# NEO-KANTIAN CONSTRUCTIVISM AND METAETHICS

by

KIRK SURGENER

A thesis submitted to the University Of Birmingham

for the degree of

DOCTOR OF PHILOSOPHY

Department of Philosophy

College of Arts and Laws

The University of Birmingham

September 2011

I’d like to thank Alex Miller for giving me all the arguments in this; Joe Morrison and Darragh Byrne for printing it; Iain Law and Hallvard Lillehammer for examining it; and Naomi Maria Callas Thompson and Natalie Ashton for proof-reading it. I’d also like to thank all the people who have argued with me about the contents over the years, including but not limited to: Jussi Suikkanen, Ben Matheson, Ben (II) Bessey, Roxanne Harmony Green, Paul ‘The Broadbean’ Broadbent, Khai Wager, Callum Hood, Anna Brown, Damian ‘John’ Lewis, Helen Louise Crane, David Papineau, Philip Goff, ‘Nikk’ Effingham, Helen Beebee, Joss Walker, Max Kölbel, Sarah-Louise Johnson, Gis Infield-Solar, Mihaela Popa, Sarah Gancarczyk, John Gingell, Kate Major, Pegah Lashgarlou, Sorana ‘Piggy’ Vieru, Melanie Parker and Emma Cecilia Bullock.

# TABLE OF CONTENTS

# INTRODUCTION

Why should I be moral? This question gets to the heart of the *normative problem*, that is, the problem of grounding the normative force of moral obligations. People who take the normative problem seriously think that even once we have determined which actions are right or which objects are good there is still a question to be raised – why should we perform right actions? Why should we pursue good objects? In some cases what morality asks of us can be *hard* and the normative problem there seems particularly pressing.

Christine Korsgaard has used the normative problem to launch arguments against two of the most popular metaethical accounts – moral realism and expressivism. She argues that reflection on the normative problem forces us to reject moral realism and expressivism, and adopt a position which 'transcends' or 'goes beyond' metaethics as it is traditionally conceived. We can call this positive view neo-Kantian constructivism.

This thesis is a sustained examination of both of these parts of Korsgaard's work – the negative attacks on moral realism and expressivism; and her own neo-Kantian constructivism. I have two ambitions for it. First, I want to get clear formulations of and evaluate Korsgaard's arguments against her metaethical competitors and for her own position. Second, by engaging in the first task I hope to offer some results that will be independently interesting to people who are interested in metaethics.

Views similar to the ones Korsgaard defends have been advocated by other philosophers. In this thesis I have concentrated almost exclusively on Korsgaard's own view. This is in part because extracting a clear formulation of Korsgaard's arguments from her work is not always easy, and requires some amount of space. Also, I am more interested in how a view like neo-Kantian constructivism contrasts with completely different views in metaethics than in the details of different types of neo-Kantian constructivism (This is, of course, to some extent a false dichotomy. One way to explore how neo-Kantian constructivism hooks up with the rest of metaethics is to explore differences within the constructivist camp. I can only say this: it is only to some extent a false dichotomy – I have felt that the best way to pursue the issues I'm interested in is to concentrate on a single view. I hope that the things that I say about metaethics are sufficiently independently interesting to compensate for this somewhat narrow focus).

My conclusion will be that Korsgaard's position, although worth engaging with, ultimately fails. Her arguments against moral realism can be resisted if we formulate the right type of moral realism. However, her complaints about expressivism, when charitably interpreted, do cause problems for the expressivist. I give a new way of interpreting her own metaethical position, but argue that it ultimately fails in its ambitions.

I begin in chapter one with an examination of attempts to dismiss the normative problem and the questions stemming from it as confused and thus safely ignored. I argue there that such attempts rely upon an overly austere conception of the tasks of metaethics, and a questionable thesis about the relationship of metaethics to normative ethics. I also begin to outline Korsgaard's argument against realism.

In chapter 2 I argue that we can get a clearer grip on Korsgaard's argument against realism if we construe it as a problem to do with the alleged motivational import of moral judgements. Variations of the claim that moral judgements are inherently motivating are often made (a claim we can call 'internalism'), and this claim causes problems for realism (against suitable background assumptions). Seeing Korsgaard's argument this way allows us to explain the affinities she claims her argument has with G.E. Moore's open question argument and J.L. Mackie's argument from queerness. I argue that the lesson we should draw is that internalism is a troubling claim for moral realism, and that we should investigate whether there are compelling reasons to accept it. I first argue that Michael Smith's two-pronged manoeuvre in favour of internalism fails, before going on to consider Mark van Roojen's more recent case for internalism – again arguing that it fails. What moral realists need to do, I claim, is establish a viable form of externalist realism, and hence they will be able to dodge Korsgaard's argument when it is construed in this manner.

I then go on (chapter 3) to offer a second interpretation of Korsgaard's argument where she ends up offering what we can call, following Mark Schroeder, a generalised anti-voluntarist argument. The upshot of this argument is that moral realism, to avoid the argument, should be *reductionist*. I then go on to consider two versions of moral realism: one externalist and non-reductionist (Cornell realism); and the other externalist and reductionist (Stephen Finlay's analytic reductivism). If either of these views is viable then moral realism is able to dodge one or both of the construals of Korsgaard's argument. I argue that Finlay's position does well along a number of dimensions, but that it invokes a methodology that is not licensed by the account of analyticity he adverts to. Cornell realism can resist two of the major lines of attack typically launched against its semantic programme and its ontological claims. Both views, I think, offer us ways to circumvent Korsgaard's arguments.

Chapter 4 looks at Korsgaard's engagement with expressivism. I flesh out her complaints against expressivism in a way that has them as getting at something like the Frege-Geach problem – the problem of explaining how the semantic properties of complex expressions are built up out of the semantic features of their simpler constituents. I survey the most popular attempts to deal with this problem and argue that they are all unpromising. This motivates a study of a new type of position – hybrid expressivism – which combines elements of cognitivist and non-cognitivist semantics in order to deal with the problems that attend each type of semantic theory individually. I argue that hybrid expressivism either fails, or is best construed as a more sophisticated version of moral realism.

Finally, I turn to Korsgaard's positive proposal, neo-Kantian constructivism. I propose a novel way of interpreting this position, where we construe constructivism as a form of cognitivist anti-realism along lines inspired by Crispin Wright's work on judgement-dependent qualities. Doing things this way allows us to both give Korsgaard most of what she wants from a metaethical theory as well as generating a clear proposal to evaluate. When things are put this way the viability of constructivism depends upon being able to give the right sort of argument for the categorical imperative. When we try to do this, we see that neo-Kantian constructivism ultimately fails.

# CHAPTER ONE: THE NORMATIVE QUESTION AND MORAL REALISM

When investigating morality, it seems as if we are not just looking for a list of things that we should or should not do. In addition, we expect to find out *why* we are beholden to the dictates of morality – to find out how morality gets its normative force. This question – why should I be moral? – Christine Korsgaard calls 'the normative question' and it provides the basis for her moral thinking. She uses this question to attempt three manoeuvres: first, to show that traditional metaethical theories like moral realism and expressivism are inadequate – they lack the resources to answer the normative question satisfactorily; second, that this failure is in part a result of the inadequacies of the typical distinctions (e.g. between cognitivism and non-cognitivism) made in contemporary metaethics; and third, that neo-Kantian constructivism (a position which 'goes beyond' traditional metaethics) *does* have the resources to provide a satisfactory answer to the normative question, and so we should accept it.

The normative question, then, is at the heart of Korsgaard's moral philosophy. Here I will try to get clear on precisely what the normative question is asking for (§1.1), before laying some additional groundwork (§1.2) that will be required for showing how Korsgaard tries to use the normative question to undermine moral realism (§1.3).

## 1.1 The Normative Question

Korsgaard contends that there are two major tasks involved in systematic moral theorising. First we want to come to understand three features of moral concepts: 1. The meaning of moral concepts, or to use Korsgaard's metaphor "what they contain" (1996, 10) – what does it mean to say that something is good or bad, or to call a person vicious or virtuous etc. 2. We want to know to which objects these concepts are appropriately applied – just which things are good or bad, which people are virtuous or vicious. 3. Where do these moral concepts come from? That is, how did we come to possess them?

In addition to these tasks (providing what Korsgaard calls a 'theory of moral concepts') the nature of moral concepts also means that there is another account we need to provide. We don't just use moral concepts to describe the world, but also to make demands upon one another:

> [E]thical standards are *normative*. They do not merely *describe* a way in which we in fact regulate our conduct. They make claims on us; they command, oblige, recommend, or guide. When I say that an action is right I am saying that you ought to *do* it; when I say that something is good I am recommending it as worthy of your choice... it is the force of these normative claims – the right of these concepts to give laws to us – that we want to understand. (Korsgaard, 1996, 9)

Morality does not just make these claims upon us, these claims sometimes have a practical effect – it seems as if we sometimes alter our behaviour on the basis of what morality

demands, sometimes to a radical degree: fiction and everyday life offer us examples of people who purportedly sacrifice their own lives for the sake of doing the right thing. Furthermore, either failing or succeeding to meet these demands can have psychological effects upon us – failing to fulfil what we take to be our duty can cause distress, for example.

Korsgaard takes these observations to demonstrate the need for two criteria of adequacy for an account of the nature of morality. First, it must be explanatorily adequate – it must be able to explain the seeming importance of morality in our lives, and the practical and psychological effects it can have on us. Second, it must be justificatorily adequate – it must be able to explain whether we are justified in giving morality such importance; whether we should make the judgements that we do; whether we should allow those judgements to have the practical and psychological effects that they do; and why morality has the *practical significance* that it seems to possess. Asking whether a theory of morality fulfils this second criterion is what asking the normative question involves.

We can see how these two criteria diverge in a case where the explanation of our moral practices works to debunk those practices in some way – an explanation where once we see how our moral practices are explained, we no longer feel they have any justification. For example we might discover that morality has some kind of genetic basis, what Korsgaard calls 'the evolutionary theory'. According to this theory: "right actions are those which promote the preservation of the species, and wrong actions are those which are detrimental to this goal." (14). Further, suppose that:

> The evolutionary theorist can prove, with empirical evidence, that because this is so, human beings have evolved deep and powerful instincts in favour of doing what is right and avoiding wrong. (14.)

If this theory were true, then it could give an account of our moral practices which was explanatorily adequate – no wonder we place such huge importance on morality, we have "deep and powerful instincts" that incline us to act morally. However, would such a theory be good enough to *justify* those moral practices?

> Suppose morality demands that you yourself make a serious sacrifice like giving up your life, or hurting someone that you love. Is it really enough for you to think that this action promotes the preservation of the species? You might find yourself thinking thoughts like these: why after all should the preservation of the species count so much more than the happiness of the individuals in it? Why should it matter so much more than my happiness and the happiness of those I care most about? Maybe it's not worth it. (14-15).

So, once we see that our moral practices are explained by a theory like this, we start to doubt whether they really are justified. Such a theory exhibits, as Korsgaard puts it, 'normative failure'. It is the existence of the second criterion (justificatory adequacy) that allows the possibility of moral scepticism. It should be obvious, even to a sceptic, that people apply moral concepts quite regularly – a moral sceptic does not deny that people utter sentences like 'Tony Blair was a morally base individual' or 'There is nothing wrong with telling a white lie.' Instead, the sceptic can simply deny that the effects these judgements have are justified – we have no good reason to take my judgement of Tony Blair's character to influence my behaviour. This is because we need an account of why I should pay attention to the demands of morality. If we can formulate an adequate answer to

the normative question then we can have something to say in response to the moral sceptic who claims that we have no reason to pay attention to morality.

Another way to get clear on what the normative question is asking us for is to distinguish it from other, closely related, questions with which it might be confused. Korsgaard does this by looking at H.A. Prichard's argument that the question 'why should I be moral?' is confused. Briefly put, Prichard claims that there are two possible types of answers to the question: 1. We give an answer involving moral notions (e.g. 'because it is your duty'), in which case we have argued in a circle; 2. We could give an answer from outside of morality (e.g. 'because doing so would make you happy') but now our answer looks irrelevant – we feel as if the reason why we should be moral can't be because it would be good for us[1]. So, the question 'why should I be moral?' only admits of answers that are either irrelevant or circular. Prichard takes this to indicate that although the normative question looks coherent, it is not (see Prichard 1912 and Korsgaard 1996, 32).

Korsgaard argues that one way to see how this is confused, and show that the normative question is live, is to look at what Prichard advises us to do in the case where someone asks 'why should I be moral?' According to Prichard, a question like this is a disguised way of

[1] This answer is both extensionally inadequate – morality seems to diverge from self-interest, at least in some cases – and inadequate in another sense: if we say that we should be moral in order to further our self-interest, we haven't yet explained why morality is binding on us, for now we need an explanation of why we should feel bound by our self-interest. The fact that people often do act in their self-interest, and that it seems obvious that one should, we can imagine Prichard saying, does not actually explain *why* self-interest is normatively binding.

asking whether a particular criterion with moral significance applies to a particular object.[2] So, suppose that someone argues that the correct moral theory is some form of consequentialism with the relevant consequences restricted to psychological states of pleasure or pain. This theory demands that we maximise the amount of pleasure in the world and minimise the amount of pain. According to this theory, good states of affairs are those with the greatest balance of pleasure over pain[3], and right actions are those that promote such state of affairs. Now, suppose somebody asks of some action that they agree to be right whether they should perform it. Prichard claims that this person is really asking whether the action promotes the greatest balance of pleasure over pain. In order to answer their question we don't have to do anything mysterious; instead we simply remind them that the morally significant criterion applies – we say to them 'But look, the action promotes the greatest balance of pleasure over pain, it must be right!' This answer is appropriate, because anyone asking such a question is confused – they are not really asking whether they should be moral, instead they are wondering whether the criterion they use to distinguish the moral really applies in this case. Korsgaard's normative question then, is, for Prichard, a generalised confusion – asking 'why should I be moral?' is to ask whether a particular moral concept ever applies to any object, it is not to ask for an explanation of the normative force of morality (see Korsgaard 1996, 38-9).

However, Korsgaard argues that the confusion is all Prichard's. The answer Prichard offers to our putative sceptic "addresses someone who has fallen into doubt about whether an

[2] It seems here as if what Prichard is doing is offering a charitable re-interpretation of our question – the question, as literally stated, is confused. However, there is a nearby question that *is* significant that we can interpret people who ask the confused question as trying to get at.

[3] Of course Prichard himself, being a non-naturalist, would not identify goodness with the distribution of pain and pleasure. Instead, he would claim (if he accepted this particular moral theory) that goodness was just necessarily connected with the distribution of pain and pleasure.

action is really required by morality, not someone who has fallen into doubt about whether moral requirements are really normative" (38). Korsgaard's diagnosis of this misfire is that Prichard takes words like 'right' and 'obligatory' to be essentially normative, by definition. They are, as Korsgaard labels them, normatively loaded. If we accept this view, then the two questions: 'Is this action really obligatory' and 'Is this obligation really normative?' collapse into one another – in order for an action to obligatory it must have normative force. We would only ever need to answer the first question. However, this invites confusion for the question 'Is this action really obligatory?' admits of a reading under which it is simply a question about the correct application of some moral notion – about whether the action *does* promote the greatest amount of pleasure over pain. Because of this potential reading, and the collapse of the second question in to the first, we imagine that once we have answered it we have completed our work. Once we know that an action promotes the greatest balance of pleasure over pain, there is no more that needs to be said about it. Korsgaard contends that this is a mistake. There is another reading of 'Is this action really obligatory' available – one where we are asking not if it promotes the greatest balance of pleasure over pain (which would make it right) but whether we should be bothered about performing right actions. So, Korsgaard argues:

> Prichard's way of approaching the matter therefore leads us to confuse the question of correct application with the question of normativity. And this actually happened to Prichard himself. For it led him to think that once we have settled the question of correct application, there can be nothing more to say about the normative question. (39)

To put it another way, Prichard's collapsing of the distinction between the question of the correct application of a moral criterion and the normative question depends upon the assumption that morality really does have normative force. But this is precisely the

assumption that the normative question is asking for an explanation, or defence, of. It is illegitimate to use this assumption as a way of showing the normative question to be incoherent, as Korsgaard's reading of Prichard has him doing.

Of course, it could turn out that something like Prichard's conclusion is right. That there is something wrong with the normative question. However, I suspect this is the sort of conclusion we could only reach after seeing where attempts to answer it get us. Is it really true, for example, that the question only admits of answers that are either circular or irrelevant? It seems rather hasty to accept *this* on the basis of what we have seen from Prichard so far. For example, Korsgaard thinks that she *does* have a good answer to the normative question, and it would be better to examine it in detail rather than dismissing the claim in advance. In addition, this Prichardian move may be unnecessary, depending upon our purposes. Korsgaard wants to use the normative question not just as a way of promoting neo-Kantian constructivism (her own view) but also to attack moral realism, non-cognitivism and the metaethical distinctions upon which these views rest. If we examine these criticisms and find that, in fact, moral realism (for example) *does* have the resources to provide a satisfactory answer to the normative question then we might lose interest in attempting to dismiss the question from afar. Even if we are troubled that there is some kind of incoherence concealed within the normative question, it would still be an interesting finding if the conclusions that Korsgaard draws from the question do not follow. Then we could remain agnostic about the status of the normative question whilst resisting Korsgaard's manoeuvres for other reasons.[4]

[4] There is another line according to which Korsgaard's project is entirely misguided, found in the work of Nadeem Hussain and Nishi Shah (2005, 2006a, 2006b). However, they argue not against the legitimacy of the normative question, but against its use by Korsgaard against certain metaethical views. The

We have seen then that the normative question seems to provide a criterion of adequacy on any systematic moral theorising. The example of the evolutionary theory of morality demonstrates that we need not only an explanation of how morality has the effect that it does (in broadly psychological terms) but also an account of normative force. Also, we have seen that the normative question needs to be distinguished from other, closely related questions, such as the correct application of moral concepts. Once we do this, we see that the most obvious kind of attack on the coherence of the normative question is potentially misguided. This gives us, I claim, good reason to examine the *uses* of the normative question before we reach any grand conclusion about its status.

Below we begin this process by first setting up a distinction we need in hand (§1.2) in order to see how Korsgaard launches an attack on moral realism (§1.3).

## 1.2 Substantive vs. Procedural Realism

Moral realism, as I shall use the term in this thesis, is a view identified by three claims:

(1) Moral judgements purport to be true or false.

---

normative question is coherent, but not pitched at the right theoretical level for Korsgaard to derive her conclusions. I tackle this argument for the irrelevancy of Korsgaard below.

(2) Sometimes these judgements are true (in other words, their truth-conditions are sometimes fulfilled and the judgements in question accurately represent moral facts).

(3) These facts hold independently of our best judgements concerning them.

There are a number of different types of moral realism, and one could design a number of taxonomies to divide them. A useful one is given by Alex Miller (2003), which I will be following here. First we can ask whether the facts that our moral judgements purport to represent are natural facts. If not, then our view will be non-naturalist, a position inhabited by G.E. Moore (1903) who claimed that moral properties are *sui generis*, simple and indefinable, but also includes the work of John McDowell (1998) who tries to shed non-naturalism of its objectionable epistemological baggage. If instead we decide that moral facts are natural facts, then we face another choice – between positions which claim that moral facts reduce to other natural facts, and positions that view moral facts as irreducible natural facts. In the first camp we have the revising definitions strategy of Peter Railton (1989); and the non-revisionist strategies of Frank Jackson and Philip Petit's analytic functionalism (1995), Stephen Finlay's analytic naturalism (forthcoming) and others. The main proponents of the second view (that moral facts are irreducible to other natural facts) have been the so-called 'Cornell realists' – philosophers like Nicholas Sturgeon (1985, 1986) who argue that moral facts earn their keep by featuring in some of our best explanations of natural phenomena.

In addition, moral realists are also usually cognitivists – moral judgements express belief-like mental states which when true conceptually guarantee that the judgement is true[5]. It is easy to see how cognitivism sits well with the three claims above – if moral judgements have descriptive content, and this content concerns the holding of mind-independent moral facts then it seems natural to hold that moral judgements are belief-like mental states rather than desire-like non-cognitive states (which don't seem like the kinds of things capable of having descriptive content). The close link between realism and cognitivism will become relevant when we consider Korsgaard's attack on the distinction between cognitivism and non-cognitivism.

Korsgaard offers another way of distinguishing between moral realisms – between procedural and substantive realism. Both views agree that there are answers to moral questions, and that there are right and wrong ways of going about answering them – some procedures are better for arriving at answers to moral questions. The substantive realist adds the claim that there are correct procedures for answering moral questions because there are independently existing moral facts that those procedures ask about.[6] So both views think that there are good and bad procedures for going about answering moral questions, but they disagree about what underpins those procedures. The substantive realist thinks the best procedure is best because it tracks the independently existing moral facts that we are aiming for in correct moral judgements. The *merely* procedural realist claims that there is no need for these independent moral facts:

[5] Note that there is nothing here saying that moral judgements *only* express beliefs. This is to allow space for so-called 'hybrid' views where moral judgements express both beliefs and desires. The clause about a conceptual link between the truth of the belief expressed by the judgement and the truth of the judgement will become relevant when discussing those views in chapter 4.

