise. Again, I agree that this is a cogent criticism: it is a mistake to **assume** that there is such a thing as impersonal value to be maximised. To do so ignores the fact that value is ultimately personal.

But aggregation does not have to be aggregation of value. The way that Hare's derivation of utilitarianism works is that his proof forces us to acknowledge that the preferences of other people provide us with moral reasons just as much as our own preferences do (ignoring complications concerning external reasons etc.). Taurek could have no objection to taking someone else's interests as providing moral reasons. It is only at that point that aggregation takes place: I recognise many (conflicting) reasons, provided by the preferences of the different people involved, and I have to resolve this conflict. I do so in the same way that I resolve other conflicts between reasons, by balancing them depending on their strength. The strength of those reasons depends on the strength of the underlying preferences (there is no other basis to determine the strengths of the reasons). Such aggregation of reasons does not presuppose impersonal value (though it does **construct** it), nor does it seem to ignore distinctions between persons in any objectionable way. Rather, as Hare (1989b: 79-80) says, his view operationalizes equal concern and respect for each person, which is exactly what Taurek says he cares about, by treating the interests of different people involved as generating moral reasons, whose strength is proportional to the importance of the respective interest.

The above arguments show that Hare's utilitarianism cannot fairly be criticised for not respecting the separateness of persons; moreover the fact that other versions of

utilitarianism **can** be criticised on these grounds shows that there is something unusual and advantageous about Hare's version. Exactly what that is becomes clearly when we consider a related objection to utilitarianism due to Korsgaard:

> Ask yourself, what is a reason? It is not just a consideration on which you in fact act, but one on which you are supposed to act; it is not just a motive, but rather a normative claim, exerting authority over other people and yourself at other times. To say that you have a reason is to say something *relational*, something which implies the existence of another, at least another self. It announces that you have a claim on that other, or acknowledges her claim on you. For normative claims are not the claims of a metaphysical world of values upon us: they are claims we make on ourselves and each other. It is both the essence of consequentialism and the trouble with it that it treats The Good, rather than people, as the source of normative claims. [Korsgaard 1993: 51]

I find these thoughts very persuasive (though I shall not defend them) except in so far as Korsgaard takes herself to have a general argument against consequentialism and utilitarianism.

In order to understand the issues here, we need to employ Kagan's (1989: 20-2) distinction between factoral and foundational issues in ethical theory. Factoral questions concern which facts or aspects of situations are morally relevant and what that moral relevance amounts to. Foundational questions concern the explanation/justification of these facts about moral relevance. When Korsgaard says that consequentialism 'treats The Good, rather than people, as the source of normative claims,' she is describing and criticising a **foundational** view, which we can call 'foundational consequentialism' or (more commonly) the 'teleological' view. But utilitarianism itself, as normally understood, is a factoral view: it says that right actions maximise expected utility, but it does not carry with it an explanation of this. One possible explanation, perhaps the one that most students of

utilitarianism first encounter, is that utilitarians identify value with utility ('welfarism') and combine this with foundational consequentialism to produce foundational utilitarianism: utility/welfare is the source of normative claims. However popular such a view may be or may have been, it is not universal to utilitarianism; in particular, it is not Hare's view. Indeed, reading the passage from Korsgaard above (up to her mention of consequentialism) she might just as well be describing with approval Hare's own view and his argument for utilitarianism. The mistake then is assuming that utilitarianism is a teleological theory; Rawls (1999: 26) explicitly makes this error. In fact Hare's variety of utilitarianism is best thought of as a deontological view (in the foundational sense that Rawls is interested in), because it effectively constructs the normative notions of reasons and rightness first, without taking the good to be prior to them. Again, insofar as we think of Korsgaard's criticisms of teleological theories as persuasive, they tend to show that a version of utilitarianism like Hare's, with its Kantian expressivist/constructivist foundations, is the best form of utilitarianism.

Dealing with the allegation that utilitarianism does not respect the separateness of persons goes a long way to showing that there is nothing unfair or unjust about inter-personal aggregation itself. It removes much if not all of the motivation for imposing some obstacle to aggregation such as Scanlon's (1982, 1998) Individualist Restriction. But there is perhaps a remaining concern that the results of applying utilitarianism are bound to show no concern for equality, and thus be unjust. Of course we can insist that since utilitarianism gives a central role to the equal consideration of interests it is egalitarian in the sense that matters, but the suspicion remains that distributive justice requires a more substantive form of equality.

Hedonistic utilitarianism notoriously faces the problem of Nozick's (1974: 41) utility monster. We can imagine a being capable of enormous amounts of happiness and suffering, and which we could easily make happier by giving it resources. If it is vastly more efficient at turning resources into happiness than any other being, then hedonistic utilitarianism will demand that all resources are given to it, since everyone else's immiseration will be outweighed by the happiness of the utility monster. Does preference utilitarianism face a similar problem? Normally the problem is presented by talking about some individuals being more efficient than others at converting resources into utility; then a utility monster is simply an enormously efficient being in this respect. What does this efficiency mean when utility is taken as preference satisfaction? Presumably the ideally efficient agent will achieve full satisfaction of her preferences at almost zero cost in resources. But in that case it will make no sense to give that agent any more resources; the efficient agent is not a utility monster.

Thinking about utility monsters shows us that in one important respect preference utilitarianism is more egalitarian than hedonistic utilitarianism: on the former theory there are no super-efficient utility monsters whose interests dominate everyone else's. The anti-utilitarian may object: 'Granted that efficiency does not create utility monsters, surely there can still be agents whose preferences are simply much stronger than everyone else's, and they will be utility monsters.' My reply is that we have no meaningful idea of the overall strength of an agent's preferences, and so it cannot be the case that one agent's preferences are stronger overall than another's. This is not intended as an empirical claim, but as a conceptual one. An intuitive way of putting the point is that one agent can attach a higher priority to a goal than another agent does, but it makes no sense to say that one agent attaches higher priorities to goals **in general** than another agent

does. Talk of priorities and preferences is interchangeable: the order of an agent's priorities is equivalently the order of her preferences. Our basic grasp of preference strength is intra-personal: the differing strengths of an agent's preferences are reflected in what that agent would choose when those preferences conflict. We can make sense of the fact that an agent's preferences are ordered because the priorities that an agent assigns to various goals will come to the fore when different actions or outcomes will satisfy different ones.

This insistence on the primacy of intra-personal comparisons of preference strength may seem a devastating point against utilitarianism, since it looks like the common refrain that inter-personal comparisons of utility are impossible. Utilitarians should say that inter-personal comparisons of utility are given sense by the basic commitment that they make to treating the preferences of individuals equally (as explained by Hare's argument in §4.1). This commitment would make no sense if it primarily meant treating particular preferences of equal strength equally, because that would presuppose a sense in which particular preferences of different people could be equal. The equal consideration of preferences that the utilitarian envisages must be a global one; the idea will be to use intra-personal comparisons of strength to make inter-personal comparisons, using a formal assumption of equal overall strength as a metric. In the simplest case, when A's preference *p* has the same strength relative to A's other preferences as B's preference *q* has to B's other preferences, we treat *p* and *q* as being equally strong. So a crucial point of preference utilitarianism is a commitment to give equal weight to each individual's preferences, considered globally. But how is this to be done?

Here it is possible to make a link between preference utilitarianism and Ronald Dworkin's (1981) Equality of Resources theory of distributive justice. Dworkin is a luck-egalitarian. So for him distributive justice involves eliminating (as far as is compatible with constraints on efficiency etc.) any inequality in resources that is the result of brute luck, which is luck not chosen by the individuals concerned. What are resources? They are the things that are to be distributed, paradigmatically money, which we can call 'external resources', plus 'internal' resources, which are features of individuals themselves that make a difference to how easy it is for them to acquire external resources and convert such resources into welfare. Thus talents, disabilities and unchosen, unwanted cheap/expensive tastes all count as internal resources. The suggestion is not, of course, that internal resources are themselves redistributed, but rather that external resources are distributed so as to compensate for inequalities in internal resources.

At this stage the account faces a compulsory question. What metric can be used to measure the relative value of different resources and to determine what counts as equality of resources? The answer, according to Dworkin, is that people are equal if no one envies anyone else's bundle of resources. This is because if anyone thought that someone else had a better bundle of resources, they would envy that person her bundle. Yet there is a problem with this envy test: some of the resources which people have in the real world are acquired through differential effort and option luck, for which individuals are responsible. Luck-egalitarians do not want to remove all sources of envy, just those sources of envy which are the results of luck rather than responsible choice. So it is inappropriate to apply the envy test to a real-world case.

The way to make an envy test applicable is to apply the envy test in an original position where life choices have not yet had their effects, and to have a veil of ignorance

that filters out brute luck. Dworkin constructs a thin veil by imagining a shipwreck scenario where the survivors wash up on an island and decide to divide the resources equally. It is not enough simply to give the individuals identical resource bundles, since the way in which the resources are divided into identical bundles may not be neutral between the preferences of different individuals:

> Suppose (to put the point in a dramatic way) the divider achieves his result by transforming all the available resources into a very large stock of plovers' eggs and pre-phylloxera claret (either by magic or by trade with a neighbouring island that enters the story only for that reason) and divides this glut into identical bundles of baskets and bottles. Many of the immigrants – let us say all but one – are delighted. But if that one hates plovers' eggs and pre-phylloxera claret he will feel that he has not been treated as an equal in the division of resources. [Dworkin 1981: 285]

Thus there is an auction procedure, in which each individual is allocated an equal number of (intrinsically worthless) clamshell tokens which they use to bid with. An auctioneer suggests lots and prices and sees if she can sell all the lots at the suggested price. Even if she can, any individual can change his mind if he is dissatisfied and bid differently or suggest different lots (Dworkin 1981: 287). We suppose that the auction eventually terminates when individuals all realise that they cannot do any better. The envy test is satisfied because each individual could have purchased the bundle purchased by any other individual and failed to do so, showing a lack of envy for anyone else's bundle.

A complication to the auction to filter out brute luck is that extra lots can be introduced which are insurance contracts. We now imagine a veil which renders individuals ignorant of their internal resources, and which allows them to take out insurance against having particular internal resources such as disabilities (and against

other contingencies that might arise). The individuals are allowed to know the statistical distribution of internal resources in the population etc., so as to calculate whether it makes sense to buy insurance against having particular disabilities etc. If we imagine that each person will make the same decisions concerning what insurance of this kind to buy, then we can use taxation to model the results of the hypothetical insurance market. Certain forms of tax represent insurance premiums; certain forms of benefits paid for by those taxes represent the insurance pay-outs. Efficiency savings from the pooling of insurance will justify systems such as nationalised healthcare. Thus the results of the hypothetical insurance market can be used to make real-world policy recommendations. It is not part of the theory that brute luck is entirely eliminated: there will be some disabilities that it makes no sense to insure against, since the cost of treatment is so high, and there will be very many cases where the optimal insurance contract involves a pay-out that fails to fully compensate for the severity of the disability. This suggests the criticism that Dworkin's theory fails an *ex post* envy test; but we have seen that such a test is difficult to formulate. It is worth noting at this point, as a hint of what is to come, that this rejection of full compensation for disabilities on efficiency grounds is reminiscent of utilitarianism.

What does the equal distribution of clamshells at the beginning of the hypothetical auction represent? My idea is that it can represent equal consideration of preferences, considered globally. Suppose that by making a bid for a lot in the hypothetical market, one expresses one's preference. It seems (at least initially) that the person who makes the highest bid for a lot is the person who wants that lot most. Each person has the responsibility of dividing up her clamshells between different lots she bids for; this is analogous to the idea that each person has a certain overall weight of preference, which is

distributed amongst the various lots. We might argue for the equivalence of Dworkin's market procedure and preference utilitarianism as follows: the proportion of her clamshells that an individual is willing to spend on a particular lot is an indication of the priority she attaches to obtaining that lot relative to other lots. Hence the fact that one individual is willing to spend more clamshells than others on a lot shows that the priority she attaches to getting that lot is higher than the priority any other agent attaches to it. Attachment to equal global strength of preferences guarantees that the agent who attaches the highest priority to a lot has the strongest preference. So giving the lot to the highest bidder will maximise preference satisfaction, as utilitarianism requires.