[6] Korsgaard does not have the independence clause in her presentation of the distinction, but it is clear that she does embrace this condition see 1996, 34-37

Procedural Realism (**PR**): there are right and wrong ways of answering moral questions (good and bad procedures for answering them).

Substantive Realism (**SR**): there are right and wrong ways of answering moral questions (good and bad procedures for answering them) because there are independently existing moral facts that those procedures aim to track.

Mere Procedural Realism (**MPR**): there are right and wrong ways of answering moral questions (good and bad procedures for answering them), and this does not depend upon the procedures tracking independent moral facts. [7]

So procedural realism claims that there are good and bad procedures for answering moral questions. Substantive realism then offers an explanation of why those procedures are good or bad (because they track or fail to track the independently existing moral facts). Mere procedural realism denies the need for these moral facts. To put it another way, **PR** claims that there are answers to moral questions *because* there are good procedures for arriving at answers to them. **SR** adds the claim that this "because" is underwritten by another, more fundamental "because" – the procedures are good *because* they track the independently existing moral facts accurately. **MPR** denies the need for this second "because" underpinning the first. As I have laid it out here, **SR** is a subset of **PR**. **MPR** is what is left of the **PR** set after you take out the **SR** views. Given this, 'moral realism' as I mean it will denote substantive realism.

[7] Korsgaard does not distinguish between PR and MPR. However, this appears to be a harmless clarification – Korsgaard wants to defend mere procedural realism (the view that denies the substantive realist's additional claim) under the banner of 'procedural realism' but her description of procedural realism is in fact compatible with substantive realism.

Procedural realism, despite its expansiveness, is not a trivial claim. It is denied by the moral nihilist, who claims that there are no answers to moral questions. However, it does include views which look slightly nihilistic. For example, an error theorist claims that all positive atomic moral judgements are false – whenever we utter one, we are in error. Typically though, the error theorist will offer us *some* way of going about answering moral questions.[8] In other words, a non-eliminativist error theorist will hold that there are 'right' and 'wrong' ways of answering moral questions that do not have to correspond to true and false moral propositions – as all positive, atomic, moral claims are false. If they can do this, then they would count as a procedural realist. Procedural realism also encompasses non-cognitivist views (for a non-cognitivist, there is an answer to the question 'Is murder wrong?', just one that does not involve the notion of independent moral facts), Korsgaard's neo-Kantian constructivism, and some forms of cognitivist anti-realism.

Korsgaard accepts procedural realism. When she offers an attack on moral realism, it is not intended as an attack undermining the claim that there are answers to moral questions. Instead, she means to attack the substantive realist claim that those answers are only available because our procedures for answering moral questions aim to track independently existing moral facts. This leaves her with the need to accept **MPR** if she wants to avoid nihilism. So, her attack is, in one sense not against realism. However, all of the above listed versions of moral realism (those positions that accept my **1- 3** above) *do* form a target for her attack. They are all committed to the truth of substantive realism.

[8] See, for example Mackie 1977

Thus we can see how wide-ranging Korsgaard wants her attack on moral realism to be. She claims that substantive realism does not have the resources to answer the normative question, and for this reason should be rejected. This means rejecting all of the above types of realism – naturalist, non-naturalist and so on, alike.

## 1.3 Korsgaard Against Moral Realism

Korsgaard contends that we can see where moral realism goes wrong if we first look at how voluntarism deals with the normative question. Voluntaristic theories about obligation claim that obligations are grounded in the commands or choices of a legislator. The most well-known variant is theological voluntarism where obligation derives from the commands or will of God. However, voluntarism has space to slot in any particular legislator – for example, for Thomas Hobbes (1651) the relevant legislator was an earthly sovereign. All unsophisticated variants of voluntarism would endorse a claim like the following:

> **VOL**: If agent *x* is obligated to perform action *a* then this is because the legislator commands, or in some other way wills, *a*.

With the role of legislator being taken by different entities. Korsgaard argues that such an account of obligation fails because it cannot provide an adequate answer to the normative question.

The voluntarist tells us that all our obligations stem from the commands of some legislator. We can then pose the following question: why am I obligated to obey those commands? According to the theory, all obligations come from the commands of the legislator, so it must be because she commands my obedience. But this cannot be right: the legislator cannot make it the case that I should obey their commands just by commanding that I do so – unless I'm already obligated to obey their commands then such a command will make no difference. The answer that the voluntarist offers to the normative question is thus circular – commands inheriting their normative status from being the commands of a particular legislator would depend upon the commands of that legislator already having normative significance. This significance can be established only by a further command. We can then repeat our question of 'what's so special about that command?' indefinitely. Thus this answer to the normative question is incoherent.

One way to avoid this incoherence would be to claim that our obligation to obey the legislator lies in something else. Pufendorf (1672), another voluntarist, claims that we have an obligation to obey the legislator when they have *legitimate authority* over us. But if we follow this path, we have in effect given up on our voluntarism. Our obligations are now explained by, or consist in, something else – the legitimacy of the legislator's authority in Pufendorf's case. And now the normative question can be just reiterated. First we will ask what is it about the legislator that gives them legitimacy, and then we can ask why that means I should obey their commands.

So, in summary, the voluntarist tries to answer the normative question by saying that the obligations stemming from the commands of a suitable legislator are justified. However,

voluntarism fails because we can ask why these commands are justified – if it is just because the legislator has commanded that we obey them, then the position is inadequate; if they inherit their justification from something else then we have given up on being a voluntarist.

Considering this shows up a dilemma when looking for the authority of obligation from a substantive realist. We can claim either:

i.) Its authority comes from morality, in which case we have argued in a circle.

ii.) Or, its authority comes from something else. In this case we can ask where that something else's authority comes from, and we are on the road to some kind of infinite regress of justification.

The voluntarist account of obligation fails because we can always ask why we should obey the legislator's command. It cannot be because they have commanded us to, because the same question arises about *that* command. The voluntarist thus fails to tell us why we should feel obligated to obey the legislator's commands, and thus fails to give an adequate account of obligation.

So the voluntarist faces a problem – attempting to root obligation in the commands of a legislator without generating a regress of justification (a 'normative regress'). Korsgaard contends that substantive realism fails as it attempts to end this regress illegitimately. The substantive realist brings the threatened normative regress to an end by fiat by positing intrinsically normative entities (facts or truths that exist independently of our procedures for answering moral questions) that are supposed to stop a repetition of the normative question.

For Korsgaard, this is a way of avoiding answering the question at all. Instead of telling us why some actions are obligatory, the realist posits intrinsically normative entities or relations found in the world – some actions are simply right, and this is because these actions are intrinsically obligatory. These normative entities are supposed to forbid further questioning – once we have discovered that certain actions are intrinsically obligatory, that will be the end of the matter.

Korsgaard holds that this does not engage with the normative question at all. What is at issue is whether there are any obligatory actions, and if there are whether they are the ones we are traditionally asked to do. In this, realism seems to be of little help. For the realist answer to the question 'why should I perform such-and-such an action?' appears to be 'because that action is intrinsically obligatory.' But this is the very thing the person asking the normative question is questioning. It appears as if the realist's answer can only be backed up by their confidence that such entities exist, whereas the person asking the normative question is asking it because they lack such confidence.

Therefore, the realist's answer to the normative question is inadequate because it is no answer at all – it merely restates the realist's confidence in the existence of intrinsically normative states of affairs or relations. The inadequacy of this line of response is revealed by asking how a realist would respond to someone who had lost their confidence in the normativity of morality. At best they would be able to explain whether a particular action was demanded by morality, not why you should act in line with the dictates of morality at all.

To summarise, Korsgaard argues that moral realism lacks the resources to generate an adequate answer to the normative question. The voluntarist attempts to answer that question by citing the commands of an appropriate legislator. But we can ask of these commands how they earn their normative justification. If the voluntarist uses some consideration other than the commands of the relevant legislator then they have given up being voluntarists (and we can, in any case, simply ask how that other consideration earns its normative force). If instead the voluntarist simply claims that the commands of the legislator are authoritative because the legislator commands that we obey them, then our question has not been properly answered – we can reiterate our question and ask why that command should matter. Korsgaard claims that at this point the moral realist engages in something akin to foot-stamping – they merely insist that obligation exists by positing intrinsically normative entities. Such a move is illegitimate (according to Korsgaard) because it completely ignores the normative question – instead of explaining why you are obligated to perform a particular action, the moral realist simply insists that you are so obligated.

In this thesis I will explore two types of response to Korsgaard's argument against moral realism. First, I will investigate whether after we bracket considerations to do with the motivational force of moral judgement Korsgaard's problem remains (§2.2). Second, we will see if there is a solution to Korsgaard's dilemma for the voluntarist which can be used by the moral realist as well (§3.2). In addition, I will explore Korsgaard's own answer to the normative question (Chapter 5). If this answer to the normative question fails then we may suspect that Korsgaard's question is illegitimately posed, and something like Prichardian scepticism towards it is merited. At the very least we will have established that the moral realist is in no worse position than the neo-Kantian constructivist.

Before we get this far, however, it is worth considering whether there is *any* potential for Korsgaard's argument to have any force against moral realism. Nadeem Hussain and Nishi Shah (2005, 2006a, 2006b) have argued that Korsgaard's argument cannot undermine moral realism as it is pitched at the wrong level of theoretical generality to have that consequence. I will consider this claim in the next section and hope to demonstrate that there is at least a *prima facie* case for taking Korsgaard's argument seriously.

## 1.4 Korsgaard's Rejection of Realism and the Distinction Between Normative Ethics and Metaethics

Hussain and Shah are interested in Korsgaard's project of attempting to transcend or 'go beyond' the distinctions of traditional metaethics. Against this project they argue that Korsgaard's arguments do not have any metaethical conclusions at all so do not license either a) rejecting any particular metaethical view or, b) attempting to transcend the traditional distinctions. If they are right, then Korsgaard's argument against moral realism must fail – moral realism is a view within metaethics, and if Korsgaard's claims cannot generate any metaethical consequences she has no reason to reject it. Hussain and Shah couch their argument in terms of Korsgaard's dissatisfaction with the non-naturalist realism advocated by G.E. Moore (1903) and H.A. Prichard (1912) so I will follow them in taking this type of moral realism as our test case.

Hussain and Shah lay out their complaint against Korsgaard as follows:

> Our general strategy will be to argue that what are supposed to be claims that conflict with realism in fact fail to do so. We will rarely attack the arguments for these claims. What we will attack instead is the argument against realism based on these claims. These claims (and arguments for them) fail, in general, to undermine realism because Korsgaard fails to show that they actually conflict with realism in the first place. They often fail to conflict because though they appear to be metaethical claims they in fact are not obviously so and indeed are most charitably interpreted as either claims within normative ethics or normative psychological claims in the philosophy of action, claims compatible with several metaethical accounts of those same claims including non-reductive realism.
>
> (Hussain and Shah 2006a, 266)

So then, their strategy seems clear. They wish to show that Korsgaard's rejection of realism is based on a mistake about the scope of her claims. Korsgaard's objection to non-naturalism fails because it is an objection with no metaethical implications, and as such cannot undermine realism (a metaethical position). Once we have seen the mistake in question, we will realise that Korsgaard cannot differentiate her position from non-naturalism (or any other metaethical position – e.g. non-cognitivism; see Hussain and Shah 2006b); she cannot successfully offer an account to rival traditional metaethical theories, or succeed in attempting to 'go beyond' traditional metaethics.

The mistake in question is "a failure to appreciate all the consequences of the traditional distinction between normative judgments and metaethical interpretations of normative

judgments." (Hussain and Shah 2006a, 266). So what is the traditional distinction, and which consequences of that distinction does Korsgaard fail to take account of?

On the first question, Hussain and Shah give us the following account. In the domain of normative ethics, they place two differing philosophical tasks:

i.) "To construct a set of principles that systematize and ground our correct moral judgments"

ii.) "To place morality within practical reason, explaining whether we have reason to do what morality demands and, if so, whether these reasons are derived from another branch of practical reason"

(Hussain and Shah 2006a, 266-7)

In contrast, the job of metaethics is to give us an interpretation of the normative claims made in the process of carrying out the above tasks. Specifically, to spell out the semantic, metaphysical and epistemological commitments entailed by our normative claims (so, for example a non-naturalist offers us a fact-stating semantics for moral discourse, an ontology of non-natural properties and some kind of intuitionist epistemology. In contrast a non-cognitivist claims that moral judgements express desires, offers an ontology of natural properties, and as they don't think there is such a thing as moral knowledge they do not need to offer a substantial epistemology. Obviously these are very crude caricatures of the most simple versions of those positions). Of course, Hussain and Shah acknowledge that discourse in either domain can impact on the other. Nevertheless, they contend that Korsgaard's failure to fully appreciate the different tasks of the two domains leads to her rejection of non-naturalism on spurious grounds.

The consequence of this distinction that Korsgaard fails to notice is an ambiguity in what the normative question is asking for. We can make a distinction between what *makes* an action wrong, and what *constitutes* the normativity in question. The example Hussain and Shah consider is brushing one's teeth:

> Thus, the fact that brushing my teeth regularly will reduce plaque may *make* brushing my teeth good (for me); however, we do not want to claim, presumably, that the property of goodness itself just is the property of reducing plaque.
>
> (Hussain and Shah 2006a, 270).

To extend our shaky analogy further, we could say that the fact that brushing teeth reduces plaque places the claim 'you should brush your teeth' within practical reason (it gives us reason to do what the imperative demands), but it does not give a 'metaethical' account of the goodness of brushing teeth (it tells us nothing about the metaphysical, semantic and epistemological commitments of the practice of ascribing goodness to tooth-brushing). Here we have two different notions on the scene: loosely speaking the 'normative-making properties' of teeth brushing; and what constitutes that normativity.

It is Korsgaard's failure to notice this consequence of the distinction between normative and metaethics that lies behind her dissatisfaction with Moore's non-naturalist realism. Korsgaard claims that the reason that the open-question has any force is because of the force of the normative question:

> That is, when the concept of good is applied to a natural object, such as pleasure, we can still always ask whether we should really choose or pursue it.

However:

> This should not lead us to conclude that the concept of good, or any other normative concept, cannot be defined in a way that guides its application. Conflation of the normative question with other questions is what drives Moore and others to the view that moral concepts must be simple and indefinable, and as a result to intuitionism.
>
> (Korsgaard 1996 43)

The problem with this conflation and the Moorean answer it leads to would be that such a conclusion would be of no help to someone asking the normative question. Such a person wants to know if the claims of morality really are justified, and to say that they are justified by the existence of intrinsically normative entities is of no help ( it is exactly the existence or relevance of such entities that the questioner is doubting). The problem in the case of non-naturalism is even stronger, as the answer to the normative question appears even sketchier. The bones of Korsgaard complaint appear to be:

> [T]hat Moore, like Pritchard, failed to distinguish the question whether a normative concept has been correctly applied from the 'normative question,' and thus that Moore mistakenly thought that because no naturalistic answer can be given to 'the normative question,' there can be no naturalistic criteria given to guide the application of a normative concept.
>
> (Hussain and Shah 2006a, 273)

However, Hussain and Shah point out that Moore did claim that there are synthetic necessary truths connecting normative and natural properties, and thus could accept a naturalistic account of the normative-making properties (see Moore 1903, 9). For example,

according to them, Moore could accept that it was a synthetic necessary truth that pleasure is good, and thus what makes a certain action good is that it produces pleasure. Nevertheless, he rejected a naturalistic account of what constitutes that goodness – good is not identical to any natural property, but instead is simple and indefinable. What this example illustrates, is that one can give an account of the normative-making properties in play, without yielding any metaethical conclusions. In this case, Moore could give a naturalistic account of normative-making properties, whilst holding a non-naturalistic view about the ontology of moral properties. By failing to distinguish between what *makes* an action good, and what *constitutes* goodness, Korsgaard misdirects her efforts. In her discussions of Moore and Prichard she attacks them for offering inadequate accounts of the placing of duty and good within practical reason, on the grounds that realism in general answers the normative question with a statement of its confidence in the existence of intrinsically normative entities, which is of no help to the person who has fallen into doubt. However, this does not show that they do not have reasonable accounts of the semantical, metaphysical and epistemological commitments of our normative claims – her arguments have no impact within that domain.

So, in summary, Korsgaard fails to comprehend the full implications of the traditional distinction between normative and meta-ethics. Thus she fails to distinguish between the question of what makes an action right, and what constitutes that normativity. Her dissatisfaction with the non-naturalists can be pinned to the fact that she believes that their answer to the metaethical question is intended to tell us something about the placing of a particular normative claim within practical reason. It may well be the case that the way in which the non-naturalist places our normative claims within practical reason is inadequate, but this does not mean that they are wrong about the metaethics. Thus Korsgaard's

objection that non-naturalism cannot adequately place our normative claims within practical reason has no metaethical import – she is wrong to reject non-naturalism on the grounds of justificatory inadequacy in the normative domain, for such questions stand apart from the metaethics.

Now, put like this, Hussain and Shah's argument may appear slightly strange. We would only expect to be unable to reject certain metaethical positions on the basis of inadequacy in the normative domain if we felt that the two areas did stand apart in such a sharp fashion. You may even start to wonder whether Hussain and Shah are securing their conclusion via a very austere conception of the tasks of metaethics.

On this second point, it's worth noting that other taxonomies dividing metaethics from other philosophical tasks give much more of an expansive role to metaethics (e.g. Miller 2003). These more expansive conceptions of the tasks of metaethics are borne out if we consider one of the implications of Hussain and Shah's own division between metaethics and normative ethics. One of the 20th century's most pervasive arguments in favour of non-cognitivism cites the motivational import that moral judgements have. In brief, the non-cognitivist argues that:

(4) Moral judgements are inherently motivating

(5) Beliefs have no motivational important, and beliefs and desires are distinct existences with no necessary connections between them

Therefore:

(6) Moral judgements do not express beliefs. Instead they express non-cognitive desire-like states (which are inherently motivating).

Now, if Hussain and Shah are right then this argument is not metaethical – it uses considerations to do with the motivational force of moral judgements (a consideration outside of the remit of metaethics, as they conceive it) to derive its conclusion. But this is surely absurd – this is a paradigmatic metaethical argument, used frequently to drive non-cognitivism, and if a taxonomy rejects it as non-metaethical then that is reason to reject that taxonomy.

Hussain and Shah could argue that this argument *is* metaethical because its second premise relies upon taking a view on the metaphysics of belief – about what kind of thing they are, and what they can do. However, if this is the way of making the argument above fall within the domain of metaethics, then we could make a parallel claim for Korsgaard's own argument against moral realism. We could represent Korsgaard's argument, very schematically, as follows:

(7) At least some moral judgements are justified.

(8) The entities available to a moral realist are incapable of justifying any moral judgements.

Therefore

(9) Moral realism is false.

As in the argument for non-cognitivism the first makes a claim about some feature of moral judgements. The second premise makes a claim about what the resources available to a moral realist can and cannot do. If we wish to spare Hussain and Shah blushes by calling premise **5** of the argument against cognitivism metaphysical, there seems no obvious reason to not treat premise **8** of Korsgaard's argument against moral realism in a similar manner.

Returning to the first point – of whether normative ethics and metaethics really do stand so far apart that we cannot use considerations of justificatory adequacy as a criterion of success for a metaethical theory – it seems implausible to think that metaethics and normative ethics really do stand apart in such a fashion. The claim that conclusions about what the normative-making properties are has no bearing on what constitutes that normativity only seems plausible if considering a single case.

What do I mean by this? Suppose you were considering some action $\varphi$, generally accepted as good. You want to know first what it is that makes $\varphi$-ing good (what its normative-making properties are) and second what constitutes that normativity. Now, if the answer to the first question was that what makes $\varphi$-ing good was the fact that it promotes the greatest balance of pleasure over pain, we wouldn't (presumably) want to say that gave us an answer to the second question – we wouldn't have given an account of what 'good' means, or its metaphysical characteristics, or how we gain knowledge about it. However, suppose now that we continued our enquiry, looking at more and more different good actions, and in each case found that what makes $x$, $y$ or $z$-ing good was the fact that they promote the greatest balance of pleasure over pain. Would it not now seem more reasonable to conclude not that we have given no account of the metaethics of 'good', but rather that good just *is* the

property of promoting the greatest balance of pleasure over pain? Suppose that after an extensive enquiry over many different cases we get the same answer – what makes ***φ***-ing good is the fact that it promotes the greatest balance of pleasure over pain. Have we still not reached something with metaethical implications? I am tempted to say that we have done quite a lot to unearth such implications, and it would be reasonable to accept a utilitarian realism about moral properties.