There is a wrinkle to this neat story, however, which prevents the above argument from going through. Suppose Alice has idiosyncratic preferences: the lot that she wants most is one that nobody else wants, so she can obtain it spending hardly any clamshells. Boris would have to spend more clamshells to get his highest priority lot. Suppose there is a lot that is Alice's third priority and Boris's second. Alice will probably get it since she has more clamshells to spend. So it appears that the auction will result in Alice getting that lot even though Boris wants it more, which looks like a failure to maximise preference satisfaction. This reveals that there is a mismatch between assigning equal weight to each individual's preferences considered globally (by allocating trading tokens equally) and maximising the satisfaction of preferences considered individually. This is because the strength of a preference depends on the greatest proportion of her total tokens that an agent **would** expend to obtain a lot *ceteris paribus*, rather than the amount that she **will** spend given the auction situation; the latter can be lower or higher, as the case of Alice and Bob reminds us. Given this conflict, we are forced to choose between the different

conceptions of utilitarianism. There is good reason for choosing the conception of assigning equal overall weight to individuals' preferences.

A toy example illustrates that considering preferences individually leads to unattractive results. Suppose that there are only three agents, Cass, Derek and Emily, and 30 objects that the agents can want to possess. How are these objects to be distributed? Suppose that Cass is equally keen on all of the objects. We can then use the assumption of equal overall preference strength to assign strengths to Cass's individual preferences. Say that each agent has 30 units of preference weight to be divided between her different preferences. If Cass's preferences are restricted to desires for individual objects, we can say that for each of the 30 objects, her respective preference has a strength of 1. Now suppose that half the objects are red and half blue. Derek is equally keen on the 15 red objects, and is uninterested in the blue objects; and vice versa for Emily. Then the strength of each of Derek's preferences for red objects is 2, and similarly for Emily with the blue objects. If we distribute the objects so that the person who wants an object most gets it, then we have to give all 15 red objects to Derek and all 15 blue objects to Emily, leaving Cass with nothing. If this is utilitarianism, it is far from egalitarian.

It is hard to reconcile the distribution proposed above with the utilitarian slogan that each is to count for one, and none for more than one. We have a strong intuition that the way to count Cass, Derek and Emily equally in this case is to give Derek 10 of the red objects, Emily 10 of the blue objects, and Cass 5 of each. And that would be the only equilibrium in the hypothetical market where the three have an equal number of tokens. The failure to count Cass can be located in a failure to respect her thwarted preferences. Each time we give Derek a red object, we impose a cost on Cass, but not on Emily; and when we give Emily a blue object, the cost again falls entirely on Cass.

The premise that individuals have preferences that are equally strong overall is misused by the maximising conception. It is treated as an empirical assumption when it is really an ethical commitment. To say that overall preferences are equally strong should just be another way of saying that they should count equally overall. But on the maximising conception it is interpreted in a way that prevents individuals' preferences counting equally. If it is not equivalent to the ethical commitment of equal treatment, where does it come from? It can hardly be a fact about preferences that has been discovered – we have said that it is needed in order to make sense of interpersonal comparisons of preference strength.

The key to the problem lies in Dworkin's insistence that preferences may be sensitive to market conditions, which reflect others' preferences:

> Under equality of resources, however, people decide what sorts of lives to pursue against a background of information about the actual cost their choices impose on other people and hence on the total stock of resources that may fairly be used by them. [Dworkin 1981: 288]

Bob imposes a higher cost on the community by getting his highest priority lot than Alice does by getting hers. Since such preferences are a matter of choice it is only fair that Alice should be rewarded for this, and Bob penalised, by her being at an advantage with respect to securing other lots. Yet the utilitarian has no right to such concepts of fairness, so apparently cannot constrain her theory in the way that Dworkin constrains his. What the utilitarian can say is that preferences are not immutable, and encouraging people to have cheap preferences is more efficient in satisfying preferences. Consider a scenario in which there are two types of resources, one which everyone wants, and one which no-one wants. In the market system there is an incentive for people to start wanting the unpopular resources, because then their preferences will be better satisfied. As the

situation changes to one in which people want different things, they will all be able to get more of what they want, and so their preferences will be better satisfied.

On the more sophisticated picture emerging from such considerations, the auction is not directly a method of calculating what utilitarianism requires, at least not on utilitarianism's standard interpretation. Let us keep fixed the claim that the equal division of clamshells involves an equal global weight of preference. Standardly utilitarianism treats preferences as fixed when deciding how to maximize their satisfaction. The market procedure does not do so: it takes it that maximizing preference satisfaction can involve some change to the preferences whose satisfaction is being maximized. Instead of looking at the amount of clamshells individuals pay in the auction as a measure of their preference strengths, we look at the auction procedure as a way of arriving at our conclusions without the need for a unit-comparable measure of welfare. The auction is justified on utilitarian grounds because it embodies the central utilitarian value of equal consideration for preferences in the only way that that makes sense – equal global weight to the preferences of each individual – and is justified by the utilitarian demand of efficiency in satisfying preferences, since it creates the right incentives.

The situation with insurance is actually rather simpler. Here individuals can be assumed to have the same preferences, so if people would choose to purchase a particular kind of insurance in the hypothetical market, we can simply read off that this insurance maximizes preference satisfaction. The crucial point here is that the veil mechanism is necessary in order to give us the information we need. Rather than just being a guarantee of impartiality, it allows us to compare preferences that would not otherwise be comparable. We want to know whether redistributing money to the disabled maximizes preference satisfaction; but we have no obvious way to find out without a metric for

comparing the preferences of the disabled to those of the non-disabled. The preferences that manifest themselves in insurance purchasing decisions from behind the veil, however, tell us just what we need to know. So here we have two different ways in which Dworkin's hypothetical market solves problems for utilitarians.

What this discussion reveals is that, under a certain interpretation, preference utilitarianism entails Equality of Resources. This is in itself a somewhat surprising result, and I think helps to show how utilitarianism in something like Hare's version can be substantively egalitarian. Of course I have not shown in detail how to adapt Hare's argument for utilitarianism to take into account what I have said about inter-personal comparison. But I do not think that there is a serious inconsistency here. One thing we need to say is that before we come to the application of (CR) we have to come to a view about preference strengths, and we do so my considering each individual preference in the context of all the preferences of the person whose preference it is, assuming equal overall preference strength as I have specified. It seems possible to do so whilst accommodating the insight that to consider the interests of different people equally we need keep track of each individual's overall level of preference satisfaction, rather than thinking about each preference in isolation. Hare's account accommodates this at least in spirit because it demands that we think about what it is like to be in a person's shoes in order to understand their preferences. I acknowledge, however, that further work is required to fully integrate the ideas discussed in this section with the Harean theory as a whole.

4.4 CONCLUDING REMARKS

In this chapter I have tried to argue that the distinctive form of utilitarianism defended by Hare is successful. I have shown how his argument for it can be tweaked to

avoid various objections, with the resulting theory one which improves on Hare's in aligning the preferences which count with interests (by excluding external preferences). I have explained why the transcendental argument for universalizability, given in Chapter 3, gives this argument much greater significance: rather than just telling us that utilitarianism follows if we have a particular set of moral concepts, it tells us that utilitarianism follows from a principle, universalizability, accepting which is a transcendental condition of agency and so inescapable in a much stronger sense than our concepts are. I have gone on to defend Hare's theory and his methodology against various objections. Here aspects of Hare's view explained in earlier chapters have been crucial. In particular the Kantian expressivist/constructivist roots of his theory explain how he can be a utilitarian without having an objectionably teleological theory, and the earlier arguments against realism rule out alternative methodologies that draw on alleged metaphysical facts about reasons. It was already argued in Chapter 3 that expressivism and Kantianism needed each other for support, because without Kantianism expressivism could not answer the regress, and without expressivism the Kantian argument could not go through (both Kant's derivation of the content of the CI, and the overall transcendental argument). The unusual Harean package of views – expressivism, Kantianism and utilitarianism – thus turns out to be mutually supporting and to show exactly how the concerns of meta-ethics, normative justification, and ethical theory can be understood in a unified way. In doing so I hope to gone some way towards answering, more convincingly than Hare did himself, Williams' (1988: 194-5) worry that 'the external view of what morality is, and the internal representation of it in moral practice' may not fit together.

That, at least, is what I hope to have shown, but of course I do not claim that this demonstration is by any means complete (though it is perhaps as complete as space has

allowed). I have just mentioned one gap to be filled in by further research: integrating my understanding of inter-personal comparison of utility with Hare's argument for utilitarianism. There are others: I have not had space to explain in detail where Kant's arguments for anti-utilitarian conclusions go wrong, or to do anything similar for contemporary Kantians. There are other rival views to Hare's that a more comprehensive study would compare, contrast and critique. Indeed even my study of meta-ethics in Chapter 1 did not attempt to survey or refute **every** rival to expressivism, except insofar as they all fit into the categories I delineated; a comprehensive defence of expressivism would spell out exactly how I deal with every serious alternative. I have chosen to rely mainly on putting forward a positive view, hoping that what I say in support of it will serve to explain the problems I see with the various alternatives, and thinking also that this will be of more philosophical interest than a more negative survey.

BIBLIOGRAPHY

Allison, Henry (2004). *Kant's Transcendental Idealism: An Interpretation and Defense*. New Haven and London: Yale University Press, Revised and Enlarged Edition.

Altham, J. E. J. (1986). 'The Legacy of Emotivism' in G. Macdonald & C. Wright eds*., Fact, Science & Morality*. Oxford: Blackwell.

Anscombe, G.E.M. (1957). *Intention*. Oxford: Blackwell.

Arkonovich, S. (2001). 'Defending desire: Scanlon's anti-Humeanism.' *Philosophy and Phenomenological Research* 63: 499-519.

Audi, Robert (1997). *Moral Knowledge and Ethical Character*. Oxford: Oxford University Press.

Aune, Bruce (1979). *Kant's Theory of Morals.* Princeton: Princeton University Press.

Ayer, A. J. (1946). *Language, Truth and Logic* (2nd edition). London: Victor Gollancz.

Barnes, W.H.F. (1933). 'A Suggestion about Value.' *Analysis* 1: 45-6.

Bentham, Jeremy (1907/1823). *An Introduction to the Principles of Morals and Legislation*, corrected edition. Oxford: Clarendon Press.

Blackburn, Simon (1981). 'Rule-Following and Moral Realism' in Stephen Holtzman and Christopher Leich eds., *Wittgenstein: To Follow a Rule*. London: Routledge and Kegan Paul, 163-87.

—— (1984). *Spreading the Word*. Oxford: Oxford University Press.

—— (1988). 'Attitudes and Contents.' *Ethics* 98.3: 501–17.

—— (1991a). 'Just Causes.' *Philosophical Studies* 61: 3-17.

—— (1991b). 'Reply to Sturgeon.' *Philosophical Studies* 61: 39-42.

—— (1993a). *Essays in Quasi-Realism.* Oxford: Oxford University Press.

—— (1993b). 'Realism, Quasi, or Queasy?' in John Haldane & Crispin Wright eds., *Reality, Representation & Projection*. Oxford: Oxford University Press.

—— (1998a). *Ruling Passions*. Oxford: Oxford University Press.