The example need not be so simple. Suppose instead that we found that actions ***x,y,z…*** possessed 4 or 5 different natural properties ***a,b,c…*** such that some proportion of ***x,y,z…*** had natural property ***a*** in common, some ***b*** in common and so on. And suppose that every action examined so far was made good by its possession of ***a*** or ***b*** or ***a & b*** and so on. I believe the obvious conclusion to draw would be that good is some disjunction of those properties. Suppose instead that what makes actions good is wildly heterogeneous – there is no natural property or disjunction of natural properties that unifies them. Then it would be tempting to conclude that goodness was some indefinable non-natural property.[9] Or, if what unified the good actions was that we were inclined to express approval of them, some version of subjectivism would be on the scene. Or, more relevantly to this case, suppose that the characteristic that all good actions shared was that to perform them would be to act

[9] The most obvious comparison here is with the case in the philosophy of mind. There multiple realisability of mental states by physical states is usually taken to block reduction of the mental to the physical. Wild heterogeneity could also give a *prima facie* case for non-reductive ethical naturalism of the type proposed by Sturgeon (1985). The wild heterogeneity indicates that moral properties may be irreducible. If we combine this with Sturgeon's argument that moral facts have a distinctive role to play in our explanation of people's moral beliefs (and earn their non-reductionist stripes because of this), then we might be drawn towards non-reductive naturalism. Without this second claim (about the role of moral facts in empirical explanation) we might be tempted to stick with the idea of the irreducibility as lending support to non-naturalism. In the case of the mind, we can note that some have claimed that, given the overwhelming evidence in favour of physicalism (e.g. the causal closure of the physical world, see Papineau (unpublished) multiple-realisability forces towards a version of non-reductive naturalism.

upon a maxim that could be willed as a law by someone acting as a universal legislator in a Kingdom of Ends (the answer Korsgaard gives), then we might be inclined to accept some formulation of neo- or plain Kantianism.

Of course, care must be taken here. There are senses in which it is arguable that the scenarios presented above are all compatible with error theory or certain types of non-cognitivism.[10] However, I think the point still stands, that - in the absence of further argumentation - our answer to 'what makes this action good?' does have some implications for the kind of answers we could give to the question 'what constitutes that normativity?' Thus, returning to Moore, it turns out that Moore may have given too much ground in admitting the existence of synthetic necessary truths linking normative-making properties to moral properties. It seems as if the natural conclusion to draw would be that the moral property of goodness is identical to the normative-making properties (singularly or in some disjunction). It then turns out that Moore was wrong in his rejection of naturalism for the same well worn reason – the possibility of synthetic identification between moral properties and natural properties, which the open-question argument (even if meets its intended target) does nothing to undermine.

Another way to see the force of my criticism is to consider the recent exchange between Frank Jackson (1998) and Jussi Suikkanen (2010). Jackson argues, against the non-naturalist, that commitment to supervenience restrictions forces the non-naturalist to admit

[10] The non-cogntivist can claim that the properties that we find are correlated with our judgements are the ones which we humans, as a matter of empirical fact, tend to disapprove or approve of. The error theorist too will happily admit the relevant natural properties to their ontology, but claim that thinking of these as *moral* cannot be sustained.

that it is in principle possible to generate a disjunctive description of all right (for example) acts. Then we can refer to the property using the predicate 'being $D_1$, or $D_2$, or $D_3$, or... $D_n$' (where $D_{1\text{-}n}$ are completely naturalistic descriptions of right acts). This conjunctive property will also be naturalistic. So, 'being right' and 'being $D_1$, or $D_2$, or $D_3$, or... $D_n$' will refer to the same set of acts. Jackson claims that, for a number of reasons, we cannot have necessarily co-instantiated properties. This means that there is only (at most) one property for 'being right' and 'being $D_1$, or $D_2$, or $D_3$, or... $D_n$' to refer to. As the second predicate is, by hypothesis, naturalistic, there isn't space for there to be an additional non-natural property of rightness.

The non-naturalist can reply that there are cases where we seem to have necessarily co-instantiated distinct properties (e.g. the property of God willing it be light, and the property of it being light – given God's omnipotence he can't will it to be light without it being light, but these are not (presumably) the very same property. See Philip Goff's (2007)). However, Jackson's worries about permitting necessarily co-instantiated properties[11] are still in play, so the non-naturalist that takes this line needs to find a way of distinguishing between cases where we have two predicates designating the same property, and those where (like, they suppose, the case of rightness) where two necessarily co-extensive predicates designate distinct properties.

[11] Basically centring around being unable to come to a determinate answer of how many properties are instantiated at any time if there can be a distinct property for each predicate that truly applies – we can gerrymander up many predicates that truly apply, but suspect that not all of them are distinct, and if we give up the ban on necessarily co-instantiated properties we need a new way of counting the properties.

This prompts Suikkanen to undertake an investigation of how and when the non-naturalist can distinguish necessarily co-extensive properties, and offers other reasons to conclude that Jackson is begging the question against the naturalist. I wish to claim here that Jackson attempts, in effect, a shortcut version of the strategy I am pursuing here. Instead of claiming that 'rightness' and 'being $D_1$, or $D_2$, or $D_3$, or... $D_n$', because necessarily coinstantiated, *must* be the same property I merely claim that if 'being right' is coinstantiated with some naturalistic predicate or suitably short disjunctive naturalistic predicate then this is *prima facie* (absent further argument) evidence that being right just is that naturalistic property. Thus, I hope to offer a weaker (and hence more plausible) argument that shares something in common with the Jackson strategy. What I think this reveals is that Hussain and Shah's complaint is off-beam. Suikkanen does not respond to Jackson merely by adverting to the difference between good-making features and what constitutes goodness. Instead he undertakes a sustained investigation of when the non-naturalist is licensed to distinguish between necessarily co-instantiated properties.

Sure enough, the non-naturalist may be able to make use of this distinction, but Jackson's argument puts some pressure on the ease with which they can do this. Suikkanen responds by trying to show where the non-naturalist has the right to this distinction, thus rendering Jackson's objection inert. I am calling attention to the fact that although Suikkanen's response to Jackson is fair enough, it does not completely close off the strategy I'm advancing here – where correlations between naturalistic properties and moral properties are taken as *prima facie* evidence in favour of their identification. Furthermore, the debate between Jackson and Suikkanen makes no sense if Hussain and Shah are right – the non-naturalist does not have to bother showing that they are entitled to the distinction between the necessarily coinstantiated properties they posit, they can instead merely state it as a bald

assertion. I think this is more evidence that Hussain and Shah's assault on Korsgaard is misplaced.

If this is right, then the distinctions that Hussain and Shah use to undermine Korsgaard's argument against moral realism are not robust enough to bear the weight required. They need to provide us with some positive reason to think that the ability of a metaethical theory to provide resources capable of meeting standards of justificatory adequacy is an unsuitable criterion. At present, there seems to be a case for claiming that providing resources that offer justificatory adequacy is an important test of a metaethical theory's viability. If Korsgaard is right about the inability of moral realism to deal with the normative question, that appears, at this stage, to be a mark against moral realism, absent any further argumentation about the relevance or otherwise of the normative question.

In summary, we have been attempting to get a grip on what Korsgaard's normative question is, what it is asking for, and what consequences follow if we start to take it seriously. Korsgaard claims that moral realism cannot take the normative question seriously and that gives us reason to abandon moral realism (where moral realism is construed as substantive rather than merely procedural). I have attempted to rebut an argument to the effect that Korsgaard's question is confused. In the next chapter we will move on to seeing how the moral realist can respond to Korsgaard's challenge, beginning with what happens to the normative question if we clear up issues to do with moral motivation.

# CHAPTER TWO: INTERNALISM

As we have seen, Korsgaard's normative question presents some kind of problem for moral realism, one that cannot be resisted through pointing to the distinction between normative ethics and metaethics. However, neither have we seen that the normative question cannot be answered by the moral realist – simply that providing such an answer could be a criterion of adequacy for a metaethical theory. Also, it is not yet clear precisely what the normative question is asking for and what an adequate answer to it would look like. This chapter investigates whether Korsgaard's concerns are grounded in the motivational effects that moral judgements are typically felt to have, and in the next chapter we will look at how the moral realist can best respond if these are genuinely the neo-Kantians concerns. An alternative way of reading Korsgaard is to see her as offering the generalised anti-voluntarist argument identified by Mark Schroeder. I will explain this argument and how it relates to Korsgaard's work, before seeing how the moral realist can respond. I will be attempting to force a dilemma on the neo-Kantian constructivist: either their concerns are rooted in the motivational effects of moral judgements (in which case the moral realisms canvassed in the next chapter provide a viable response to those concerns) or they are getting at the generalised anti-voluntarist, an argument to which the moral realist and neo-Kantian constructivist are equally vulnerable, and to which the moral realist has a response.

Here, I will first illustrate some connections between Korsgaard's concerns and issues to do with the motivational force of moral judgements. She claims that her normative question explains the force of Moore's open question argument and Mackie's argument from

queerness – two arguments where we can give an interpretation in terms of motivational force.

## 2.1 The Normative Question: Moore, Mackie and Internalism

We saw above how Korsgaard thinks that the normative question causes trouble for the moral realist. When we ask the realist 'why should I be moral?' they cite the existence of a moral fact, a fact that some action or another is intrinsically obligatory. This fact is not suitable for answering the normative question, however, so the realists answer to the normative question fails. Because the cited fact is not suitable, and can't get a grip on the agent asking the normative question in the right kind of way, we need to cite another to ground that fact and then we are on the way to some kind of infinite regress.

However, it is not yet entirely clear what the unsuitability of the moral realist's moral facts consists in. What is it about a moral fact that stops it being an appropriate response to the normative question? Here I will suggest that we can offer an interpretation of Korsgaard's challenge to the realist that makes the problem clearer, and is commensurate with her remarks on Moore's open question argument and Mackie's argument from queerness.

I will begin by laying out the connections Korsgaard believes obtain between the normative question and the open question argument (§2.11) and the argument from queerness (§2.12).

Then I will offer an outline of judgement internalism (§2.13) and its connection to the open question argument (§2.14) and the argument from queerness (§2.15). With these elements in place I will be in a position to explore the connection between them and show how the moral realist can attempt to escape from Korsgaard's argument against them (§2.16).

## 2.11 The Normative Question and Moore's Open-Question Argument

Korsgaard contends that the normative question explains the force of both the open question argument and the argument from queerness. Turning to the open question argument first: Moore (1903) argued that 'good' could not be synonymous with any naturalistic predicate[12], for whatever definition you give of 'good' in naturalistic terms it is always an open question whether 'good' applies to a state of affairs to which that naturalistic predicate does. Therefore, 'good' cannot be equivalent to any naturalistic predicate as a matter of conceptual necessity. Moore can be criticised on a number of grounds – that he begs the question against the naturalist, that he presupposes a very austere notion of what conceptual analysis can achieve, and that the argument can be avoided by moving away from definitional naturalism to synthetic naturalism (see Miller 2003, 15-18 for a good discussion). However, Korsgaard argues that his argument has some merit, and what merit it does have it acquires from the force of the normative question.

[12] Moore also argues that the argument works for 'metaphysical' predicates. Here he has in mind supernatural theories that link morality to God's commands.

Moore argues from a particular datum, that the question 'Is this action, which falls under the naturalistic predicate N[13], really right?' is open. That is, sincerely asking the question does not betray any conceptual confusion. Compare this question to the similar one in the case of a potentially analytic truth: 'Is this man, who is an unmarried eligible male, really a bachelor?' in that case, if any, competence with the terms involved suffices to provide an answer, and if you sincerely asked the question that would betray some conceptual confusion on your part. Moore then goes on to derive a metaphysical conclusion from this evidence against conceptual connections between moral and naturalistic terms (that goodness is a non-natural property). Korsgaard agrees that the question 'Is this action, which falls under the naturalistic predicate N, really right?' appears to be open. However, she argues that the openness of this question is not due to the lack of a conceptual connection between rightness and the naturalistic analysis. Even if such a connection held, the open question would still appear open. This is because when we ask the open question what gives the appearance of openness is not the openness of that question, but the openness of a nearby question with which it is confused.

This question is 'Is this action, which is right, really obligatory?'. This is, in effect the normative question. Although Korsgaard is never explicit on this, she must be thinking that when we ask the normative question we are making something like the mistake she attributes to Prichard. What we are doubting, in asking the question, is whether the action in question is really obligatory. Like Prichard, we confuse 'being right' with 'being really obligatory', and thus express our question by asking whether the action in question is really right. But this is not what we are interested in – instead, we are interested in whether I

[13] Where N is the naturalist's analysis of rightness.

should choose to perform right actions, in whether they are normatively binding for us. In a way, Korsgaard is arguing that Moore's diagnosis of the normative question is superficial – it doesn't get at what is really driving our scepticism over whether the action in question is to be performed.

Were the normative question closed we could still get into a dispute over the question: 'is this action, which is N, really good?' However, it would be because we were arguing about whether we have the *right* definition of 'good', not over whether such a definition is possible: so I suggest a naturalistic definition and you disagree about that particular proposal – perhaps you think another naturalistic definition is right, or that we should adopt some kind of supernatural definition (perhaps that 'good' means 'commanded by God'). However, what you would not suggest is that we give up on the whole enterprise (at least not for the reasons Moore thinks we should: we might still think that 'good' was unanalysable, but for different reasons – perhaps because we don't think there are any general principles linking moral facts to natural facts.[14]

It is the lack of attention to the relationship between the normative question and the question in the open question argument that leads Moore to his non-naturalism, where goodness is a simple, indefinable, *sui generis* non-natural property:

[14] See Dancy (2000, 2004 and 2006) for something like this position, and Jackson, Pettit and Smith (2000) and McKeever and Ridge (2004) for replies on this point).

> Moore argued that no matter what analysis we give of 'good', it is an open question whether the objects picked out by that analysis are good. And he concluded that 'good' must therefore be unanalyzable, and further that therefore we can only know which things are good through intuition. But the force of the open question argument clearly comes from the pressure of the normative question. That is, when the concept of the good is applied to a natural object, such as pleasure, we can still always ask whether we should really choose or pursue it. This should not lead us to conclude that the concept of the good, or any other normative concept, cannot be defined in a way that guides its application. Conflation of the normative question with other questions is what drives Moore and others to the view that moral concepts must be simple and indefinable, and as a result to intuitionism. (Korsgaard, 1996 43).

If Moore had realised what the open question was trying to get at – the normative question – he would realise that it is not naturalism that is faulty, but moral realism more generally. There is the problem of explaining the normative force of our obligations, and citing moral facts to ground those obligations doesn't help – whether the facts in question are natural or non-natural.

## 2.12 The Normative Question and Mackie's Argument from Queerness

John L. Mackie (1977) advances an error theory of moral judgements (that is, a theory that claims that all positive, atomic moral judgements are systematically false) via two claims:

one conceptual and one metaphysical. The first, conceptual claim, is that moral discourse is cognitive – moral judgements express beliefs which are about the instantiation of moral properties. However if such moral properties existed they would be metaphysically 'queer'. Mackie's metaphysical claim is that no such entities exist. So, we have an area of discourse that is in the business of expressing beliefs about the instantiation of a type of property that does not exist. Therefore the discourse is radically in error, and all positive, atomic moral judgements are false.

But what does the queerness of moral properties consist in? What makes them so outlandish that we know that they cannot exist? Mackie claims that moral properties, if there were such things, would have to have built in 'to-be-pursued-ness' or 'not-to-be-done-ness'. Some kind of magnetic force that pulls creatures like ourselves towards and away from objects and actions that instantiate them. Such features of a property cannot be made to fit into a naturalistic conception of the world, Mackie argues, so if we are serious about naturalism we must abandon our commitment to them and accept an error theory.

Korsgaard argues that the point Mackie makes is a real one, but not the one he intends to make. The realist, in answering sceptical challenges has to build intrinsic obligation[15] into the fabric of the world. However, this is inadequate as:

[15] Korsgaard's analogue of Mackie's 'to-be-pursued-ness' and 'not-to-be-doneness'

> If someone falls into doubt about whether obligations really exist, it doesn't help to say 'ah, but indeed they do. They are *real* things'. Just now he doesn't see it, and herein lies his problem. (Korsgaard, 1996 38).

The realist way of responding to the sceptic who doubts whether anything is obligatory doesn't really engage with their worry at all. It is of no help to tell them that obligation is a real part of the world. So Mackie is right to note that there is something wrong with trying to build obligation into the world. He thinks that doing so would add something irredeemably weird to our ontology, and is not worth the cost. However, if Korsgaard is right, then the real point that the argument from queerness gets to is the inadequacy of trying to build morality into the world at all. It wouldn't help us with what we need an account of obligation to do – placate someone who has fallen into doubt over whether the requirements of morality are obligatory.

If Korsgaard is right, then Mackie's problem would remain even if moral properties did not have to be so metaphysically queer. John McDowell (1998) argues that secondary qualities have something like the 'magnetic' element Mackie finds objectionable in moral properties – redness has 'to-be-seen-as-red-ness' built into it, for example. Using this comparison you can try to develop a companions in guilt style defence of moral properties[16]: moral properties would be no more weird than properties that you already believe exist (e.g. redness), so unless you drop your commitment to the existence of those properties you already believe in, you have no reason to think that moral properties don't exist. If this

[16] For a fuller explanation of companions in guilt strategies and their application to the moral domain see Lillehammer 2007.

method did work[17] then we would have shown that moral properties did not have to be any more queer than more metaphysically hygienic properties like redness[18]. However, Korsgaard must claim, even if we managed this manoeuvre, we still wouldn't have solved the problem Mackie is getting at – what is wrong with building morality into the fabric of the world is not what it would look like if we got there, but the fact that such an attempt is redundant, it doesn't help us at all.

So, we have seen that Korsgaard thinks that the open-question argument and Mackie's argument from queerness both gain any force they have from the force of the normative question. I will suggest that there is one way to interpret Korsgaard's argument against realism that makes sense of all these claims (although we will see that it is an interpretation that Korsgaard would not accept – my aim is not to produce something Korsgaard would agree with whole-heartedly, rather merely a way of taking her complaint that could appeal to people who feel there is something in what Korsgaard says, but do not want to follow her all the way to the rejection of realism). All three arguments (Korsgaard's argument from the normative question, the open question argument and the argument from queerness) could be intimately tied up with the stance we take on the link between moral judgement and motivation. In the next section I will outline one stance that could underlie all three complaints – judgement internalism. I will then draw the connections between that stance and the three arguments before showing how a realist can start to respond to all three arguments.

[17] I take no stand on whether it does here – instead I am concerned with what would follow for Mackie's problem if it did and Korsgaard was right about what Mackie's problem really reveals.

[18] Of course, you still might think there is something queer about redness.

## 2.13 Judgement Internalism

Judgement internalists about motivation claim that there is a conceptual (or *internal*) connection between making a moral judgement and being motivated to act in accord with that judgement.[19] That is, it is conceptually impossible for an agent to make a moral judgement (say that something is good) without being appropriately motivated (in the case of goodness, towards obtaining that good thing). As Michael Smith puts it: "*believing I should* seems to bring with it my *being motivated to*" (Smith 1994, 60, emphasis in original). The internalist places great emphasis on the action-guiding nature of moral judgement – to be a moral judgement at all, it seems, a judgement must have practical import.

[19] Many types of internal connections have been proposed. The main division is between internalism about motivation, and internalism about reasons for action. I will be focussing on the former type. In addition, some internalists defend a connection between something other than moral judgements and motivation. Belief internalism claims there is a conceptual connection between moral belief and motivation (see Stratton-Lake 1997). Existence internalism posits a connection between the existence of a moral obligation and motivation (see Brink, 1989). Hybrid internalism claims the connection is between the recognition of a genuine obligation and motivation. I avoid discussing these distinctions between internalism, and instead focus on the more schematic 'judgement' for five reasons. 1. Some of these internalisms are prejudicial against some metaethical views – you cannot be a non-cognitivist and a belief internalist, for example. As I wish to use judgement internalism when it comes to discussing non-cognitivism and hybrid metaethical views (views which incorporate elements of cognitivism and non-cognitivism) in chapter 4 it is better to stick with the more agnostic judgement formulation. 2. Existence internalism is prima facie implausible – it seems hard to see how the mere existence of an obligation could motivate someone if they were ignorant of the existence of a moral obligation. 3. Hybrid-internalism also face a problem with ignorance – it is possible to be motivated by a moral judgement, even if you are wrong about what morality requires and thus there is no genuine obligation corresponding to the judgement. 4. I will go on to argue that there is a certain type of conceptual possibility that militates against all forms of internalism, so the details of what type of thing is linked to motivation (judgement, beliefs, moral obligations or the recognition of a genuine obligation) is not important here, and so it is better to lay the groundwork using the metaethically neutral judgement formulation. 5. Giving serious space to these different types of internalism would give us an unmanageably large taxonomy of internalisms and if the details can be passed over safely, then we should do so.

Internalism comes in varying degrees of strength. Strong internalists claim that moral judgements provide over-riding motivation. Weak internalists claim that moral judgements merely provide some motivation, that can be over-ridden by other (prudential, aesthetic, etc) concerns. For the weak internalist there is no problem acknowledging that people can fail to act in accord with their moral judgements when they have other motivations which over power the motivation brought by the moral judgement. In this sort of case, the strong internalist would have to claim that the agent never made a genuine moral judgement at all. Weak internalism is thus easier to defend. What I shall have to say in this thesis will apply equally well to both types of internalism, so when I use the term 'internalism' I will mean the weaker, more *prima facie* plausible claim.