—— (1998b). 'Wittgenstein, Wright, Rorty and Minimalism.' *Mind* 107(425): 157-81.

—— (1999). 'Is Objective Moral Justification Possible on a Quasi-realist Foundation?' *Inquiry* 42.2: 213-27

—— (2009). 'Truth and A Priori Possibility: Egan's Charge Against Quasi-Realism.' *Australasian Journal of Philosophy* 87.2: 201-13.

Bradley, Richard & Christian List (2009). 'Desire-as-belief revisited.' *Analysis* 69.1: 31-7.

Brink, David (1989). *Moral Realism and the Foundations of Ethics*. Cambridge: Cambridge University Press.

Broad, C. D. (1933). 'Is Goodness the Name of a Non-Natural Quality?' *Proceedings of the Aristotelian Society* XXXIV: 249-68.

Bromwich, Danielle (forthcoming). 'Motivational Internalism and the Challenge of Amoralism.' *European Journal of Philosophy*.

Bykvist, Krister & Jonas Olson (2009). 'Expressivism and Moral Certitude.' *Philosophical Quarterly* 59(235): 202-15.

Carroll, Lewis (1995/1895). 'What the Tortoise Said to Achilles.' *Mind* 104(416): 691-3.

Carnap, Rudolf (1956). 'Empiricism, semantics, and ontology' in his *Meaning and Necessity: a study in semantics and modal logic*. Chicago: University of Chicago Press, 203-21.

Carson, Tom (1993). 'Hare on utilitarianism and intuitive morality.' *Erkenntnis* 39.3: 305-31.

Chalmers, David (2005). 'The foundations of two-dimensional semantics', in M. Garcia-Carpintero & J. Macia, eds., *Two-Dimensional Semantics: Foundations and Applications*. Oxford: Oxford University Press.

Clarke, Samuel (2003/1706) *A Discourse Concerning the Being and Attributes of God, the Obligations of Natural Religion and the Truth and Certainty of the Christian Revelation*. Whitefish MT: Kessinger Publishing.

Cohen, G. A. (2003) 'Facts and Principles.' *Philosophy and Public Affairs* 31: 211-45.

Copp, David (1995). *Morality, Normativity, and Society.* Oxford: Oxford University Press.

Cudworth, Ralph (1996/1731) *A Treatise Concerning Eternal and Immutable Morality*. Cambridge: Cambridge University Press.

Curry, Oliver (2005) *Morality as natural history*. PhD thesis, London: LSE Research Online. Available at: http://etheses.lse.ac.uk/2/

Dancy, Jonathan (1993). *Moral Reasons.* Oxford: Blackwell.

—— (1996). 'In Defense of Thick Concepts.' *Midwest Studies in Philosophy* 20.1: 263-79.

Daniels, Norman (1979). 'Wide reflective equilibrium and theory acceptance in ethics.' *The Journal of Philosophy* 76.5: 256-82.

Darwall, Stephen, Allan Gibbard & Peter Railton (1992). 'Toward Fin de siècle Ethics: Some Trends.' *Philosophical Review* 101.1: 115-89.

Dreier, Jamie (2006). ‘Negation for expressivists: A collection of problems with a suggestion for their solution.’ *Oxford Studies in Metaethics* 1: 217-33.

Dworkin, Ronald (1977). *Taking Rights Seriously*. Cambridge MA: Harvard University Press.

—— (1981). ‘What is equality? Part 2: Equality of resources.’ *Philosophy & Public Affairs* 10.4: 283-345.

—— (1996). ‘Objectivity and truth: You'd better believe it.’ *Philosophy & Public Affairs* 25.2: 87-139.

Egan, Andy (2007). ‘Quasi-realism and fundamental moral error.’ *Australasian Journal of Philosophy* 85.2: 205-19.

Eklund, Matti (2011). ‘What Are Thick Concepts?’ *Canadian Journal of Philosophy* 41.1: 25-49.

Elstein, Daniel Y. (2007). ‘Against Sonderholm: Still Committed to Expressivism.’ *Proceedings of the Aristotelian Society* CVII: 111-6.

Elstein, Daniel Y. & Thomas Hurka (2009). ‘From thick to thin: Two moral reduction plans.’ *Canadian Journal of Philosophy* 39.4: 515-35.

Enoch, David (2006). ‘Agency, Shmagency: Why Normativity Won’t Come from What Is Constitutive of Action.’ *Philosophical Review* 115.2: 169-98.

Frankena, W. K. (1939). ‘The Naturalistic Fallacy.’ *Mind* 48: 464-77.

Fullinwider, Robert K. (1977). ‘Fanaticism and Hare's Moral Theory.’ *Ethics* 87.2: 165-173.

Geach, Peter (1960). ‘Ascriptivism.’ *Philosophical Review* 69: 221–5.

—— (1965). ‘Assertion.’ *Philosophical Review* 74: 449–65.

Gibbard, Allan (1986). ‘An expressivistic theory of normative discourse.’ *Ethics* 96.3: 472-485.

—— (1988). ‘Hare’s Analysis of “Ought” and its Implications’ in Douglas Seanor & N. Fotion eds., *Hare and Critics*. Oxford: Clarendon Press.

—— (1990). *Wise Choices, Apt Feelings*. Oxford: Clarendon Press.

—— (1996). ‘Projection, Quasi-Realism, and Sophisticated Realism.’ *Mind* 105(418): 331-5.

—— (2003). *Thinking How to Live*. Cambridge MA: Harvard University Press.

—— (2006). ‘Normative Properties’ in Terry Horgan & Mark Timmons eds., *Metaethics after Moore*. Oxford: Oxford University Press.

Gregory, Alex (ms). ‘Might Desires Be Beliefs About Normative Reasons?’

Hajdin, Mane (1990). 'External and Now-For-Then Preferences in Hare's Theory.' *Dialogue* 29: 305-10.

Hale, Bob (1993). 'Can There Be a Logic of Attitudes?' in John Haldane & Crispin Wright eds., *Reality, Representation & Projection*. Oxford: Oxford University Press.