We also need to get clear on the scope of the internalist claim. Smith admits that agents can fail to be motivated by their judgements under certain circumstances – if they are suffering from weakness of will or "other similar forms of practical unreason" (61). The connection between judgement and motivation he advances is a *defeasible* one – it can fail when we are in the grip of some factor that threatens our practical rationality. In light of this Smith offers the following, which he labels the practicality requirement:

> **PRAC:** If an agent judges that it is right for her to φ in circumstances C, then either she is motivated to φ in C or she is practically irrational.[20]

[20] One issue I do not have space to get into is how the internalist should characterise this practical irrationality clause. They must say something substantive, or we will be worried that the **PRAC** ends up being trivial – simply stating that people will be motivated by their moral judgements unless something stops them being motivated by their moral judgements. However, when internalists like Korsgaard and Smith do give a substantive characterisation of conditions that threaten practical rationality they merely gesture towards conditions like depression and great anxiety. This, arguably, fails to engage at all with the relevant philosophy of psychology literature (see, for example, Levy 2011).

This seems more plausible than claiming that the link between moral judgement and motivation is indefeasible – it allows for the possibility of an agent making a genuine moral judgement without feeling the typically associated motivation. Thus I will be mainly directing my efforts at the defeasible version of internalism.

Nevertheless, **PRAC** is a strong claim. The internalist holds that this is not merely a contingent matter of fact about typical human psychology (the claim is not that as a matter of empirical fact the motivational states of practically rational human agents are in line with their moral judgements). Instead it is a claim of conceptual necessity – it is conceptually impossible to make a moral judgement without being motivated to act in accordance with that judgement (absent any practical irrationality).

The externalist simply denies the internalist's claim. They believe that one can make a moral judgement, fail to be motivated by that judgement, and such a case would not (necessarily) be a mark of practical irrationality. They admit the existence of two sorts of conceptual possibility that the internalist cannot. First, an amoralist: an agent that makes a genuine moral judgement without feeling any motivation to act in accordance with that judgement, where that judgement does not betray any practical irrationality. Second, the internalist also makes a claim about the direction of the associated motivation, and not just its mere existence. So it is not possible to judge that something is evil, say, and find that that gives you motivation to pursue it. Thus the internalist cannot accept the conceptual possibility of an immoralist, or an agent we could call (following Richard Joyce, 2001)

'pure evil'[21]. The immoralist makes genuine moral judgements, which provide them with motivation to act in a contrary way to typical moral agents (they pursue evil and avoid good), where these judgements do not betray any practical irrationality. Denying the coherence of these conceptual possibilities is bold[22], but it is something that the internalist is committed too.

But what does this have to do with the open-question argument?

## 2.14 Internalism and the Open Question Argument

Recall that Moore argues that we cannot give a naturalist account of moral properties (that is, an account that identifies moral properties with natural properties, however we construe 'natural' properties) due to the force of the open question argument. What ever naturalistic analysis we give of goodness, it's always possible to doubt that analysis, and falling into such doubt does not betray any conceptual confusion. Thus goodness must be a non-natural property.

[21] You might suspect that this possibility is ruled out by the fact that for an internalist a moral agent has to be motivated to act in *accordance* with their moral judgements, and Joyce's satanic character does not do that. However, Joyce could say that his agent of pure evil does act in accord with his moral judgement – he responds with the right direction of motivation for his view on the world. Whether or not the agent of pure evil is distinct from the amoralist will not matter here.

[22] They look, *prima facie*, like quite plausible conceptual claims – even if we thought that facts about human psychology meant that there never actually are any amoralists or immoralists

Now the problems with this argument are legendary – chief among which is the move from showing that no *definition* of goodness in terms of natural properties can be right, to claiming that goodness cannot *be* some natural property, or collection of natural properties. If we thought, with Moore, that in order for a proposition to be necessary it also had to be *a priori* and analytic then this move would be licensed. However, the philosophical history of the twentieth century has taught us that we need to treat these three notions (*a priority*, necessity, analyticity) with care, and they may not exhibit the close connection Moore needs for his argument to work.[23]

If we leave this aside it does seem that Moore was on to something even though he exaggerates the force of his argument. Contemporary meta-ethicists have claimed that Moore gives us something like an argument structure, which needs to be filled in with some detail. James Lenman (2006) suggests that "naturalistic understandings of moral concepts do indeed omit something central to them" (1). So, there *is* something wrong with a naturalistic account of moral properties, but as of yet Moore hasn't told us what this is. Stephen Darwall, Alan Gibbard and Peter Railton (1992) suggest that the missing ingredient is judgement internalism. With this in mind, they offer a revised open question argument which can be represented like so:

> (1) There is a conceptual or internal link between making a moral judgement and being motivated, *ceteris paribus*, to act as that judgement prescribes [judgement internalism]. Absent some weakness of will or other psychological affliction, judging that a type of action is morally good entails being motivated to perform

[23] For a survey of other problems with Moore's argument see Miller (2003) ch. 2; for a detailed explanation of how Moore's historical position blinded him to the deficiencies of his argument see Soames (2003).

actions of that type. Someone with no psychological afflictions etc who apparently judges that a type of action is morally good but consistently claims that he has no motivation to perform actions of that type doesn't grasp the concept of moral goodness.

(2) Competent and reflective speakers of English are convinced that they are able to imagine clear-headed (and otherwise psychologically healthy) beings who judge that R (some naturalistic property) obtains but who fail to find appropriate reason or motive to act in accordance with that judgment.

(3) If there were no conceptual link between judging that R obtains and being motivated to act accordingly, we would expect competent and reflective speakers of English to have the conviction described in **2**.

(4) So, unless there is some other explanation of the conviction described in **2**, we are entitled to conclude that there is no conceptual link between judging that R obtains and being motivated to act accordingly.

(5) So, unless there is some other explanation of the conviction described in **2**, we are entitled to conclude that the judgement that R obtains isn't a moral judgement.

(6) So, unless there is some other explanation of the conviction mentioned in **2**, we are entitled to conclude that the property of being morally good is not identical or reducible to the property of being R as a matter of conceptual necessity. (adapted from Miller, 2003)

This revised open-question argument has some attractive features. First, it gives us the missing ingredient in Moore's argument (judgement internalism); second, it is less strong than Moore's argument, in a favourable way – the argument does not definitively establish that naturalism is false, it just sets out what the naturalist would have to do to defend their

position (find another explanation of the conviction mentioned in **2**); third it has long been suspected that Moore's own meta-ethical position falls prey to his argument, and we can see why using the revised argument – substituting '*sui generis*, indefinable, non-natural property Q' for 'naturalistic property R' seems to make not a jot of difference to the argument.

This third feature is particularly germane for attempting to generate an interpretation of Korsgaard's rejection of realism where it is underwritten by a commitment to judgement internalism. Korsgaard intends her own argument to have force against both naturalist and non-naturalist moral realism, and that is precisely what the revised open question argument offers us. Now we can turn to the connection between Mackie's argument from queerness and judgement internalism.

## 2.15 Internalism and the Argument from Queerness

We saw that Mackie thinks that we should embrace an error theory about moral judgements because moral properties would have to be untameably queer. But we can ask Mackie, what is so queer about them? After all, don't coloured objects have a 'to-be-seen-as-coloured-ness' built into them? Well, one reason to think that the kind of obligation the realist supposes is contained in the fabric of the world would be queer is if it were intrinsically motivating – just perceiving that it is there would be enough to motivate you to act in accordance with its prescriptions. And this does appear rather odd – how can merely

believing that some state of affairs obtains, by itself, move you to action? But, why should we accept this picture of what moral properties are like? It seems as if such a picture is motivated by some sort of commitment to judgement internalism. If we decide to rid ourselves of this commitment, then it starts to look like moral properties could be the perfectly ordinary, everyday properties the naturalist realist claims they are. Thus, by embracing externalism, we can generate a decent response to Mackie's argument from queerness, whilst retaining our realist leanings (if we have them).[24]

In fact, we do not have to speculate about whether Mackie intends the queerness of the 'to-be-pursued-ness' of moral properties to consist in their motivational import. Richard Joyce (2001) has recently argued for an error theory using precisely this strategy. His first move is to show that moral discourse is committed to judgement internalism, then show how this would make moral properties irredeemably queer, forcing us to reject their existence and accept an error theory. However, as should be plain, this move only works if we are compelled to accept judgement internalism.

Now we have the elements in place to connect up judgement internalism, the argument from queerness, the open question argument and Korsgaard's normative question.

[24] Another way to gloss Mackie's objection is that he objects to the contention that the instantiation of moral properties can generate categorical reasons for actions (a thesis called 'rationalism'). However, we can then ask 'what's so queer about categorical reasons for action?'. One plausible answer to this is that, given a form or reasons internalism, categorical reasons necessitate a particular motivational effect. In other words, if we accept a form of reasons internalism then rationalism entails judgement internalism, which is in turn problematic for the reasons outlined above.

## 2.16 The Normative Question, Internalism, the Argument from Queerness and the Open-Question argument

We have seen above that the argument from queerness and the open question argument may depend upon judgement internalism to get their force. The argument from queerness relies upon judgement internalism either directly (if we follow Richard Joyce's reading of the argument) or indirectly if we think that it is rationalism causing moral realism problems. The revised open-question argument solves some of the problems of its unrevised ancestor – we do not end up begging the question against the naturalist, instead we point to a feature of morality that requires explanation: the conviction that clear-headed, competent speakers have that the open question really is open. But this revised argument crucially depends upon judgement internalism. Both arguments share a feature in common – they point to some aspect of moral discourse which is deeply connected with motivation, and claim that moral realism does not have the resources to explain that feature of moral discourse. The moral facts required to explain the motivational force of moral judgements would be too queer (according to the subscriber to the argument from queerness) or we could imagine a competent judge making the judgement without feeling the appropriate motivation (move **2** in the revised open-question argument).

We have also seen how Korsgaard thinks that the force of the open question argument and the argument from queerness depend upon the force of the normative question. So we have two arguments that plausibly are given their force by judgement internalism. Korsgaard thinks that what gives them both their force is the normative question. However, what it is precisely that the normative question is asking for is fairly unclear. We are looking for an interpretation of it that will allow us to get a better grasp on Korsgaard's problem with

realism. My suggestion is that the normative question gets its force from judgement internalism.

The assimilation is easiest to see if we compare Korsgaard's normative sceptic to an amoralist. Korsgaard's normative sceptic is the agent who asks 'why should I be moral?' in general, and 'but why should I do that?' of any act they judge to be a moral duty in particular. The normative question is asking us to provide a response to that kind of question, asked by someone who has fallen into doubt about the demands of morality. The amoralist is an agent who makes a genuine moral judgement but fails to feel motivated to act in accord with that judgement.

We can see the similarity between the two cases if we consider how the internalist moral realist responds to both cases - the normative sceptic and the amoralist - starting with the latter. The internalist does not concede that an amoralist is a genuine conceptual possibility. The agent who lacks motivation cannot be making a genuine moral judgement. Once someone makes a moral judgement, in effect, there is no more work to be done in explaining why they will be motivated to act in accordance with that judgement – that they are follows by conceptual necessity. Now, the normative sceptic, as presented by Korsgaard, is an agent that appears to make a genuine moral judgement, but then asks why they should act in accord with that judgement in particular, or with moral judgements in general. The internalist moral realist, it seems, will treat this kind of question in the same way as they treat the putative amoralist – by stonewalling it. Once you have made a moral judgement, the internalist moral realist will say, you will have the appropriate motivation as

a matter of conceptual necessity, so what are you asking for? You've made the right judgement, and you want to do what the judgement prescribes, so what's the big deal?

To someone who takes the possibility of normative scepticism seriously this manoeuvre will look altogether too quick. We find ourselves faced with a real problem, it seems – if the normative question has any force, and yet the realist simply ignores it. The best they can do is restate the normative fact that obtains in any particular case. Now it looks like Korsgaard's charge against the realist – that they simply ignore the normative question – is vindicated. Of course, it is open to the realist to claim that they ignore this kind of question for good reason – that their conceptual claim is true. But if you start from the opposite direction – by acknowledging the force of the normative question – you could instead doubt whether the internalist moral realist's conceptual claim is true, or whether they have the resources at their disposal to underwrite it: whether citing another fact to the amoralist will get them to see that that fact has practical significance.

This manoeuvre – of responding to both the normative sceptic and the amoralist by repeating the internalist's conceptual claim – will look equally inadequate to someone convinced of the genuine conceptual possibility of an amoralist. Far from retreating from the claim that an amoralist is possible when told the internalist's conceptual claim, instead they will simply doubt the truth of that claim.

So then, the case of the amoralist and the normative sceptic share some features – the internalist moral realist responds to both by adverting to their claim that there is a conceptual connection between moral judgement and motivation to act. This means that the

case of the amoralist is not a genuine possibility, and that the normative sceptic can be answered by simply resupplying them with the relevant moral fact – if they come to judge matters aright, they will see the pointlessness of their question. It looks, then, that what grounds the internalist moral realist's dismissive treatment of the normative question and the sceptic who asks it is the same thing that grounds their dismissive treatment of the conceptual possibility of amoralism – judgement internalism.

Two points need to be stressed here – although there are parallels between the realist's treatment of amoralism and normative scepticism, such that it's reasonable to diagnose these treatments as reflecting their commitment to internalism, what Korsgaard thinks is wrong with this realist treatment is slightly different to what the externalist thinks is wrong. I said above that taking the normative question seriously could lead you to make one of two moves against the internalist moral realist: either you could question the internalist's conceptual claim; or you could argue that moral realism doesn't have the resources available to underwrite that claim. The externalist concerned with amoralism takes the first option – they take the possibility of amoralism to militate against the internalist's conceptual claim. Korsgaard, on the other hand, would have to take the second option – it's not that internalism is false, it's that moral facts aren't suitable materials for explaining the intimate connection between moral judgement and the will.

In this, Korsgaard's normative question again parallels the argument from queerness and the revised open-question argument. The argument from queerness and the revised open question argument are not, as they stand, arguments against the internalist's conceptual claim – instead they argue that moral realism cannot explain that claim. The only resources

moral realists have available to explain the motivational force of moral judgement are moral facts and the instantiation of those facts. If what I am saying here is right then the normative question, the argument from queerness and the open question argument are all ways of making the same point – these resources aren't adequate.

The second point is that this reading of what Korsgaard's normative question is getting at may not be one that Korsgaard would share, a point I shall return to at §2.3. However, to some extent this may not be too damaging for my case. I started by acknowledging that the normative question has some force – it feels live, and seems to require some answer. If the interpretation I offer is viable, then we have secured an explanation of these seemings. In a way, my diagnosis of what the normative question is getting at is structurally similar to Korsgaard's diagnosis of what the argument from queerness and the open question argument are getting at. Both of these arguments, Korsgaard contends, are making a point. However, what underwrites the point they are getting at is the force of the normative question, and not what the authors concerned thought was involved. So, we have an explanation of why the argument from queerness and the open question argument look compelling – they are getting at something important, in a roundabout way. What I am advising is that we extend this sort of story to Korsgaard's normative question – it is getting at something important, but it is not really clear what that is supposed to be. If we read it as concerning the ability of moral realism to account for judgement internalism, then we have an explanation of why the normative question looks compelling. So, although this line of reasoning may not be welcome to Korsgaard herself, I hope readers of Korsgaard who initially find her normative question compelling might consider this interpretation illuminating.

We can now present my interpretation of Korsgaard's argument against moral realism, from the normative question, (very schematically) as follows:

(7) Moral realism is committed to judgement internalism.

(8) Moral realism does not have the resources to account for judgement internalism.

Therefore:

(9) Moral realism is false.

This is very schematic as, if I'm right above, we could fit any of the argument from queerness, the open question argument or Korsgaard's argument against realism into this framework. The normative question, like the considerations from queerness and the open-question argument, apply at premise **8**.

The important point is that we have gotten clear on what Korsgaard's complaint against realism might be. This allows us to understand how a moral realist could start to respond. They could try to deny premise **8** – perhaps there are some resources available to a moral realist that haven't been considered. For example, perhaps moral judgements have both belief like and desire like components, capable of explaining the judgement internalist's conceptual claim.[25] However, there is a simpler method available to the moral realist – they could simply deny premise **7** by claiming that there is no conceptual connection between moral judgement and motivation to act.

[25] Discussion of this possibility and a similar looking proposal called hybrid expressivism are found in chapter 4.

Such a move would not be *ad hoc*. We have already seen that the internalist's conceptual claim rules out the possibility of an amoralist. This seems implausibly strong. And if we do have good reasons to reject judgement internalism then moral realism will be impervious to not only Korsgaard's argument from the normative question, but also the revised open-question argument and the argument from queerness.

The picture would look like this – just as the externalist moral realist can accept the genuine possibility of amoralism, they can also take the normative question seriously. They can accept that the job of explaining why we should obey the demands of morality is a serious one, and attempt to give such an explanation. What they are not forced to do is what Korsgaard argues they are – ignore the normative question.

This strategy will only begin to work, however, if externalism is viable. Is there any reason to think that moral realism must be committed to internalism? If it were then the strategy I've just sketched would be unavailable and, for all I've shown, Korsgaard's argument from the normative question would have the scope and strength she claims for it.

## 2.2 Internalism vs. Externalism

So we have seen how it would be really great if we could do without internalism – it would give the moral realist a strategy against Korsgaard, Moore and Mackie. But can we go one

better, and give reasons for thinking we can? We can settle this issue via two routes – are there any internal problems with internalism, and are there any compelling reasons to ignore the externalist's hypothesis?

Michael Smith (1994, 1997) argues against externalism via two steps: first by attempting to deflate the worry about the conceptual possibility of an amoralist; and second by arguing that the externalist about motivation can only explain why the motivation of a good and strong-willed person falls in line with their judgements by turning them into a moral fetishist. Responses to Smith have mainly focussed on the second half of this strategy[26]. In the next section I will instead attack the first part of Smith's manoeuvre by arguing that the amoralist gives us *prima facie* evidence against judgement internalism, which Smith's argument does not dispel. Then I will consider his 'fetishism' argument and how the realist should respond, before finally detailing a number of residual problems beyond the conceptual possibility of an amoralist for judgement internalism. I hope to show that we have no good reason to accept internalism (the fetishism argument fails), and that internalism should be rejected.

In section 2.21 I will lay out the case against internalism based on the amoralist (and how it has traditionally been responded to). In 2.22 I will outline Smith's response to the amoralist before showing how he missteps in section 2.23.

[26] See, for example, Brink (1997), Lillehammer (1997), Miller (1996), Stratton-Lake (1999).

## 2.21 The Amoralist and the Inverted-Commas Response

So we have seen how it would be really great if we could do without internalism. But we can go one better than that – there are good reasons to suspect that internalism is false. Recall the strength of the internalist's claim – it is one of conceptual necessity. It is conceptually impossible to make a moral judgement without being motivated to act in accordance with that judgement (absent any practical irrationality).

It is at this point that the externalist steps in with their amoralist challenge. Imagine, they ask us, a practically rational agent who makes some moral judgement (e.g. that meat-eating is wrong) but who feels no motivation to act in accordance with that judgement (they feel no pull towards refraining from eating meat – we will call such a person an amoralist). Now, you might think that such people do actually exist; but remember the strength of the internalist's claim – it is one of conceptual necessity. Thus, it is compatible with externalism for there never to actually be any amoralists. But, it doesn't seem that there is anything *conceptually impossible* about such an agent existing. So, here we have the challenge to the internalist's thesis. If the amoralist is conceptually possible, then internalism is false.

One way of defending internalism against this attack is to claim that when the amoralist uses the moral terms that we do they are using them in a different way to us: in some kind

of inverted-commas[27]. So when the amoralist says that some state of affairs is good, they don't mean the same thing as we do when we say that a state of affairs is good. Instead they mean that that state of affairs is "good", where "good" has a different semantic content to 'good'. So, for instance, when the amoralist says that a state of affairs is "good" they might mean something like it is 'judged to be good by others'. They are using the same linguistic token as us (the symbol 'good'), but it possesses different meaning, depending upon whether we are an amoralist or moralist. Under this analysis, it turns out that the amoralist isn't really making any moral judgements at all, so there is as yet no counterexample to the claim that moral judgement conceptually necessitates motivation. By giving this 'inverted-commas' account of the amoralist's use of (seemingly) moral terms, we can preserve the truth of internalism.

Brink (1989) contends that this does not take the amoralist challenge seriously enough, a point which I will extend here. We can modify the example a little to one where the amoralist distinguishes between what other people judge to be right and wrong, and what they think is *really* right or wrong. The amoralist's ability to make this distinction (between what others judge to be wrong, and what they think is really right and wrong) gives us *prima facie* evidence that the proposed inverted commas account does not accurately capture the content of the amoralist's judgements. In fact, we could strengthen Brink's case even further by imagining an amoralist who has the inverted-commas account explained to them, but then rejects that as a description of their own practice. There doesn't seem to be anything conceptually impossible about these cases and if that is right then the inverted-commas account fails – we have an amoralist making judgements about what they think is

[27] This strategy was first suggested by R.M. Hare (1952) in reply to a possible attack on his claim that all moral language had imperative content – those using moral terms without imperative force were really only using them within inverted-commas.

*really* right or wrong, but yet feeling no motivation to act in accordance with those judgements. It seems as if it is possible to do this with any suggested interpretation of the amoralist's use of moral terms. If this is correct, then the inverted-commas response to the amoralist challenge fails.

## 2.22 Smith's Response to the Amoralist Challenge

Smith claims that what these latter amoralist examples are relying upon is the failure of one particular account of how the amoralist is using moral terms. The inverted-commas theorist offered one possible way in which the amoralist could be failing to make real moral judgements – by "good" they mean 'judged to be good by other people'[28], for example. It's making this further analysis of the judgements involved which causes trouble for the inverted-commas theorist, as at this point the externalist can point to an amoralist who distinguishes between judgements of *that* particular type, and the judgements they make when they use moral terms. What was important about this type of reply to the amoralist was the claim that the amoralist is not really making moral judgements at all; the inverted-commas theorist creates a hostage to fortune when they then go on to try and offer an informative account of what exactly it *is* that the amoralist is doing. Instead, Smith contends "the *very best* we can say about amoralists is that they try to make moral judgements but

[28] Another problem with this account is that 'good' features in the analysis of the amoralist's judgement. But the amoralist does not grasp the meaning of 'good' – they only understand the meaning of "good", the meaning of which we are attempting to characterise! Smith could fix this problem by making the amoralist's practice more obviously anthropological – the amoralist by "good" means 'people utter the sound [good] when called upon to make a judgement of the object's moral status', or similar.

fail." (1994, 68). To insist that the amoralist *is* making a genuine moral judgement would be to beg the question against the internalist. Smith offers the following comparison to demonstrate that this move is not *ad hoc*.

Imagine, he asks, someone who is blind from birth (and thus incapable of having visual experiences) but is able to distinguish between differently coloured objects reliably (perhaps the sensitivity of their fingers allows them to discern the different surface-reflectance properties involved). This person uses colour terms with the same extension as our colour terms – they apply, for example, the predicate 'green' to the same set of objects as us. Now, it appears to be a live issue whether this person is really making colour judgements or not. This debate, Smith contends, parallels the one between the internalist and externalist. Here it is worth quoting Smith's remarks at length:

> One side says that a subject has mastery of colour terms (moral terms), and really makes colour judgements (moral judgements), only if, under certain conditions, being in the psychological state that we express when we make colour judgements (moral judgements) entails having an appropriate visual experience (motivation). The other side denies this holding instead that the ability to use a term whose use is reliably explained by the relevant properties of objects is enough to credit her with mastery of colour terms (moral terms) and the ability really to make colour judgements (moral judgements). Having the appropriate visual experience (motivation) under appropriate conditions is an entirely contingent, and optional, extra The debate is a real one, so how are we to decide who wins?

> Imagine someone objecting that those who say that the capacity to have certain visual experiences is partially constitutive of mastery of colour terms do not take 'seriously' enough the challenge posed by people who can reliable say 'Grass is green'. 'Fire-engines are red', and so on, while yet being completely blind. Suppose the objector insists that since blind people can reliably use colour terms in this way, it just follows that they have full mastery of colour terms. Would the objection be a good one? I do not think so. For the objection simply assumes the conclusion it is supposed to be arguing for. It assumes that blind people have mastery of colour terms, something that those who think that mastery requires the capacity to have the appropriate visual experiences under the appropriate conditions deny.
>
> It seems to me that Brink's amoralist challenge is flawed in just this way. He puts a prejudicial interpretation on the amoralist's reliable use of moral terms. He assumes that the amoralist's reliable use is evidence of her mastery of those terms; assumes that being suitably motivated under the appropriate conditions is not a condition of mastery of moral terms. But those who accept the practicality requirement do not accept the account of what it is to have mastery of moral terms that makes this prejudicial interpretation of the amoralist's use of moral terms appropriate. (69-70)

So, the inverted-commas theorist got themselves in trouble when they attempted to tell us precisely what it is that the amoralist does when they say that something is 'good'. Instead, we should merely claim that they are trying to make a moral judgement and failing. This move is not *ad hoc*, because if the externalist insists that the amoralist *is* making a genuine

moral judgement, that is only because they are putting a prejudicial spin on the amoralist's use of moral terms.

## 2.23 The Strength of the Amoralist Challenge

This seems to misrepresent the strength of the amoralist challenge. Smith expects us to consider the colour case and then conclude that the externalist is giving the amoralist's use of moral terms a prejudicial interpretation. We are expected, I think, to conclude that the amoralist case leaves the two parties all square: the externalist will interpret the amoralist as making genuine moral judgements, whereas the internalist will interpret them as failing to make genuine moral judgements[29]. This, though, merely reflects their pre-standing commitment to their respective positions. The example itself doesn't carry any independent weight. To insist that it does refute internalism is to merely presuppose the truth of externalism, so will be entirely unconvincing to the internalist.

However, this seems to underestimate the force of the amoralist challenge. This is because of Smith's admission that the 'acolourist' (the person using colour terms without having visual experiences) can use colour terms with the same facility as a person with normal colour vision: "she uses colour terms with the same extension as our colour terms" and "we

[29] In parts of his (1994) Smith seems to be suggesting that the considerations he raises not only leave the externalist and internalist all square, they in fact *favour* the internalist. I can see no motivation for having this reaction, so I will restrict myself to the weaker, more obviously motivated claim that the considerations leave the matter a tie.

can even imagine, if we like, that her colour judgements are far more accurate and reliable than those made by sighted folk" (p. 69). By analogy the same should go for the amoralist – they apply 'good', 'bad', 'right', 'wrong' etc to the same cases as the moralist (someone who makes moral judgements and feels the associated motivation).

Smith thinks that the possibility of this amoralist does nothing to militate against his internalist thesis. But, it does seem that the conceptual possibility of such an agent does at least give us *prima facie* evidence that there is something amiss with the internalist's main claim. It seems reasonable to endorse the following principle:

> If a subject S reliably applies 'F' to items that fall within its extension then this is *prima facie* evidence that S grasps the concept expressed by 'F'

I suspect it would be possible to mount a good defence of this principle from any particular account of what constitutes grasping a concept.[30] To allay any fears, let me stress the weakness of the claim. The principle does not claim that speakers' use of terms are our only, or even our best evidence for whether or not their use of those terms is in part explained by their grasping a particular concept. All it claims is that, absent any confounding variables, if someone can reliably use a term (that is, apply it with regularity to the items that actually do fall under its extension) then this gives us some evidence that they grasp the concept that the term expresses. In effect, I am simply claiming that ascribing conceptual competence to the amoralist is the best explanation of their linguistic practice.

[30] For example, on a Fregean way of doing things the concept **F** will be a sense which determines an extension. Then, if a speaker applies a term to the same objects as are within the extension of that sense then this indicates that they have grasped some sense that determines that their use of the term applies to those objects. As this sense determines that the exact same objects fall under its extension as the genuine sense of 'good' then we have good grounds to conclude that they are identical, and the amoralist does really grasp the meaning of 'good'.

But of course this principle is explicitly defeasible – reliable use of a term only gives us *prima facie* evidence of concept possession. One way to defeat the evidence the reliable use of a term gives us is to tell some sort of story about how S is applying the term accurately (this would be to try to offer a *better* explanation than the one that cites the amoralist's conceptual competence). Consider the colour case again. In the analogy as Smith sets it out we are told what it is that the acolourist is doing when they are making a seemingly standard colour judgement (they are using their tactual sensitivity to pick out the surface-reflectance properties of objects). It is this alternative story that increases the plausibility of the analogue of the internalist's claim in the colour case. Absent this story about what the acolourist (or amoralist) is actually doing it seems reasonable to conclude that the possibility of an amoralist, who applies moral terms with the same facility as a moralist, gives us some evidence against the plausibility of the internalist's position.

I am not claiming that the case of the amoralist definitively proves the falsity of internalism, instead I am advancing a weaker claim. If the internalist wants to agree with the externalist that you can have an agent who applies moral terms to the same class of things as a moralist but who lacks any motivation to act in accord with those judgements, then it seems as if they have already given too much ground. The existence of such an agent would give some grounds for doubting the internalist claim. It would be different if the internalist were prepared to give us an alternative account of what precisely the amoralist *is* doing. Then we would have an analogue of the explanation given in the colour case. This explanation would give us good reason to overturn the *prima facie* evidence provided by reliable term application. However, Smith has explicitly abandoned such a strategy – on the grounds that

it creates a hostage to fortune: the externalist can merely retell their amoralist story, modified to take into account the internalist's suggestion.

So it seems as if the internalist is stuck between the horns of a dilemma. If they give us a detailed account of what exactly it is that the amoralist is doing, then they open themselves up to a revised amoralist case. If they refrain from offering such an explanation (as Smith is keen to do) then their admission that the amoralist can apply moral terms with the same facility as a moralist gives us *prima facie* evidence against the internalist position. Smith's claim that the externalist begs the question against the internalist by presenting the amoralist as a counterexample to the practicality requirement is particularly harsh. You may think that the challenge presented by the possibility of amoralism is not particularly strong, but if we are not given an explanation of what the amoralist is actually doing, then it seems perfectly legitimate for the externalist to use the case in this way.

It seems as if Smith is trying to plot a course between two, irreconcilable, positions. He does not want to assign incoherence to the amoralist case too glibly- he is at pains to reject the externalist charge that internalism doesn't take the amoralist seriously enough. On the other hand, he does not want us to take the amoralist (as the externalist presents it) as a genuine conceptual possibility. In the end he may simply be forced to insist that the case, as set out by the externalist, is simply incoherent. Whether this is an adequate response to the externalist is not my main concern here (although I am about to try and assess it in a rather sketchy manner). However, it has notably less dialectical force than Smith's original positioning of the amoralist challenge.

One cost of being forced into taking this line (that the amoralist case is simply incoherent) is that the internalist then faces an obligation they need to discharge – to explain what gave us the mistaken impression that the case was coherent. The internalist might be tempted to fall back on their claim that the amoralist is *trying* to make a moral judgement, but fails. However, absent any explication of how they are failing, to pin the blame on the amoralist's conceptual competence strikes me as *ad hoc*.

To sum up so far, I have argued that a lot can be gained from reconsidering our commitment to any form of judgement internalism. In addition, we have a positive reason to abandon it – the conceptual possibility of the amoralist. Smith attempts to deflate this argument, but I hope to have shown that the way in which he does this fails – the amoralist's facility with moral terms gives us *prima facie* evidence that they are making genuine moral judgements. Finally, I presented a sketch of a worry about Smith's description of the amoralist case. If he says too much about it, he opens himself up to a revised amoralist case (which he tacitly acknowledges in his own treatment of the inverted-commas theorist) if he says too little, then it becomes hard to see where he gets the resources to explain away our mistaken judgement about the conceivability of the amoralist.

## 2.24 The Argument from Fetishism

I have shown that the seeming conceptual possibility of an amoralist gives us *prima facie* evidence against internalism. However, if externalism has unacceptable consequences this would give us reason to revise our previous judgement – we would say that we were wrong

about the conceptual possibility of amoralism. It looked plausible to start with, but now we see that it leads to certain untenable consequences we will be forced to admit that the amoralist is not a genuine conceptual possibility and that internalism is true. With this conclusion in hand we could not offer externalist moral realism as a response to Korsgaard's concerns. So does externalism have unacceptable consequences?

Smith claims that externalism cannot give an adequate explanation of the fact that "*a change in motivation* follows reliably in the wake of a *change in moral judgement*, at least in the good and strong-willed person" (1994, 71). In fact, he argues that the only explanation available to the externalist turns the virtuous and strong-willed into moral fetishists. Suppose, Smith asks, that I engage you in a discussion about which party I am going to vote for in the next election. I start by thinking I should, and feeling motivated to, vote for the libertarian party. After our discussion, I realise that what I should do is vote for the social democrats – not because they will better enact the policies I agree with, but because our discussion has changed what I value in a political party; my moral judgement about the values of the parties has changed. In good and strong-willed people, we would expect my motivation to change accordingly – I should now want to vote for the social democrats. Why does this change reliably occur?

The internalist has a rather straightforward explanation – there is a conceptual connection between judgement and motivation, so if my judgement changes, my motivation must too.

> Moreover, and importantly, note that defenders of the requirement [the practicality requirement, which the internalist endorses] are in a position to insist that what an

> agent is thus motivated to do when she changes her moral judgement is precisely what she judges it right to do, where this is read *de re* and not *de dicto*. (73)

But what of the externalist? For them, they cannot make an appeal to the contents of moral judgement:

> The externalist says the connection between moral judgements and motivation is contingent, so he cannot say that it obtains in virtue of the contents of the moral judgements themselves. What accounts for an agent's moral motivation must then be that agent's motivational dispositions, more specifically the contents of her desires (Lillehammer, 1997 188)

What we would need to explain why there is a reliable connection between judgement and motivation in the good then would be some kind of standing desire to do what the agent judges to be right. The motivational content which explains the reliable link between judgement and motivation must be "a motivation to do the right thing, where this is now read *de dicto* and not *de re*." (Smith, 1994, 74).[31]

In other words, the internalist who endorses the practicality requirement can explain how one can change the contents of one's moral judgements, and then care non-derivatively about those contents. So I change my opinion on who the best party to vote for is and my motivation follows straightforwardly from that new judgement – I now care directly about the values of the social democrats. This is because it is a conceptual constraint on making a sincere moral judgement that I have the appropriate motivation. The externalist however has to say that my concern for the contents of my new judgements can only be derivative – that

[31] Smith originally sets out the case in terms of *de re* versus *de dicto* desires, but later uses a distinction between derived and underived desires. I stick with the original terminology.

is, following from my general desire to do the right thing (where this desire is read *de dicto*). So I don't care about social democratic values directly in themselves, only as a way of satisfying my more general desire to care about the right kind of things.

But, Smith claims, this picture cannot accommodate certain platitudes which are essential elements of the structure of moral discourse. We should care about the contents of our moral judgements directly, in a non-derivative kind of way. To only care about them because of our more general desire to do the right thing "is a fetish or moral vice, not the one and only moral virtue" (75)

> Good people care non-derivatively about honesty, the weal and woe of their children and friends, the well-being of their fellows, people getting what they deserve, justice, equality and the like, not just one thing: doing what they believe to be right, where this is read *de dicto* and not *de re*. *(*75*)*

So externalism when explaining the reliable connection between judgement and motivation in the good and strong-willed has to make recourse to a general standing desire to do the right thing, not directly as whatever that thing is in itself, but derivatively as being the thing that happens to fall under the description 'the right thing to do'. This is a type of moral fetishism, and so externalism fails to explain this reliable connection in a way that is compatible with central platitudes about the moral psychology of virtuous agents. This is because "[j]ust as it is constitutive of being a good lover that you have a direct concern for the person you love, so it is constitutive of being a morally good person that you have direct concern form what you think is right." (76). As in Bernard Williams' (1981) famous example where the advocate of impartiality ascribes to the man considering saving his drowning wife a thought too many, the externalist does this with every moral motivation of

the good person, which ends up "alienat[ing] her from the ends at which morality properly aims" (Smith, 1994 76).[32]

Lillehammer (1997) responds by 1.) rejecting the claim that caring about the right thing, where this desire is read *de dicto* is a kind of moral fetish and 2.) pointing out that the externalist is not barred from making reference to a desire to do the right thing read *de re*.

On the first point, imagine a woman (Lillehammer asks us) who feels her affection for her husband waning. Attending a party, an opportunity to pursue an affair presents itself. She realises that pursuing this relationship would be wrong, because of the effect on her husband's feelings – but right now she has no concern for her husband's feelings. Fortunately, she has a standing desire to do the right thing (where this is read *de dicto*) and refrains from being unfaithful to her husband. This demonstrates that a desire to do the right thing (where this is read *de dicto*) can play an important role in the psychology of the good and strong willed. (192)

[32] To put it another way for the externalist who explains the reliable connection between virtue and being motivated to pursue right actions there will be an extra step between noting an action's right-making features and acting – the agent will notice that an action is an instance of helping their children, say. They also accept that helping one's own children is the right thing to do. They can then conclude that that act is the right thing to do, and this judgement combined with their standing desire to do the right thing (whatever that is) leads them to perform the action. Smith thinks this is unduly fetishistic, for the virtuous person should be moved to act by the action in question being an instance of helping their children, not by it being an instance of helping your children which is an instance of a right action. The internalist does not need this last step, and it's the externalist's reliance on it which leaves them open to the charge of fetishism.

We can also imagine scenarios where it would be close to obscene to require an underived concern for the contents of a moral judgement:

> Consider next the case of the father who discovers that his son is a murderer, and who knows that if he does not go to the police the boy will get away with it, whereas if he does go to the police the boy will go to the gas-chamber. The father judges that it is right to go to the police, and does so. In this case it is not a platitude that a desire to do what is right, where this is read *de re*, is the mark of moral goodness. If what moves the father to inform on his son is a standing desire to do what is right, where this is read *de dicto*, then this could be as much of a saving grace as a moral failing. Why should it be an *a priori* demand that someone should have an underived desire to send his son to death? (192)

However, this looks unfair to Smith if we consider the objection as a way of clarifying what, for Smith, the relevant desires are directed towards. Smith's claim is that moral virtue requires caring non-derivatively about a whole range of things: honesty, justice, equality, the well-being of family and friends, and so on. So, moral virtue requires a non-derivative concern for the right-making features of actions. What it does not require is a non-derivative desire for the death of your son. Instead, you need to have a non-derivative concern for the relevant right-making features. In this case, these will be concerns to do with justice as well as the well-being of your offspring. In this way we can avoid, when offering a style of explanation Smith would class as non-fetishistic, attributing to the father a desire which looks morally repugnant.

If this is the way that Smith avoids this objection (by arguing that the relevant desires that a virtuous agent possesses should be directed towards the right-making features of action, thus avoiding the need to attribute a *de dicto* desire to do the right thing in cases where a *de re* desire would look monstrous) then this opens up space for another objection, or rather two closely related objections. First: we began by being told by Smith that externalists turn virtuous agents into fetishists by having to attribute to them a very narrow, but general, moral concern – they only care about doing the right thing (whatever that thing happens to be). Instead, being virtuous involves responding to a range of morally significant features, and caring non-derivatively for a number of particular things. So we should care about our friends' and family's well-being, equality, honesty, justice, and so-on. But it turns out that really what we should care about are the right-making features of our actions.

In the case of the father reporting his son to the police we should not attribute to the father the desire to have his son executed. Instead, he should have the desire to promote justice, or some other suitably similar desire. Now it seems as if the internalist is not too far away from fetishism themselves. The desires they need to attribute to the virtuous agent are themselves directed towards fairly abstract, and general, ends – like promoting justice, equality and so on. We can imagine someone balking at the following scenario: suppose that the father in the example above in the end decides not to report his son to the police, perhaps because he feels that considerations to do with his son's well-being outweigh the fact that justice would be served by reporting his son (it could be that the father thinks that the death penalty is wrong, so justice would not be very well-served by reporting the son, and he's convinced that his son's actions would not be repeated), and tells his son about the dilemma he went through. Although his son may be happy that he has evaded death, would he be impressed with his father's reasoning? When his father explains that he chose to not

report his son, and thus protect his life, because justice wouldn't be well served by reporting him can't his son complain that he should have acted out of love? Or, when a friend tells you that your partner is being unfaithful to you because they have a strong commitment to honesty, would it be inappropriate to respond 'What, you only told me this because you want to do the honest thing *whatever that happens to be*?'. The point is that if only caring about a right action under its aspect as 'the right thing to do' (where this is read *de dicto*) is unacceptably fetishistic, then is caring about a right action under its aspect as 'the honest thing to do' (read *de dicto*) significantly less fetishistic?