Hare, R.M. (1952). *The Language of Morals*. Oxford: Oxford University Press.

—— (1963). *Freedom and Reason.* Oxford: Clarendon Press.

—— (1970). 'Meaning and Speech Acts'. *Philosophical Review* 79.1: 3–24.

—— (1981). *Moral Thinking: Its Levels, Method, and Point*. Oxford: Clarendon Press.

—— (1988). 'Comments' in Douglas Seanor & N. Fotion eds., *Hare and Critics*. Oxford: Clarendon Press.

—— (1989a). *Essays in Ethical Theory*. Oxford: Clarendon Press.

—— (1989b). *Essays on Political Morality*. Oxford: Clarendon Press.

—— (1989c). 'Reply to Ingemar Persson.' *Theoria* 55: 171-7.

—— (1993). 'Could Kant Have been A Utilitarian?' *Utilitas* 5.1: 1-16.

—— (1999). *Objective Prescriptions and other essays*. Oxford: Clarendon Press.

Harsanyi, John (1975). 'Can the maximin principle serve as a basis for morality? A critique of John Rawls's theory.' *The American Political Science Review* 69.2: 594-606.

—— (1988). 'Problems with Act-Utilitarianism' in Douglas Seanor & N. Fotion eds., *Hare and Critics*. Oxford: Clarendon Press.

Heathwood, Chris (2012). 'Could Morality Have a Source?' *Journal of Ethics and Social Philosophy* 6.2: 1–19.

Hooker, Brad (2000). *Ideal Code, Real World*. Oxford: Clarendon Press.

Horgan, Terry & Mark Timmons (1991). 'New Wave Moral Realism Meets Moral Twin Earth.' *Journal of Philosophical Research* 16: 447-65.

—— (1992). 'Troubles for New Wave Moral Semantics: The Open Question Argument Revived.' *Philosophical Papers* 21: 153-175.

—— (2006). 'Cognitivist Expressivism' in their (eds.) *Metaethics after Moore*. Oxford: Oxford University Press, 255-98.

Horwich, Paul (2001). 'A defense of minimalism.' *Synthese* 126.1: 149-65.

Hume, David (1975/1777). *Enquiry concerning the Principles of Morals*, ed. L. A. Selby-Bigge, 3rd edition revised by P. H. Nidditch. Oxford: Oxford University Press.

—— (2000/1739). *A Treatise of Human Nature*, David Fate Norton and Mary J. Norton, eds. Oxford: Clarendon Press.

Jackson, Frank (1998). *From Metaphysics to Ethics*. Oxford: Clarendon Press.

—— (2001). 'Responses.' *Philosophy and Phenomenological Research* 62: 653–64.

Jackson, Frank & Philip Pettit (1996). 'Moral Functionalism, Supervenience and Reductionism.' *Philosophical Quarterly* 46: 82-6.

—— (1998). 'A Problem for Expressivism.' *Analysis* 58: 239-51.

—— (2003). 'Locke, Expressivism, Conditionals.' *Analysis* 63: 86-92.

Jackson, Frank, Graham Oppy & Michael Smith (1994). 'Minimalism and Truth Aptness.' *Mind* 103(411): 287-302.

James, William (1891). 'The Moral Philosopher and the Moral Life.' *International Journal of Ethics* 1.3: 330-54.

Jenkins, C. S. (2005). 'Realism and independence.' *American Philosophical Quarterly* 42.3: 199-209.

Kagan, Shelly (1989). *The Limits of Morality*. Oxford: Clarendon Press.

—— (1991). 'Replies to My Critics.' *Philosophy and Phenomenological Research* 51.4: 919-28.

—— (1998). *Normative Ethics* (revised edition). Boulder: Westview Press.

Kalderon, Mark (2005). *Moral Fictionalism*. Oxford: Oxford University Press.

Kant, Immanuel (1996/1783). 'Review of Schulz's Attempt at an introduction to a doctrine of morals for all human beings regardless of different religions' in Mary J. Gregor, ed. & trans., *Practical Philosophy, the Cambridge Edition of the Works of Immanuel Kant.* Cambridge: Cambridge University Press.

—— (1996/1785). *Groundwork of The metaphysics of morals* in Mary J. Gregor, ed. & trans., *Practical Philosophy, The Cambridge Edition of the Works of Immanuel Kant.* Cambridge: Cambridge University Press.

—— (2001/1786). 'What does it mean to orient oneself in thinking?' in Allen W. Wood, ed. & trans., George di Giovanni, trans., *Religion and Rational Theology, The Cambridge Edition of the Works of Immanuel Kant.* Cambridge: Cambridge University Press.

—— (2002/1790). *Critique of the Power of Judgment*, ed. & trans. Paul Guyer, trans. Eric Matthews, *The Cambridge Edition of the Works of Immanuel Kant*. Cambridge: Cambridge University Press.

Kelly, Thomas & Sarah McGrath (2010). 'Is reflective equilibrium enough?' *Philosophical Perspectives* 24.1: 325-59.

Kirchin, Simon (2010). 'The Shapelessness Hypothesis.' *Philosophers' Imprint* 10.4: 1-28.

Kitcher, Patricia (2004). 'Kant's Argument for the Categorical Imperative.' *Noûs* 38.4: 555-584.

Klein, Peter (1998) 'Foundationalism and the Infinite Regress of Reasons,' Philosophy and Phenomenological Research LVIII: 919-26.

Korsgaard, Christine M. (1993). 'The reasons we can share: an attack on the distinction between agent-relative and agent-neutral values.' *Social Philosophy and Policy* 10.1: 24-51.

—— (1996). *Sources of Normativity.* Cambridge: Cambridge University Press.

—— (2003). 'Realism and Constructivism in Twentieth-Century Moral Philosophy' in *Philosophy in America at the Turn of the Century*, supplementary volume (APA Centennial Supplement) of the *Journal of Philosophical Research*.

Kripke, Saul (1980). *Naming and Necessity.* Oxford: Blackwell.

Lang, Gerald (2001). 'The Rule-Following Considerations and Metaethics: Some False Moves.' *European Journal of Philosophy* 9: 190-209.

Le Guin, Ursula K. (1973). 'The Ones Who Walk Away from Omelas' in Robert Silverberg ed. *New Dimensions 3*. New York: Nelson Doubleday.

Lenman, James (2003a). 'Disciplined Syntacticism and Moral Expressivism.' *Philosophy and Phenomenological Research* 66.1: 32-57.

—— (2003b). 'Non-cognitivism and the dimensions of evaluative judgement' in Jamie Dreier & David Estlund eds., *Brown Electronic Article Review Service*. www.brown.edu/Departments/Philosophy/bears/homepage.html

Lewis, David (1986). *On the Plurality of Worlds*. Oxford: Basil Blackwell.

—— (1988). 'Desire as belief.' *Mind* 97(387): 323-32.

—— (1996). 'Desire as belief II.' *Mind* 105(418): 303-13.

Mackie, J. L. (1977). *Ethics: Inventing Right and Wrong.* Harmondsworth: Penguin.

Majors, Brad (2005). 'Moral Discourse and Descriptive Properties'. *Philosophical Quarterly* 55(220): 475-94.

McDowell, John (1979). 'Virtue and Reason.' *The Monist* 62: 331–50.

—— (1981). 'Non-cognitivism and Rule-Following' in Stephen Holtzman and Christopher Leich eds., *Wittgenstein: To Follow a Rule*. London: Routledge and Kegan Paul, 141-62.

—— (1987). 'Projection and Truth in Ethics.' Lindley Lecture, University of Kansas.

McGinn, Colin (1999). 'Reasons and Unreasons.' *The New Republic* May 24: 34-8.

McNamara, Paul (2010). 'Deontic Logic.' *The Stanford Encyclopedia of Philosophy (Fall 2010 Edition)*, Edward N. Zalta (ed.), http://plato.stanford.edu/archives/fall2010/entries/logic-deontic/

Mele, Alfred R. (1996). 'Internalist moral cognitivism and listlessness.' *Ethics* 106: 727-53.

Miller, Alexander (2003). *An introduction to contemporary Metaethics.* Cambridge: Polity Press.

Moore, G.E. (1903). *Principia Ethica*. Cambridge: Cambridge University Press.

Mulgan, Tim (1997). 'Two conceptions of benevolence.' *Philosophy & Public Affairs* 26.1: 62-79.

Murphy, Liam (1993). 'The demands of beneficence.' *Philosophy & Public Affairs* 22.4: 267-92.

Nagel, Thomas (1988). 'The Foundations of Impartiality' in Douglas Seanor & N. Fotion eds., *Hare and Critics*. Oxford: Clarendon Press.

Narveson, Jan (1978). 'Liberalism, Utilitarianism, and Fanaticism: R. M. Hare Defended.' *Ethics* 88.3: 250-9.

Nozick, Robert (1974). *Anarchy, State, and Utopia*. Oxford: Blackwell.

Oddie, Graham (2005). *Value, Reality, and Desire*. Oxford: Clarendon Press.

O'Neill, Onora (1989). *Constructions of Reason*. Cambrigde: Cambridge University Press.

Persson, Ingmar (1983). 'Hare on Universal Prescriptivism and Utilitarianism.' *Analysis* 43: 43-9.

—— (1989). 'Universalizability and the Summing of Desires.' *Theoria* 55: 159-70.

Powell, Brian K. (2006). 'Kant and Kantians on "the Normative Question".' *Ethical Theory and Moral Practice* 9.5: 535-44.

Price, Richard (1948/1787). *Review of the Principal Questions in Morals*, 3rd edition. Oxford: Oxford University Press.

Prior, Arthur (1960). 'The runabout inference ticket.' *Analysis* 21: 129-31.

Putnam, Hilary (1975). 'The meaning of "meaning"' in K. Gunderson, ed., *Language, Mind and Knowledge*. Minneapolis: University of Minnesota Press.

—— (2002) 'The Entanglement of Fact and Value' in his *The Collapse of the Fact/Value Dichotomy and other essays.* Cambridge MA: Harvard University Press.

Rabinowicz, Wlodek (1989). 'Hare on Prudence.' *Theoria* 55: 145-51.

—— (2009). 'Preference Utilitarianism by Way of Preference Change?' in *Preference Change*. Dordrecht: Springer, 185-206.

Rabinowicz, Wlodek & Bertil Strömberg (1996). 'What if I were in his shoes? On Hare's argument for preference utilitarianism.' *Theoria* 62: 95-123.

Radzik, Linda (1999). 'A Normative Regress Problem.' *American Philosophical Quarterly* 36.1: 35-47.

—— (2000). 'Justification and the Authority of Norms.' *Journal of Value Inquiry* 34.4: 451-61.

Rawls, John (1999). *A Theory of Justice* (2nd edition). Cambridge MA: Harvard University Press.

—— (1980). 'Kantian Constructivism in Moral Theory.' *The Journal of Philosophy* 77.9: 515-72.

Ridge, Michael (2003). 'Certitude, robustness, and importance for non-cognitivists' in Jamie Dreier & David Estlund eds., *Brown Electronic Article Review Service*. www.brown.edu/Departments/Philosophy/bears/homepage.html

—— (2006). 'Sincerity and expressivism.' *Philosophical Studies* 131.2: 487-510.

Roberts, Debbie (2011). 'Shapelessness and the Thick.' *Ethics* 121.3: 489-520.

Scanlon, T.M. (1982). 'Contractualism and Utilitarianism' in Amartya Sen & Bernard Williams eds., *Utilitarianism and Beyond*. Cambridge: Cambridge University Press.

—— (1988). 'Levels of Moral Thinking' in Douglas Seanor & N. Fotion eds., *Hare and Critics*. Oxford: Clarendon Press.