This leads directly to the second, related, objection. The first objection was that it is hard to see how replacing a commitment to doing the right thing (read *de dicto*) with a greater number of commitments directed at similarly abstract ends can halt a charge of fetishism (if one is warranted). This problem is amplified if we consider what happens when we plug in different moral theories to the case. Compare two theories of value: a hedonist claims that the only right-making feature is maximising pleasure and minimising pain. In contrast, a pluralist about value will claim that there are many different ways for an action to be made right – they can value friendship, love, benevolence, beauty etc in themselves: not merely as things that increase overall pleasure. According to the hedonist a virtuous agent is one who maximises overall pleasure, and the most effective way to do this may be for that agent to only have a concern for maximising overall pleasure (if they are quite rational and proficient at working out consequences). In contrast the pluralist will claim that being virtuous involves caring about a lot of different right-making features (whichever ones their theory of value posits). Now it seems like the hedonistic virtuous agent is closer to fetishism than the pluralist. The hedonist only cares about doing one thing – the action that maximises pleasure, whatever that is. The pluralist cares about all kinds of different values. It could be

that this tells against the hedonist theory – that just caring about maximising pleasure *is* fetishistic. However, this sort of objection would miss the target. It ignores the possibility that the hedonist theory is *true*. If it really were the case that the only thing that matters, morally speaking, is maximising pleasure then it seems hard to see how caring only about pleasure would be fetishistic. So, let's take an agent who only cares about maximising pleasure. If hedonism is true then it would be strange to describe them as fetishistic. But now it seems like a question in moral psychology (whether externalism commits us to making the virtuous fetishistic) depends upon which value theory is true. This creates a hostage of fortune for Smith – allegations of fetishism only make sense against a background value theory. The point is not that the hedonist theory might be true – it may be that the hedonist theory is implausible on independent grounds. Instead, the point is that it is unfortunate to have this kind of dependency underwriting Smith's charge of fetishism. Smith is trying to lay out an argument against externalism *in general*. As such, if the argument from fetishism depends upon externalism being coupled with a particular kind of value theory than the argument has failed in that attempt. If we can avoid the fetishism argument by merely altering our theory of value then we may do that rather than commit ourselves to internalism and all the problems *that* entails.

To summarise, Lillehammer alleges that there are cases where having a *de dicto* desire to do the right thing can either be a component of a robust moral psychology (as in the case of the wife considering betraying her husband) or can actually be a 'moral saving grace'. Smith can argue that there aren't any cases where having the *de dicto* desire is such a saving grace – the way to do this is to clarify Smith's picture. The non-fetishistic desires that the virtuous agent needs are ones directed towards the right-making features of actions. However, now the contrast between the characterisation of the virtuous agent offered by the internalist and

the one offered by the externalist starts to look less stark. A concern with realising abstract moral ends is central to both their accounts. Just as someone could complain when a person they are close to does something for them out of a desire to do the right thing, they could also complain when they do it out of a desire to increase their friend's well-being, whatever that involves (you can imagine the friend asking 'But didn't the fact that you were doing it for *me* matter?' The reply 'No, I did it because you are a friend, and I want to do whatever makes my friends happy [where this is read *de dicto*]' is hardly satisfying). The point can be put another way – if it is fetishistic to just care about doing the right thing, is just caring about doing the just thing, or the benevolent thing any less fetishistic? In addition, it also turns out that the case against externalism now depends on which theory of value is correct. A certain formulation of hedonism would identify a virtuous agent by their possession of a desire to do whatever maximises pleasure, whatever that action is. This looks, structurally, similar to Smith's fetishist. However, if this theory of value is correct, then it seems bizarre to call such an agent a fetishist. The source of this problem is the same as before – to avoid the objection that having a *de dicto* desire to do the right thing can sometimes be required Smith has to claim that the non-derivative desires of a morally virtuous agent are non-fetishistic when directed at the relevant right-making features. But this makes the argument from fetishism depend upon our account of what the right-making features are. Both objections rely on the same point – that it is hard to see how merely varying the number of right-making features can make a difference to an accusation of fetishism. The fetishistic agent (according to Smith) only cares about one type of alienatingly general and abstract thing – doing the right thing, whatever that is: that's what makes them a fetishist. The non-fetishistic agent in contrast cares about a number of general and abstract considerations. It seems implausible to suggest that it is the mere number of considerations an agent cares

about that makes caring about them alienating (and hence fetishistic) or not. This is what both objections seek to expose[33].

What of Lillehammer's second point - that Smith has not shown that the externalist cannot attribute to people desires to do the right thing, where this is read *de re*? Externalism claims, remember, that there is no conceptual connection between moral judgement and moral motivation. What this means is that it is not a mark of irrationality if my moral judgement changes and my motivational contents remain the same, but:

> Externalism is also consistent with the fact that *de re* concerns for what is right can be acquired by experience, education and reflection. I might change my previously mercenary attitude towards human life after experiencing the horrors of war and thus come to care in an underived way about other people's suffering. I might be brought to love my country after having its values inculcated in me at school. Or I may undergo a process of reflection and acquire a belief that it is right of me to perform a certain action, whereupon that belief causes a desire in me to do what I now think is right, where this is read *de re*, not *de dicto*. The externalist does not deny that moral beliefs directly cause desires to act in accordance with those beliefs. Sometimes they do and sometimes they don't. The crucial point is that it is not necessarily a mark of irrationality when they don't. (193)

[33] We also might be concerned that if Smith's argument turns on the number of sources of value a non-fetishistic should bear in mind then the argument makes pluralism about value a platitude about moral discourse rather than the substantive thesis it is.

To put it another way, it is unclear why we should think that the externalist makes virtuous agents into fetishists by elevating a vice (caring for the good only derivatively) into the "one and only moral virtue" (Smith, 1994, 75). Remember that this desire (for doing the right thing, where this is read *de dicto*) is only posited by the externalist to explain why their motivations reliably follow their moral judgements. That is, to explain what happens when the virtuous person changes their mind. Does this mean that they only care about doing the right thing derivatively in all cases? It seems unlikely – we should only think this if the externalist is committed to thinking that the only non-derivative desire that a virtuous agent has is to do the right thing (where this is read *de dicto*). If this were the case then we would have to make use of the *de dicto* desire to explain all of the virtuous agents moral actions – they are all motivated by their *de dicto* desire to do the right thing, whatever that may be, along with the belief that a particular action is the right thing to do. However, there is no reason to think that the externalist is committed to supposing that the only non-derivative desire a virtuous agent has is the desire to do the right thing, whatever that may be. Externalism is compatible with the virtuous agent having any number of non-derivative desires – what they are committed to is merely the claim that they also have a standing desire to do the right thing (where this is read *de dicto*) in order to explain what happens when their moral commitments change.

This gives us an alternative picture to the one offered by Smith. It seems as if Smith thinks that the externalist can only attribute one non-derivative desire to the virtuous agent. It is the fact that explanations of all their moral decisions have to go via this desire that makes such an agent look fetishistic – the non-derivative desire taints all the desires which depend upon it. However, this is not the picture the externalist needs. Instead, they can claim that the *de dicto* desire to do the right thing is only one non-derivative desire amongst many.

True, the other non-derivative desires will be hostages to fortune – if they are found to be incompatible with the non-derivative desire to do the right thing (read *de dicto*) then it is them that will have to change. This does not mean, though, that these desires have their non-derivative status threatened – being revisable is not equivalent to being derivative.

Perhaps Smith can claim that this picture is fetishistic enough to be unacceptable. The mere fact that your other non-derivative desires are capable of being overturned by the non-derivative desire to do the right thing is a kind of fetishism. If we consider when other non-derivative desires come into conflict with the non-derivative desire to do the right thing this charge will start to look a little strange. Conflict arises when you discover that securing one of the things you care about is incompatible with doing the right thing. The externalist claims that in the virtuous agent the non-derivative desire to do the right thing overpowers the other non-derivative desire. So, their picture entails that other non-derivative desires are capable of being overturned *when those desires are incompatible with doing the right thing*. This hardly looks fetishistic to me.

In summary we have seen how Smith attempts to undermine externalism – it can only explain the reliable connection between moral judgement and motivation in virtuous and strong-willed agents by turning them into moral fetishists. We have looked at two clusters of arguments that militate against Smith's argument from fetishism. The first attempts to conclude that having a *de dicto* desire to do the right thing can be an element in a robust moral psychology, or even necessary to avoid having a monstrous desire. The second cluster attempts to conclude that externalists are not committed to using the *de dicto* desire to do the right thing to explain all a moral agent's moral actions – it is compatible with

externalism that a virtuous agent has any number of non-derivative desires for a number of morally significant things.

I gave some reason to doubt the strength of the first type of arguments – Smith can avoid the objection that having a *de dicto* desire to do the right thing is sometimes required by clarifying what the contents or a virtuous agent's non-derivative desires should be. However, if we make the contents of these desires general and abstract enough to avoid the problem we face another – the contrast between Smith's characterisation of a fetishist and a genuinely virtuous agent starts to disappear. In any case even if this first set of considerations fails to defuse the argument from fetishism the second would be sufficient. It does not matter (for our purposes) whether internalism is committed to a mode of explanation that Smith deems fetishistic (although this would be enough to show that charges of fetishism shouldn't be taken seriously in the debate between externalists and internalists) if externalism can adopt a mode of explanation Smith endorses as non-fetishistic quite freely. If externalism is compatible with a virtuous agent having any number of non-derivative desires, and having a standing desire to do the right thing does not betray any fetishism, as I believe is the case, then Smith's argument from fetishism fails.

Where does this leave us *vis-a-vis* Korsgaard? I started to develop a line of interpretation of Korsgaard's argument against moral realism that depends for its success as an argument against moral realism on showing that moral realism is committed to judgement internalism.

We have seen that one widely discussed argument to that effect is defective. Are there any other reasons to suppose that moral realism is committed to internalism?[34]

## 2.25 van Roojen on Rational Amoralism

Mark van Roojen (2010a) has recently offered a battery of considerations designed to favour judgement internalism. His case, like Smith's, proceeds via two steps. First, he attempts to deflate the worry due to Brink's amoralist. Not, as Smith does, by attempting to show that there is something wrong with the amoralist as presented by Brink but instead by showing how judgement internalism can accommodate rational amoralism – the amoralist, as characterised by Brink, is practically rational and semantically competent which van

[34] One argument we can extract from Korsgaard that I won't go into is found in her (1997). There she seems to take issue with what we might call a *hydraulic* conception of action. What moves me, according to an externalist, is my having the right kind of belief and desire that fit together in the right kind of way to produce an action – my belief that my house is on fire together with my desire to not burn leads to me fleeing my house. The problem with this, Korsgaard contends, is it leaves the agent themselves out of the picture. There is no role for their recognition of their reasons in this picture. In effect, what we get from externalism is a *causal* explanation of why people act, but not an explanation that involves *justification*. Thus it seems that the problem with externalism and any other view that takes up the hydraulic conception of action is that they conflate causal explanation and justification (in a way that someone like Richard Rorty argues is repeated in various places in the history of philosophy (1979)). Against this we could assert that the alleged confusion is no confusion at all. Brian Leiter (2005) has made the point that philosophers like Marx, Freud, Nietzsche and Edmund Gettier debunk views in philosophy by showing that there is something wrong with their causal history (for example, Nietzsche explains the source of our moral beliefs in terms of feelings of *ressentiment*). These arguments, Leiter contends, have value – because one way for us to recognise that a view is unjustified is to see how it arose through a causal process that we think is unreliable. What this amounts to doing is denying that the genetic fallacy is always a fallacy, and thus we can learn useful things about justification by looking at causal explanations. In the moral case, we could say, on behalf of the hydraulic model, that the agent *is* included in the model, as their beliefs and desires form part of the causal story that produces their actions. This is merely to give the very broad outlines of how to develop a response to Korsgaard, but I shall not be able to pursue this line further.

Roojen accepts. Second, he makes a positive case for judgement internalism which revolves around the support judgement internalism can give to, and get from, rationalism[35] – the thesis that moral requirements are rational requirements. If rationalism were true, according to van Roojen, it would explain how morality gets its rational authority. In addition we need to accept judgement internalism in order to account for 'translation-style' thought experiments. So, if van Roojen's package works we'd have a substantive conclusion – we could explain how morality gets its special authority and the data we get from certain thought experiments, but only if we accept judgement internalism. This package might be attractive enough for the moral realist to commit themselves to judgement internalism.

I will argue that van Roojen's case fails. His attempt to accommodate rational amoralism leads him to weaken the judgement internalist's thesis in such a way that it is compatible with externalism (under one reading of one of the terms involved), and cannot fill the explanatory role that internalism is usually designed to fill. Also, the explanation offered for the rational authority of morality is shaky. I will conclude that van Roojen offers us little reason to give up externalism.

In order to explain van Roojen's position some ground needs to be cleared. I will outline, very sketchily, the elements we need to get a clear enough picture of van Roojen's account: a statement of the major theses; van Roojen's conception of relativised rationality and rightness; and a (very) brief explication of how typical responses to Frege's puzzle work. I will first lay out these elements before then explaining how they add up to a case for

[35] This being a different thesis to the one under the banner 'rationalism' as used in laying out Mackie's argument from queerness above. For the rest of this chapter I will use 'rationalism' in the way van Roojen does.

internalism. This means that the next section may seem rather disjointed and perhaps even irrelevant. Hopefully this appearance will dissolve in the section after.

## 2.26 Rationalism, Internalism, Relativised Rightness and Frege's Puzzle

van Roojen characterises rationalism as the claim that "the requirements of ethics are requirements of practical reason" (2010a, 495). Such a claim has implications for the connection between morality and motivation. If we take rationality to involve being motivated to do what is rational, and avoid the irrational, then rational agents will be motivated to perform the duties morality requires from them. It also seems to account for what we can call morality's 'rational authority' – acting morally is an end for all rational agents, as acting morally is what being a rational agent involves. However, rationalism also seems to entail that those who act immorally are practically irrational. This claim comes under pressure when we consider two types of cases – those distant from us in terms of time, space or culture. They "often have a divergent conception of what morality requires. If we are right about what morality requires then they are wrong. Yet it seems unfair to accuse them of irrationality as opposed to some other sort of mistake." (495). This is because they may have had no (easy) way of knowing what morality requires. The second type of case is the agent who claims to be unmoved by a moral judgement (Brink's amoralist). Such agents don't always seem (as Brink contends and van Roojen agrees) irrational to us.

van Roojen intends to show how these putative counterexamples to rationalism (cases where those who aren't motivated by moral requirements or judgements) can actually be accommodated by it, when rationalism is properly construed. Also, rationalism can explain two plausible forms of internalism, with the result that "a plausible internalism and a well-formulated rationalism are mutually supporting theories" (496).

What are the plausible forms of internalism? First we have existence internalism, which is the following thesis

**EX-INT**: Having a moral obligation to φ is necessarily a reason to φ[36]

This thesis connects moral obligations with reasons for action. It is supported by the fact that when someone asks for a reason to perform a particular act "it is appropriate and not obtuse to explain that the action is morally right... No further answer to the why question would normally appear to be needed" (498). One problem immediately looms – some rational agents won't be satisfied when you explain to them that some action is right: they may acknowledge that it is right, but ask 'What is that to do with me? What reason do *I* have to do it?'. van Roojen rightly points out that this does not have to be too troubling – just because someone is capable of ignoring a reason, or questioning is its existence, doesn't mean it doesn't exist. If they believed **EX-INT** then they would believe they had a reason: after all, they have acknowledged that the action is morally right and if they believed **EX-INT** they would see that they therefore had a reason to perform it. However, the problem can be reiterated – the person unconvinced by the answer to the question 'What reason do I have?' that cites the moral rightness of the action is a rational agent. If they are rational,

[36] van Roojen also offers a formulation in terms of propositions: true moral propositions necessarily give us a reason to act in the way they commend. These niceties need not bother us here.

they should respond to the reasons they have. So the fact that they are not responding to the reason shows that they do not have one, and thus **EX-INT** is false. Part of van Roojen's ambition is to respond to this kind of worry.

**EX-INT** is straightforwardly entailed by rationalism – if moral requirements are requirements of practical rationality then they will necessarily give us reasons (at least, as is plausible, rational requirements are reason-giving).

The other plausible internalism is a version of judgement internalism. As we have seen already, judgement internalism connects moral judgements with motivation. The formulation van Roojen favours (borrowed, with modifications (inspired by Korsgaard, 1986), from Jamie Dreier, 1990) is:

> **JUD-INT**: If an agent judges that it is right to φ in circumstances C then normally she is motivated to φ in C, or she is practically irrational.[37]

**JUD-INT** posits a connection between mere moral judgement and motivation, so it admits cases that cause problems for linking it up to rationalism – where an agent makes a false moral judgement they will be motivated (according to **JUD-INT**) to act in accord with it, whereas the combination of rationalism and **EX-INT** makes that look irrational (they should be motivated to act the other way, if they were responding to the reasons they had properly). Also **JUD-INT** seems to be undermined by cases like Brink's amoralist and

[37] I have altered van Roojen's statement of the thesis a little to make it closer to the practicality requirement (see 2.13) offered by Smith – the differences are ones that van Roojen endorses in the text. I have also changed van Roojen's 'believes' to 'judges' – simply to keep the discussion neutral between cognitivists and non-cognitivists: it's possible for a non-cognitivist to share the conclusions van Roojen offers if they are willing to construe accepting the requirements of practical reason as involving some suitable non-cognitive attitude.

Joyce's agent of 'pure evil'. Both characters, remember, fail to be appropriately motivated by their seemingly sincere and genuine moral judgements. van Roojen hopes he has the resources to deal with these problems too.

The next element in our picture is relativised rationality/rightness. One way to get a grip on this notion is to think about the familiar distinction between objective and subjective oughts. This is roughly the distinction between what you should do given full information and time to engage in reasoning, and what you should do given the information and reasoning capacities you have. A toy, so-called 'mine case', will illustrate the point. Suppose you are an engineer working at a mine which is flooding. The mine has two shafts, A and B. You know a lot about this mine and the implications of its flooding but one thing you do not know is which shaft the team of ten miners employed by the mine is working in today. They are either all in A, or all in B. Now, you could block the bottom of one of these shafts, stopping the water getting into it. If you do this, the water will be diverted to the other shaft, filling it completely. If you block the shaft with the miners inside you would save all their lives. If you block the wrong shaft, though, all ten miners will drown. If you do nothing (block neither shaft) then water will fill up both shafts a little, but only enough to drown the miner at the bottom of the shaft – it won't reach high enough to drown the other nine. What should you do? Block A, block B, or block neither?

We can distinguish between two relevant oughts: what you objectively ought to do, and what you subjectively ought to do. In some sense, you ought to block the shaft with the miners in it – flooding the other and saving all ten miners. This is what you objectively ought to do. However, given your situation, it would be ridiculous to hold you to this

standard – you don't know which shaft the miners are in, so there is no way you can reliably pick out the right shaft to block: you'd just be taking a wild guess. What you subjectively ought to do is block neither shaft: that is the safest course for you to take, given the information you have. We could say that that action has the highest expected utility, though that would be to commit ourselves to a type of consequentialism, whereas the distinction between objective and subjective oughts should be recognised by any moral theorist.[38]

van Roojen notes that a simple binary division between objective and subjective oughts is too coarse-grained. This is demonstrated by the way I have talked about the cases above: I introduced the notion of a subjective ought by saying it was what you ought to do given the information you have. But then I moved to talking about the information you have available to you. These are two different notions – there could be readily available evidence that you could pay attention to, but which you glibly ignore – meaning it is not part of the information on which you base your decision. In fact we can make various grades of distinction, between: what you ought to do given the information you have; what you

[38] It may look like what is bearing weight here is not a distinction between subjective and objective oughts, but instead an application of the principle that ought implies can. However, the relevance of ought-implies-can is not straightforward. It is within your power to do that which is objectively required – you can physically block the right shaft. It's just that you'd be taking a wild guess at which shaft is the right one. It could be that ought implies can is relevant because in another sense of 'can' you can't block the right shaft – you are not in an epistemic position to pick the right one. But, even if ought implies can is needed to explain why this distinction holds, it doesn't mean that the distinction isn't real. To get ought implies can into the picture we need to admit that what information we have available to us plays a role in determining what we should do, subjectively speaking, and this is all we need to make the distinction between objective and subjective oughts. It seems like in this case we can square two seemingly contradictory judgements: if someone blocked A, guessing (rightly as it happens) that the miners are there, then we could say that they did the right thing (at least in the objective sense) whilst heaping blame on their shoulders: a simple story about how this could happen is to say that certain judgements about rightness and wrongness can track the objective oughts involved, while judgements to do with blameworthiness track the relevant subjective oughts (of course, in any actual case how these two types of ought interact with our judgements is likely to be quite complicated).

should do given the information easily available to you; what you should do given the information that you can, in principle, access; and so on. Also relevant here are the reasoning and evidence gathering powers of the agent concerned – we could distinguish between what you should do given the information you have, and what you should do given the information you have plus a bit more cognitive effort put into working out what follows from the information you have.

This point even applies to *a priori* reasoning. You may be in a situation in which you could work out something morally relevant if you engaged in some *a priori* reasoning. However, it's not generally true that you are obligated to work out what is entailed by what you believe: such an effort would be a waste of time, given that there is likely to be a number of trivial, true, propositions entailed by what you believe. It wouldn't even be worth your while if you just worked out the non-trivial implications – there is no guarantee you will have any use for this new knowledge and besides, don't you have better things to do? So you could find yourself in a situation where you fail to believe something *a priori* equivalent or entailed by something you believe yet where you are still rational – if you haven't actually gone through the relevant deductions yourself.[39]

We can see how these considerations play out in our mining case. Objectively (what you should do given full information, sufficient time for reasoning and sufficient cognitive capacities) you should block the right shaft. Then there are a whole raft of subjective oughts. What you should do given the information you have. But perhaps you could work

[39] This point is well brought out in Fred Dretske and John Hawthorne's (2004) discussion of epistemic closure principles. What this discussion reveals is how much is involved in working out what follows from the information you believe, even in cases dealing with direct entailment.

out which mine shaft the miners are in – then there will be an ought corresponding to what you ought to do given the information available. Suppose that the relevant test would take too much time to carry out – then we can identify another subjective ought: what you should do given the information available and the information you could gather given the time you have available. Or perhaps working out the right shaft involves some *a priori* reasoning – we could have oughts corresponding to what you should do given the cognitive capacities for reasoning you have, and another for what you should do given the cognitive capacities you should have made the effort to develop. Or perhaps you know that if you start down the path of *a priori* reasoning you will get distracted by maths problems, run out of time, panic, and then make a stupid decision.