—— (1998). *What We Owe to Each Other*. Cambridge MA: Harvard University Press.

—— (2009). *Being Realistic about Reasons*. The Locke Lectures.

Schneewind, Jerome B. (1997). *The Invention of Autonomy.* Cambridge: Cambridge University Press.

Schopenhauer, Arthur (1965/1841). *On the Basis of Morality*, trans. E. F. J. Payne. Indianapolis: Bobbs-Merrill.

Schroeder, Mark (2005). 'Cudworth and Normative Explanations.' *Journal of Ethics & Social Philosophy* 1.3.

—— (2008a). 'How Expressivists Can and Should Solve Their Problem with Negation.' *Noûs* 42(4): 573-99.

—— (2008b). *Being for: Evaluating the semantic program of expressivism.* New York: Oxford University Press.

Schueler, G.F. (1984). 'Some Reasoning about Preferences.' *Ethics* 95: 78-80.

Searle, John (1962). 'Meaning and Speech Acts.' *Philosophical Review* 71: 423–32.

Sepielli, Andrew (2012). 'Normative uncertainty for non-cognitivists.' *Philosophical Studies* 160: 191–207.

Shafer-Landau, Russ (2000). 'A defense of motivational externalism.' *Philosophical Studies* 97: 267–91.

Singer, Peter (1972). 'Famine, affluence, and morality.' *Philosophy & Public Affairs* 1.3: 229-43.

—— (1988). 'Reasoning towards Utilitarianism' in Douglas Seanor & N. Fotion eds., *Hare and Critics*. Oxford: Clarendon Press.

Sinnott-Armstrong, Walter (2000). 'Expressivism and Embedding.' *Philosophy and Phenomenological Research* 61.3: 677-93.

Skidmore, James (2002). 'Skepticism about Practical Reason: Transcendental Arguments and Their Limits.' *Philosophical Studies* 109: 121-41.

Smart, J. J. C. (1956). 'Extreme and restricted utilitarianism.' *The Philosophical Quarterly* 6(25): 344-54.

Smith, Michael (1986). 'Should We Believe in Emotivism?' in Graham Macdonald & Crispin Wright eds., *Fact, Science & Morality Essays on A.J. Ayer's 'Language, Truth and Logic'*. Oxford: Blackwell.

—— (1987). 'The Humean theory of motivation.' *Mind* 96: 36-61.

—— (1994). *The Moral Problem*. Oxford: Blackwell.

—— (2002). 'Evaluation, uncertainty, and motivation.' *Ethical Theory and Moral Practice* V: 305-20.

Snare, Frank (1975). 'The Open Question as a Linguistic Test.' *Ratio* 17: 122-9.

Soames, Scott (2003). *Philosophical Analysis in the Twentieth Century.* Princeton: Princeton University Press.

Sonderholm, Jorn (2005). 'Why an Expressivist should not Commit to Commitment-Semantics.' *Proceedings of the Aristotelian Society* CV: 403-9.

Stocker, Michael (1979). 'Desiring the Bad: An Essay in Moral Psychology.' *The Journal of Philosophy* 76: 738-53.

Stevenson, Charles L. (1937). 'The Emotive Meaning of Ethical Terms.' *Mind* 46: 14-31.

Strawson, P.F. (1962). 'Freedom and Resentment.' *Proceedings of the British Academy* 48: 1-25.

Streumer, Bart (2008). 'Are There Irreducibly Normative Properties?' *Australasian Journal of Philosophy* 86.4: 537-61.

Stroud, Barry (1968). 'Transcendental Arguments.' *Journal of Philosophy* 65: 241-56.

Sturgeon, Nicholas (1986). 'What difference does it make whether moral realism is true?' *The Southern Journal of Philosophy* 24.S1: 115-41.

—— (1991). 'Contents and Causes: A Reply to Blackburn.' *Philosophical Studies* 61: 19-37.

Suikkanen, Jussi (2009). 'The subjectivist consequences of expressivism.' *Pacific Philosophical Quarterly* 90.3: 364-87.

Svavarsdóttir, Sigrún (1999). 'Moral Cognitivism and Motivation.' *Philosophical Review* 108: 161–219.

Taurek, John M. (1977). 'Should the numbers count?' *Philosophy & Public Affairs* 6.4: 293-316.

Unwin, Nicholas (1999). 'Quasi-Realism, Negation, and the Frege-Geach Problem.' *Philosophical Quarterly* 49(196): 337-352.

Van Roojen, Mark (1996). 'Expressivism and Irrationality.' *Philosophical Review* 105.3: 311-335.

Väyrynen, Pekka (2012). 'Thick Concepts: Where's Evaluation?' *Oxford Studies in Metaethics* 7: 235-70.

—— (2013). 'Grounding and Normative Explanation.' *Proceedings of the Aristotelian Society Supplementary Volume* LXXXVII: 155-78.

Vendler, Zeno (1988). 'Changing Places' in Douglas Seanor & N. Fotion eds., *Hare and Critics*. Oxford: Clarendon Press.

Weintraub, Ruth (2011). 'Logic For Expressivists.' *Australasian Journal of Philosophy* 89.4: 601-16.

Whitehead, Alfred North (1929). *Process and Reality.* New York: Macmillan.

Williams, Bernard (1985). *Ethics and the Limits of Philosophy*. London: Fontana Press.

—— (1988). 'The Structure of Hare's Theory' in Douglas Seanor & N. Fotion eds., *Hare and Critics*. Oxford: Clarendon Press.

Winch, Peter (1965). 'The universalizability of moral judgements.' *The Monist* 49.2: 196-214.

Wittgenstein, Ludwig (1922). *Tractatus Logico-Philosophicus*, trans. C. K. Ogden. London: Routledge and Kegan Paul.

Wood, Allen W. (1999). *Kant's Ethical Thought.* Cambridge: Cambridge University Press.

Wright, Crispin (1992). *Truth and Objectivity.* Cambridge MA: Harvard University Press.