The point is that the distinction between objective and subjective oughts is too coarse-grained. We need a far greater range of subjective oughts, corresponding to different dimensions to do with our reasoning powers and the information we have. In addition, as sometimes what information we can access and what deductions we have made will depend upon historical factors – the experiences we have been through – these factors can act as another dimension to distinguish yet more types of subjective ought. What it makes sense to do in a given situation depends upon the agent's personal history, what information they have available, what time they have available and their reasoning powers. We also need to recognise that there are second order rational requirements – requirements to do with reasonable evidence gathering, which again will be conditional upon the agent's personal history, mental capacities and so on.

We are not in the business, though, of constructing a taxonomy of subjective oughts. What is relevant for our purposes is to recognise that rationality – what it makes sense to do – can be relativised to a whole gamut of factors. If van Roojen is right about the close connection between rationality and moral rightness, then the same will be true for rightness.

The final element we need is a grasp of Frege's puzzle and traditional responses to it. Frege noted that identities involving co-referring expressions can have cognitive significance. Being told that a = a does not seem to us to be any kind of cognitive advance. But being told that a = b does seem cognitively significant (at least sometimes), even though 'a' and 'b' refer to the same object. So, being told that 'The Morning star is the Morning star' doesn't hold any cognitive significance; being told 'The Morning Star is the Evening Star' does – it's a discovery that took some empirical work, after all. Another phenomenon which points towards this feature of identities involving co-referring expressions is the fact that it is rational to doubt the identity. In fact, you could believe contradictory propositions involving the co-referring expressions: you might believe that you went to school with Robert Zimmerman, whilst believing that you didn't go to school with Bob Dylan, even though both names designate the same object. How is this possible?

There are two general pictures van Roojen is interested in: Millians hold that the meanings of referring expressions are the referents those expressions refer to. So the sentences 'I went to school with Bob Dylan' and 'I went to school with Robert Zimmerman' have the same meaning as the constituents of those sentences have the same referents and thus the same meaning. If someone were to accept the second but deny the first then they would have

contradictory beliefs. But, for the Millian, this will be a fairly widespread and benign phenomenon[40].

In contrast, the Fregean will argue that 'Bob Dylan' and 'Robert Zimmerman' present the same referent under different modes of presentation. They have different senses which play a role in determining the meaning of the sentences involving these expressions. So someone who accepted the truth of 'I went to school with Robert Zimmerman' but denied the truth of 'I went to school with Bob Dylan' would not, necessarily, have inconsistent beliefs. Their beliefs would have different contents. However, given that 'Robert Zimmerman' and 'Bob Dylan' refer to the same object, one of their beliefs would be false. There is nothing necessarily irrational about having false beliefs. So on the Fregean account the rationality of wondering whether, and the cognitive significance of learning that, the Morning Star = the Evening Star is preserved and explained.

Now we have all the materials necessary – a statement of the relevant theses, an account of relativised rationality/rightness and a brief sketch of responses to Frege's puzzle – to explicate van Roojen's defence of rationalism and the two internalisms.

[40] Or they could attempt a 'guise-theoretic' explanation of the phenomena. See Sainsbury and Tye (2011).

## 2.27 Internalism and Rational Amoralism

Remember the main problems that the package of rationalism, **EX-INT** and **JUD-INT** had: there seem to be cases where people rationally ignore what is morally required – either because they are not in a position to know what is really morally required and instead are moved by a false moral theory, or because they are unmoved by their sincere moral judgements (as in the case of Brink's amoralist). van Roojen wants to apply the machinery just detailed to solve these problems and show how internalism is compatible with rational amoralism and immoralism.

First we can use the ideas of relativised rightness and relativised rationality to explain the first type of case, rational immoralism, where an agent is moved to do the wrong thing when they believe it right (because they are not in a position to know what is really morally required). van Roojen sets out how a sprinkling of relativised rationality can help here, and is worth quoting at length:

> The general idea is to account for various kinds of rational immorality by noting that judgements of irrationality are usually or often made relative to one of the subjective senses of rationality. People who do what is objectively wrong will not be counted as irrational in one good sense so long as what they did made sense relative to the information that they have. Thus there is a sense in which those who do what is objectively wrong can still be rational though in one of the subjective senses.

> One sort of rational immorality which a rationalist should have no trouble admitting involves actions which are rational because the agent lacks certain empirical information which would, if available, have changed what made sense to do. Clearly such agents are not subjectively irrational; they are doing what makes sense given the evidence they have. But this result is compatible with the chosen action being irrational relative to fuller information that the agent might have possessed. By equating what is right with what is objectively rational in light of full information, we can truly say of such cases that the agent did something objectively morally wrong, but rational given what she knew.
>
> (512)

Given that agents rarely (if ever) have full information it's easy to explain why we feel that there is a sense in which those not in a place to know what is genuinely morally required are nevertheless decent people – they did what made sense to them, given what they knew at the time. We have an explanation of why, in cases like the mining case, our intuitions about the wrongness of an action and the blameworthiness of the agent can diverge.

We can extend this solution to the problem of squaring **EX-INT** with our intuitions about those within the grip of a false moral theory by involving the other dimensions along which we relativise rationality. For example, knowing what is objectively required could be *a priori* accessible from information you actually possess. However, you might not be blameworthy for not acting on what is *a priori* entailed by what you know if you haven't gone through the relevant deductions. Though there will be another sense of rationality (rationality relative to what you know and what is entailed by what you know) under which what you do is irrational. Another dimension we can relativise rationality along is according

to requirements on evidence gathering – you could be subjectively rational relative to the information you have, yet subjectively irrational relative to the information available to you if you don't make a reasonable effort to gather relevant information.

It's van Roojen's contention that a defender of rationalism should make use of all of these notions of rationality when they posit a connection between the requirements of morality and reasons to act. They just need to be careful about which notion they are using at any particular time. Thus they can build a connection between subjective rationality and subjective rightness – the agents who are rationally immoral still did what made sense to them given their position, so there is no threat to the connection between rationality and morality here.

There is one worry that immediately looms: I began by claiming that one of the attractions of the package that van Roojen offers is that it explains the rational authority of morality. You might gloss the rational authority of morality as the claim that it always makes sense (or is always a requirement of rationality) to do what is objectively right. This claim is compatible with what we have said so far in most cases – for some agents it makes sense to intend to do what is objectively morally required, as this is one of the aims they have. But there are cases that cause problems. van Roojen offers what we can call the ice-cream case modelled on Smith's ill-tempered squash player (Smith 1995, 109-31; van Roojen 2010a, 515). Imagine I am aware that I have a disposition to a certain type of irrationality – that I am weak-willed in the face of ice-cream. I am also attempting to lose weight, so it makes sense for me to avoid excessive quantities of ice-cream. Small quantities would be just fine, and in fact might even give me the psychological boost required to continue with my

otherwise dreary diet of porridge (with salt) and sugar-free Irn-Bru. However, I cannot manage small quantities – if I have any ice-cream in the freezer I will, in a moment of weakness, scoff the lot. Thus it makes sense for me to not have any ice-cream in my freezer.

In this case, then, there is a divergence between what is rationally required subjectively speaking and what is rationally required objectively speaking. Given that I have a disposition to weakness of will when confronted with ice-cream I should not have ice-cream in my freezer; what makes sense given my position is to refrain from having any – this is what is subjectively rational. However, if I were more rational I wouldn't have this disposition to weakness of will in the face of ice-cream. Stocking ice-cream and consuming small quantities of it would further my aims better. So, what is objectively required is that I have ice-cream in my freezer (or, it is at least rationally permissible).

If, then, we can have cases where what is objectively rationally required can conflict with what is subjectively rationally required then we seem to have a threat to the rational authority of morality: it won't always make sense to do what is objectively rationally required in cases where intending to do so would lead to an action worse than the one I can perform given my actual dispositions.

van Roojen, after introducing these types of cases, argues they are no threat to the rational authority of morality:

> Rationalism continues to identify objective moral rightness with objective rationality and to recommend this as a rational end for agents, except in special

> cases [like the ice-cream case]. It just recognises that sensible ways to aim at that goal are partly a function of facts about the agent's subjective situation. An agent is only rarely in a position that requires choosing between doing what makes sense to do in light of her evidence, time and so on and doing what she thinks would make sense to do if she had full information. In normal cases these questions collapse into one another from the first-person agential perspective; the agent is trying to do what she thinks is objectively right *by* doing what her subjective situation suggests it makes sense to do given that goal and her evidence, time, and so on.
>
> So while the abnormal cases are important because they block us from saying that it is always subjectively rational to intend to do what is objectively right, they are unusual and exceptional. In an overwhelming majority of situations subjectively rational agents will and should intend to do what is objectively right. And this gives objective rightness its rational authority and grounds rational criticism of agents who don't meet this requirement.

Van Roojen seems to think that because the cases where what we should intend differs from what is objectively right are few and far between the truth of rationalism stands. It being typically the case that it makes sense (is rational) to do what is objectively right is enough to ground the rational authority of morality.

I think this won't do. The problem is not that there might be a large number of cases where it does not make sense to do what is objectively right. Van Roojen may be right that such cases are rare (although a scenario like the ice-cream case doesn't seem preposterous enough to guarantee a low rate of incidence). The problem is that the way van Roojen

allows there to be such cases robs the combination of rationalism and **EX-INT** of any substantive content. The element in the ice-cream case that allows what it makes sense to do to diverge from what is objectively rational is one of my motivational dispositions – when I see ice-cream, my desire to eat it overcomes my other aims. It's this motivational disposition which means I shouldn't keep any ice-cream in my freezer. But, if we are allowed to include pre-existing motivational dispositions in our specification of an agent's subjective position then we run into trouble. For now what we are saying is that what it makes sense to do, and what reasons we have that (if rational) we will be motivated to respond to in the right way is a function (in part) of what motivational states we happen to have. But this is just to say that we will be motivated by our motivational states. This is a trivial claim that is compatible with externalism about motivation. This is what I mean when I say that the case for the rational authority of morality offered by van Roojen's package is shaky.

van Roojen could respond that we cannot allow *all* motivational dispositions into our specification of the agent's subjective position. Only dispositions due to weakness of will, or other condition which threatens our practical rationality is allowed into the picture. Such a reply depends upon showing that there is a principled distinction available between motivational states owing to weakness of will, and those which are not connected to weakness of will in the right kind of way. The relevant motivational disposition, which causes me to eat lots of ice-cream when I have it, doesn't seem that different from other motivational dispositions. However, we can allow that this response is viable as I will argue that there are more serious problems accounting for **JUD-INT**.

van Roojen admits that the notion of relativised rationality/rightness will not be able to explain all cases of rational wrong-doing. Brink's amoralist, for example, has no problem admitting that they are morally required to do something. Instead they simply lack motivation to act in accord with their (seemingly) sincere moral judgements. We have seen how this type of amoralism can be used to press a case against internalism: the amoralist doesn't seem practically irrational, yet is unmoved by their sincere moral judgements. Smith responds by arguing that the amoralist's facility with moral terms is no indication that they genuinely grasp moral concepts. However, I argued that the amoralist's facility with moral terms does give us *prima facie* evidence that the amoralist is conceptually competent. This is a conclusion van Roojen endorses – it is possible for an amoralist to make a genuine moral judgement, as it is possible to be competent with moral concepts without having the appropriate motivation. He contends though that this conclusion is compatible with the plausible formulation of judgement internalism – **JUD-INT**.

According to van Roojen, the belief that an action is morally required is '*a priori* equivalent'[41] to the belief that an action is rationally required. If we assume that it is irrational to act in a way that one thinks is irrational then this should mean that rational amoralism is impossible – if you believe something is obligatory, you will see that it is rationally required, and if you are rational you should be moved by that.

However, it is possible for a rational person to be unmoved by their judgement that something is right if they are not in a position to recognise the identity between the property

[41] This is *a priori* presumably because the truth of rationalism is *a priori.* If the amoralist did a bit more armchair philosophy, then they could work out what they are rationally required to do from what they morally ought to do.

of rightness and the property of being rational, where the failure to recognise the identity is not due to any rational failing. And this is where the story about Frege's puzzle comes in: the solutions to Frege's puzzle show how it can be rational, even when conceptually competent, to doubt whether a = b even where 'a' and 'b' happen to be co-referential. The Fregean says that the two expressions have different senses, and learning that they co-refer is a substantial discovery. The Millian insists that anyone who doesn't believe the identity has contradictory beliefs. But, for them, having such contradictory beliefs does not betray any irrationality or conceptual incompetence.

So, then, we have an explanation of how the amoralist goes wrong – they don't believe that the property of rightness is identical to the property of rationality, even though it is. This does not betray any irrationality though, as it's easy to see how they could think this. According to van Roojen's Fregeans 'right' and 'rational' will have different senses, and hence different meanings, even though they designate the same property. Thus you can be conceptually competent with the two concepts – grasp their meanings – without this guaranteeing that you will believe the relevant identity. For the Millian 'right' and 'rational' mean the same thing so if you didn't believe that they are identical you would be making a mistake, but this does not betray any irrationality.

It is a consequence of this picture that if the amoralist did a bit of philosophy and came to believe the truth of these theses we should expect to see a change in them. Either they would start to feel motivated by their moral judgements, or they would give them up. If this picture looks strange, then this tells against the case van Roojen makes. But there are other, more significant problems looming.

van Roojen considers the possibility of all, or most of us, being ignorant of the relevant identity. Then it would not be true that "If an agent judges that it is right to φ in circumstances C then normally she is motivated to φ in C, or she is practically irrational." (**JUD-INT**). Normally an agent would not realise that being right is the same property as being rational, and so could very easily fail to feel motivated (I mean, we could interpret 'normally' as 'if they are up to date with the philosophical literature on internalism', taking metaethicists as the relevant comparison class. Such an interpretation would be stretching the meaning of 'normally' far, far beyond breaking point.). Thus **JUD-INT** would be false in such a scenario (a sad fate for an allegedly conceptual truth).

van Roojen argues that it is impossible for all of us to be unaware of the relevant identity, in a way that falsifies **JUD-INT**, as the coherence of ascribing conceptual competence with moral concepts to the amoralist (someone who lacks the appropriate motivation) depends upon a background of moral agents who *do* feel appropriately motivated by their moral judgements. How does this idea work?

van Roojen makes an analogy with the cases taken by Tyler Burge (1979) to support social externalism about content. In our speech community 'arthritis' refers to a painful disease in the joints. If someone believes 'I have arthritis in my thigh' then they have a false belief. This is because the contribution a term like 'arthritis' makes to the semantic value of a sentence is fixed by the practices of the relevant experts in our speech community – in this case most likely medical practitioners. However, if the person who believed 'I have arthritis in my thigh' belonged to a different speech community, one where the practices of the

relevant experts made it the case that 'arthritis' means a painful disease of the joints *or the thigh*, then they would be expressing a different thought. We can tell this is the case because their belief would then be true – the proposition that their propositional attitude is directed towards would be different, meaning that the sentence expressing that propositional attitude differs in truth value. In our community, someone who says 'I have arthritis in my thigh' when they have a pain in the thigh expresses a false belief. In the other community, with different linguistic practices, they express a different, true, belief. Cases like this are supposed to demonstrate the social externalist's thesis that social context plays a role in determining the contents of thoughts and the sentences that expresses those thoughts.

What does this show us about the moral case? Well, in both cases (the amoralist and the person misusing 'arthritis') "we are willing to attribute a thought the truth conditions of which would seem to entail that the speaker is expressing something ruled out by the correct analysis of the terms used to express the belief." But "in each case we are willing to do so against a background in which most competent speakers would not avow those attitudes... If the amoralist were isolated from communities in which the term 'right' was used to commend we would not have attributed a thought about rightness to her, just as we would not have attributed a thought about arthritis to the medically ignorant patient in a community where doctors did not use the term 'arthritis' to pick out exclusively a disease of the joints." (van Roojen, 2010a 519).

This analogy shares similarities with the cases Timothy Williamson (2007) uses in arguing against the existence of a certain type of analytic truth. Williamson's target is the thesis that there are some truths that competent speakers must assent to, merely in virtue of

understanding the terms involved, or to put it another way, merely in virtue of being competent with the concepts expressed by the terms involved. He uses as an example of a putative analytic truth the sentence 'Every vixen is a vixen' (if anything was going to be an analytic truth, this looks like a fairly safe candidate). Then he introduces two competent English speakers, Peter and Stephen. Peter takes the universal quantifier to be existentially committing – he takes the truth of 'Every F is a G' to depend upon the truth of 'There is at least one F'. In addition, Peter doubt's whether there really are any vixens – he has consulted various websites, and has concluded that the existence of foxes is a conspiracy masterminded by the government to encourage bolshie types to protest against fox-hunting rather than engage in more effective revolutionary action. Thus he denies the truth of 'There is at least one vixen' and hence, given his deviant interpretation of the universal quantifier, also denies the truth of 'Every vixen is a vixen.' Thus here we have a case of a competent English speaker who denies the truth of the supposed analytic truth 'Every vixen is a vixen.' On what grounds can we say they are competent? Well, it looks like there isn't anything about Peter's grasp of the concept VIXEN that leads to Peter's denial of the sentence 'Every vixen is a vixen.' Instead he merely has a wacky view about how the universal quantifier works (although a view advocated by some competent philosophers) and some dud empirical information (it's not really true that all the evidence for the existence of foxes has been manufactured by the government; although it could have been, for all our grasp of the concept VIXEN has to say on the matter). Williamson's point is that enough competence in other uses of a concept can convince us that that user grasps the concept involved, even if there are some cases where they go wrong. They count as a member of a speech community in virtue of the fact that they tend to get things right, making it possible for them to be credited with conceptual competence even where they personally diverge from that speech community. Thus we should not hesitate to attribute to Peter competence with the concept VIXEN even though he denies the truth of 'Every vixen is a vixen.' Thus 'Every vixen is a

vixen' cannot be an analytic truth in the relevant sense – it is not a sentence the understanding of which guarantees assent. Stephen also denies the truth of 'Every vixen is a vixen', but because of a deviant view on vagueness[42]. (Williamson 2007, ch. 4).

What Williamson's cases demonstrate is how it is possible to attribute to a speaker conceptual competence even where they make false judgements about the applications of those concepts – they count as being competent users of the concept by showing enough competence in other uses of the concept to entitle them to membership of a speech community that does get things right, by and large. Competence, for Williamson, is thus holistic.

Now we can see the reason why van Roojen formulates **JUD-INT** as he does. **JUD-INT** tells us, remember, that "If an agent judges that it is right to φ in circumstances C then normally she is motivated to φ in C, or she is practically irrational." The reason for the 'normally' constraint should now be clear – it is possible for someone (like the amoralist) to make a genuine moral judgement without feeling motivated to act by that judgement, but only against a background of a community who typically do feel motivated by their moral judgements:

> [W]e have a certain sort of necessary connection between the attitudes of normal speakers in a community but of a sort that does not require all members of that community share the attitudes. The explanation is that the designatum of a speaker's

[42] Although see David Chalmers' 2010 for theses closely related to the analyticity thesis which Williamson attacks which he argues are untouched by Williamson's arguments, and Jonathan Schaffer's 2010 for a similar reply. We will also be returning to these themes when we come to discussion of Stephen Finlay's analytic naturalism.

> terms can depend on the practices of the community in which she is a member, and the content of her thoughts expressed using those terms can depend on the same facts about the same community. She may flout the norms of her community and yet harbour thoughts which are partly constituted by the very norms she flouts.
>
> (van Roojen, 2010a 519)

However, all that has been shown so far is how it is possible to credit someone with competence with moral concepts even when their judgements are not accompanied by the motivation which is normally necessitated by making those judgements, not that judgement internalism is true. We have an argument for the compatibility of **JUD-INT** and Brink's amoralist, but no positive case for **JUD-INT**. Here the comparison with Williamson's cases is apposite – in those cases we have strong, independent ground to reject the interpretations offered by Peter and Stephen: most would agree that interpreting the universal quantifier as existentially committing is a bit silly. In addition we are offered a story about what the deviant speaker *is* doing when making his deviant judgement. What independent grounds do we have for supposing that **JUD-INT** is true?

Well, van Roojen claims that "the actions and practices of most normal speakers in treating rightness as sufficient for rationalizing and justifying an action make it the case that 'rightness' designates the same property as 'practically rationally permitted'." (519-20). However, this will not be sufficient for van Roojen's case – the externalist can agree that the actions and practices of most normal speakers treat rightness as sufficient for rationalising and justifying an action. They just will not agree that this makes it the case that 'rightness' designates the same property as 'practically rationally permitted'. Instead they

will claim that most normal speakers can take it for granted that they and their fellows share an interest in doing the right thing – being psychologically normal they will care about harm done to other people and so on. Another way to put the point is that an externalist can endorse the claim that "If an agent judges that it is right to φ in circumstances C then normally she is motivated to φ in C"; it is just that they will not construe this as a conceptual constraint on making a moral judgement, and will read 'normally' as being related to something like psychological or statistical normality. The point is that we have nothing like the strong reasons we have in the Williamson case to claim that it is the amoralist who is getting things wrong, as a matter of conceptual incompetence.

So what does give us reason to accept **JUD-INT**? Van Roojen places a lot of weight on the results of thought experiments to do with the translation of seemingly moral terms.[43] How do these thought experiments proceed? Richard Joyce lays out an example succinctly enough to be worth quoting at length:

> Suppose we have undertaken a radical translation of the language of an alien community very much like our own, and we have translated nearly all of it to our satisfaction, except for a few normative expressions. They have some words that operate rather like our moral terms 'good,' 'obligatory,' etc. (call their words 'schmood' and 'schmobligatory,' etc.). If something is schmood then it is thought probably to promote or sustain alien well-being. Schmood acts are expressive of concern and respect. Schmood things are considered important. People considered schmood are praised, an absence of schmoodness is disciplined. And so on. Yet we

[43] In fact, in 2010b in reply to Russ Shaffer-Landau van Roojen appears to accept that it is mainly the considerations to do with translation of seemingly moral terms that is driving the case for **JUD-INT**.

> find that this population has an aberrant twist: when someone is considered to have judged an action φ to be schmood, this is not considered to have a bearing on whether that person is motivated to see φ done. The agent who has convinced us that he sincerely judges φ to be schmood (and judges no other available action to be schmood), and yet, calmly and with no explanation, feels utterly unmotivated in favor of φ, raises no eyebrows, produces no puzzlement in this society. To judge something schmood, in short, is not necessarily to be in favor of it.
>
> (Joyce, 2001, 26-7)

Joyce argues that we would be hesitant to translate 'schmood' as 'good'. There is something weird about the prospect. This hesitancy that we feel in translating the alien's 'schmood' as 'good' shows that our 'good' is intimately connected with motivation in the way that the alien's 'schmood' as it stands.

As it stands, this line of reasoning will not be convincing to the externalist as they have a ready explanation of our hesitancy to translate 'schmood' as 'good.' Something important is left underspecified by the case. The externalist holds that as a matter of psychological fact most moral agents are motivated by their moral judgements. We are not informed whether the alien community are psychologically similar to us. It would seem that they are, given that most of their practices involving schmoodness involve treating schmoodness as we do – it is connected up with judgements to do with human well-being, and practices involving praise and discipline in ways similar to goodness. So there is something strange about the fact that they lack the motivational element usually connected with human judgements of goodness. The case, as presented, pulls us in two different directions – the alien speakers are both psychologically like us, and yet psychologically unlike us. In fact, we may doubt

whether practices like punishing the unschmood and praising the schmood could develop, absent the typical response goodness that human beings exhibit. Perhaps this strangeness explains our reticence to perform the translation.

Now it could be that the more sophisticated translation thought-experiments offered by Dreier (1990) and Horgan and Timmons (1992) and mentioned by van Roojen alleviate this sort of worry. I won't seek to press this point directly further, as there is a more damning way of getting at the nub of the problem for the internalist.

We can see this if we remind ourselves of the internalist's ambitions. Recall that in the discussion of fetishism it was apparent that one alleged advantage of internalism was that we can give an explanation of the motivational push of moral judgements merely in terms of their contents. Because feeling appropriately motivated is a conceptual constraint on making a judgement with moral content we do not have to cite anything other than that content to explain why moral judgements have the motivational import that they do. However, the lesson of van Roojen's discussion of social externalism is that the amoralist *is* conceptually proficient. They can make a genuine judgement with moral content without feeling motivated to act in accord with that judgement. So there must be something extra that moralists have and that the amoralist lacks that explains why the former are motivated and the latter are not. The most obvious difference between them is that the amoralist lacks the right motivational state. But if this is the relevant difference between moralists and amoralist **JUD-INT** is again robbed of any substantive content – the externalist can agree that moral agents are motivated to act in accord with their moral judgements when they have, as a matter of psychological fact, the appropriate motivational element in their

psychology. They just don't think that lacking that element makes the judgement the amoralist makes a non-moral judgement. And van Roojen agrees with this claim.

van Roojen could object that this misses the point – although any individual can proficiently apply moral concepts and make a genuine moral judgement without feeling a corresponding motivation there is a conceptual link between moral judgement and motivation at a *societal level* that underwrites the amoralist's deviant moral judgements. But again the objection rears its head – there must be some difference between the amoralist and the moralist that explains this difference in motivation. If the difference is that amoralists simply lack the appropriate motivation then **JUD-INT** is not a substantive thesis – it's something the externalist can endorse. What the internalist needs to do is argue that there is something different which is wrong with the amoralist. Perhaps the internalist can claim that what is wrong with the amoralist is that they exhibit some sort of weakness of will. This runs into two problems: first, the amoralist case was supposed to be a distinct kind of rational amoralism, not attributable to weakness of will – a contention that van Roojen accepts and his picture is supposed to account for. Going this route collapses amoralist rational amoralism into another type of rational amoralism. Second, this kind of response relies upon being able to show that there is a principled difference between lacking the right motivational response due to weakness of will, and lacking it for some other reason.

The upshot of all this is that van Roojen's picture does not move the debate between internalism and externalism along very far. The internalist's case still turns on explaining rational amoralism via weakness of will, rather than a mere lack of the right motivational state on the part of the amoralist. Internalism also offered us the hope that we could explain

the rational authority of morality, but the materials that van Roojen offers us actually shows that this claim is shaky, and again depends on the viability of explaining the amoralist's situation using weakness of will. We seemed to have started from a standoff between internalism and externalism: the internalist insists that the amoralist either makes no genuine moral judgement or betrays some sort of practical irrationality, whereas the externalist insists that the amoralist makes a genuine moral judgement and their lack of motivation doesn't betray any practical irrationality. What van Roojen seems to offer us is in fact an elimination of one of the explanations of what the amoralist is doing available to the internalist – he agrees with the externalist that the amoralist is making a genuine moral judgement. If I am right then what we are left with is a disagreement over whether the amoralist betrays any practical irrationality. We have certainly been offered nothing like the compelling reasons available to someone like Williamson to support the claim that a character like Peter is making a mistake in taking the universal quantifier to be existentially committing.

What I hope to have shown here is that we have little pressing reason to endorse a form of judgement internalism of the type likely to cause problems for a moral realist. Thus if we embrace externalist moral realism then we will have a view that avoids Korsgaard's criticism if we interpret her as being concerned with the inadequacy of realism in explaining the motivational import of moral judgements. However, this is not yet to demonstrate there is a viable position available in this area. In the second half of chapter three I will look at two views that attempt to occupy this terrain and evaluate whether they are independently plausible. Before we reach that point, I wish to suggest there is another way of reading Korsgaard's complaint against realism. If there is another way of taking her argument, then it should be useful to have an account of that other reading before we evaluate how moral

realism does against her complaints. It is to that alternative way of reconstructing Korsgaard's argument that I now turn.

# CHAPTER THREE: THE GENERALISED ANTI-VOLUNTARISM ARGUMENT AND MORAL REALISMS

At this point it might be worth summing up what I have been trying to argue thus far. Korsgaard argued that the normative question (why should I be moral?) demands an answer. Moral realism, she claims, cannot provide an adequate answer. Therefore we should reject moral realism. She also claims that it is the force of the normative question that explains the force of Moore's open question argument, and Mackie's argument from queerness.

However, it's not particularly easy to get a grip on what Korsgaard's problem with realism is, precisely. I argued that if we interpret the normative question as a question about motivation then we can get an interpretation of Korsgaard's argument against realism which meets a number of aims. First, it is clearer what the problem is supposed to be; second the problem is located in the right kind of area; and we have an explanation of the link between the normative question and the open-question argument and the argument from queerness that Korsgaard posits.

On this interpretation, a normative sceptic (someone who doubts the normative force of morality) who asks the normative question is asking what is it about the fact that something is right that should move them to action. Korsgaard continually complains that the moral realist's answer to the normative question is inadequate because they advert to "just another fact" to explain why people are bound by morality. But if someone is unmoved by their duty to φ, then telling them that it really is a fact that they have a duty to φ will not help.

And, on the motivational reading of Korsgaard's complaint, this is still the problem. If the normative sceptic is not motivated by their duty, then telling them that it is a fact that that is their duty is hardly likely to motivate them. This is because merely believing that a fact obtains will not, if we assume a Humean account of moral psychology, move someone to action. So, on this reading of Korsgaard's rejection of realism, realism goes wrong in the place Korsgaard says it does – they try to use a moral fact at a point in the argument with a normative sceptic where the appeal to just another fact is of no use.

We also saw that we can give plausible interpretations of both the argument from queerness and the open-question argument where the arguments include judgement internalism as a premise. These interpretations are not only independently plausible, they also match Korsgaard's ambitions in terms of scope. The original open question argument, remember, was designed to only militate against naturalist forms of moral realism. But Korsgaard wants her normative question (which she claims lies behind the open question argument) to have force against naturalist *and* non-naturalist moral realism. We would then have a problem explaining how the normative question explains the force of the open question argument – the two would differ in purported scope, and we would have to finesse the connection between them. Fortunately, with the revised open-question argument that incorporates judgement internalism as a premise, no such finessing is required. The revised open-question argument turns out to have force against both naturalist and non-naturalist moral realism, so it's range matches Korsgaard's ambitions.

So, if we interpret Korsgaard's rejection of realism in terms of motivation we get to be able to explain a number of features of her views. This is still an advantage even if Korsgaard

herself would explicitly reject this characterisation of her argument against moral realism. This is because we would have an argument that could be endorsed by those who feel that there is something to Korsgaard's attack on realism, and it deserves proper scrutiny, but who are worried first about where it leads, and that there is still something slippery going on within the argument. This slipperiness comes from the fact that for Korsgaard's argument to work (under this interpretation) it would need to be true that moral realism *is* committed to judgement internalism. If it were not, then the realist could respond to Korsgaard simply by pointing out that there is a breed of externalist moral realism available.

This is why we have been looking at whether there is any compelling reason to think that moral realism has to be committed to judgement internalism. We saw that Smith's argument from fetishism, and van Roojen's treatment of rational amoralism don't provide us with an overwhelming case for internalism. However, it's one thing to say that there is no compelling reason why the realist has to endorse internalism, quite another to say that there is a viable version of externalist moral realism on the table. In section 3.3 I will look at two versions this type of moral realism, and see if they are at least plausible (plausible enough that if we are faced with abandoning realism altogether or picking one of them, it makes sense to pick one of them). In the course of that treatment I will cover another type of consideration thought to commit realists (and metaethicists more generally) to internalism – so called 'twin-earth' thought experiments involving moral terms.

However, there is a problem for taking this way of interpreting Korsgaard – it involves an argumentative strategy she disavows. In her discussion of Mill (an externalist moral realist) she claims that the normative question is not asking us why we should be motivated by our

moral judgements. Instead, it is asking us why we should endorse those judgements as genuinely normative (1996 §2.4). We can see that these questions can come apart, she says, when we consider a case where we find ourselves overwhelmed by our moral judgements – we can't help but be moved by them. Even in this case, she claims, we can still ask whether we should endorse those judgements. And *this* is what the normative question is aiming to get at. So, the line I am taking, although it has certain advantages, is not a line Korsgaard would endorse. As I have already pointed out, this is not necessarily fatal. The interpretation of Korsgaard's argument against realism I've offered could be accepted by those who think that Korsgaard is on to something, but need to see more details. However, what it means is that the argument identified in this interpretation is not *Korsgaard's* argument as she intends it. Is there, then, another way of reconstructing Korsgaard's argument so that we get something close to the spirit of Korsgaard's remarks about externalist moral realism, has some force against realism, and accords with the other ambitions Korsgaard has for her normative question? And what follows if we do find such an interpretation?

In the proceeding material I have made much of Korsgaard's claim that the normative question (and the argument built off the back of it) explains the force of the open question argument and the argument from queerness. Perhaps a better understanding of Korsgaard's argument against realism can be had by looking at the other affinity Korsgaard claims for the normative question – between it and an argument against voluntarism.

## 3.1 Voluntarism

A voluntarist, remember, claims that obligations derive from the commands (or will, or decisions) of a legislator. The most popular variant is theological voluntarism, where the relevant commands are God's. The voluntarist endorses something like the following claim:

> **VOL**: If agent *x* is obligated to perform action *a* then this is because the legislator commands, or in some other way wills, *a*.[44]

So when you are obligated it is because a legislator (typically God) has commanded that you act in a particular way.

But the voluntarist runs into a problem very quickly. We can ask of the voluntarist 'why am I obligated to obey the commands of that legislator?' According to the theory, all obligations come from the commands of the legislator, so it must be because she commands my obedience. But this cannot be right: the legislator cannot make it the case that I should obey their commands just by commanding that I do so – unless I'm already obligated to obey their commands then such a command will make no difference. The question becomes 'why am I obligated to obey the command that I am obligated to obey the commands of the legislator?'. The voluntarist could reply that you are so obligated because the legislator commands that you are. But this seems to add nothing – we can ask the same question about that command. And now the voluntarist has stepped on the path to an infinite regress.

[44] Adapted from Schroeder's (2005)

Instead they could claim that you are obligated to obey God's commands for some other reason. Pufendorf, for example, claims that we have an obligation to obey the legislator when they have *legitimate authority* over us. But if we follow this path, we have in effect given up on our voluntarism. Our obligations are now explained by something else – the legitimacy of the legislator.

Thus the voluntarist faces a dilemma – either they try to explain our obligation to obey the legislator's commands using another of those commands (which is of no help, and leads to an infinite regress of commands, each one of which is never properly explained); or they try to explain that obligation using something else (in which case they are no longer a voluntarist). We can call this problem the 'Cudworth problem' after the 17th century Cambridge Platonist Ralph Cudworth who originated it.

The other interpretation of Korsgaard available is to focus on her claim of affinity between the argument against realism built on the normative question, and the Cudworth problem for voluntarism.

Realists, she claims, attempt to end the regress threatened by voluntarism by fiat - by positing intrinsically normative entities that are supposed to stop a repetition of the normative question. For Korsgaard, this is a way of avoiding answering the question at all. Instead of telling us why some actions are obligatory, the realist posits intrinsically normative entities or relations found in the world – some actions are simply right, and this is because these actions are intrinsically obligatory. These normative entities are supposed to forbid further questioning – once we have discovered that certain actions are intrinsically

obligatory, that will be the end of the matter. Korsgaard contends that this kind of response to the Cudworth problem for voluntarism (and hence the normative question) does not really engage with the problem at all. What we were trying to get an account of was *why* some actions are obligatory. To be simply told that some actions are intrinsically normative is not an explanation of the normative force of the obligation to perform that action – instead it is merely a statement of the realist's confidence in the existence of moral facts.

If we emphasise this line of Korsgaard's thought, then what is wrong with realism is that it attempts to deal with the Cudworth problem voluntarism faced by stamping it's foot. We ask a voluntarist why we are obligated to do what God commands, and they have no good answer. We ask a realist why we should do what is right, and they simply insist that it's a fact that a certain action is obligatory. Neither response is adequate.

## 3.2 Voluntarism Reconsidered

We have seen that voluntarism faces a problem answering the normative question – we can ask why we should obey the commands of a legislator, to which the voluntarist seemingly had no good reply. But what is it precisely about the voluntarist's theory that is causing the problem? Remember, we had the voluntarist offering the following thesis (for simplicity I will discuss the problem in terms of theological voluntarism, where the relevant legislator is God).

**VOL**: If agent *x* is obligated to perform action *a* then this is because God commands, or in some other way wills, *x* to perform *a*. [45]

The problem arises when we ask where the authority of God's commands come from. One way to make this clear is to follow Mark Schroeder (2005) by asking what is it that makes God's commands (that the voluntarist says I should obey) different from, say, Jimmy Savile's commands (which, presumably, I should ignore). If God's commands aren't special in some way, then it is hard to see how the theory could even hope to be true. But what does this difference consist in? The voluntarist seems to be committed to saying that you ought to do what God commands you to do, whereas Jimmy Savile's commands generate no such obligations. So the voluntarist claims that:

**AUTHORITY VOL$_{God}$**: For every *x* (*x* ought to do what God commands)

Is necessarily true, whereas:

**AUTHORITY VOL$_{JS}$**: For every *x* (*x* ought to do what Jimmy Savile commands)

Is not necessarily true[46].

But, if the voluntarist is committed to **AUTHORITY VOL$_{God}$** then the argument Korsgaard provides starts to bite. That is because we can ask for an explanation of the obligation contained in **AUTHORITY VOL$_{God}$** – *why* ought I do what God commands? Well, according to **VOL** above, for every obligation, you should fulfil that obligation

[45] This way of laying out the problem is given by Schroeder (2005) 2-4

[46] **AUTHORITY VOL$_{JS}$** does not have to be false – as a contingent matter of fact Jimmy Savile could have commanded all and only those actions God does. But Jimmy Savile's commands in this case don't create my obligations, they merely track the truth about which obligations I have. God's commands create my obligations, so my obligations are linked to them in every possible world.

because God commands you to. So why should I do what God commands? Because God commands it. But this is not an adequate answer – if I have no reason to obey God's commands anyway, why would God commanding me to make any difference what so ever? So, the argument emerges out of the voluntarist's attempt to explain the authority of God's commands. It seems as if such an explanation, which goes via something like **AUTHORITY VOL$_{God}$**, which, together with **VOL**, leads to some sort of incoherence.

However, at this point we should start to be suspicious. It seems as if the argument generalises to any general explanatory moral theory. As Schroeder puts it, any such theory (which aims to give a unified answer to the question 'why ought I to do *a*?') will endorse:

> **THEORY**: For all agents *x* and action-types *a*, whenever *x* ought to do *a*, that is because *x* stands in relation R to *a*.

What relation R is will depend upon the particular normative theory that you endorse. But, whatever relation R is for a particular theory, that theorist will argue that relation R is special in some way. There are a number of ways in which you could be related to an action, most of which generate no obligations at all. Relation R on the other hand, must command some kind of authority: "but what does this authority consist in, to explain why being related by R to an action can obligate you to do it? It must be this: that you *ought to do* whatever action you are related to by R." (Schroeder, 2005, 4). But now we face the same problem as the voluntarist:

> Now that [the obligation to do what you are related to by R]... is exactly the sort of thing that **THEORY** is supposed to explain – why you ought to do something. But if we *need* this, in order for the explanations offered by **THEORY** to work, then it is hardly the sort of thing that **THEORY** *could* explain. Imagine: if it were not

> already the case that being related by R to an action obligated you to do it, then being related by R to the action, doing-whatever-you-are-related-to-by-R, would not make a difference. And if it were already the case, then it would not matter whether you were related by R to it or not. (4)

So it turns out that *any* general explanatory moral theory falls prey to this style of argument. For our purposes, this might seem a positive result – after all, Korsgaard not only wants to rule out voluntarism as a viable position, but *any* form of substantive realism. In fact, recall that the problem faced by the substantive realist was that they simply refused to offer an explanation of the source of obligation. The above argument seems to demonstrate that if the moral realist were to try to offer such an explanation it would fail for the same reasons as the voluntarist. So we might think that this generalised anti-voluntarist argument, combined with Korsgaard's normative question form quite a nice dilemma. Either the realist refuses to explain the authority of the obligations they insist are real – in which case they are failing to engage with the normative question at all; or, they offer an explanation, which falls down as a result of the generalised anti-voluntarist argument above.

However, there are two points to be made here. Korsgaard shares explanatory ambitions herself – she believes that her own neo-Kantian constructivist position can successfully answer the normative question without falling into the kind of normative regress which dogs the moral realist. The account she offers however seems to have the form of **THEORY**. We will see (in chapter 5) how she argues that you are obligated to act on that maxim which you can at the same time will to be a universal law, as an equal legislator in a kingdom of ends. This seems to fit well with **THEORY** (with 'act only on that maxim...' describing the relevant relation). Nevertheless, Korsgaard argues that the specific form of her theory avoids the problems faced by the moral realist. If so then we would have an

instance of a general explanatory theory that endorses **THEORY**, so it would seem that (if Korsgaard is right) there is at least one way of avoiding the generalised anti-voluntarist argument.

Second, what the generalist anti-voluntarist argument attempts to show is extremely strong – any general explanatory moral theory is incoherent. We may feel inclined to accept such a theory for other, seemingly persuasive reasons. So is there any way that the anti-voluntarist argument (and hence the generalised anti-voluntarist argument) can be resisted?

Remember, we originally started by asking what the authority of God's commands consist in. One way of glossing this was to ask how God's commands are different to Jimmy Savile's. The voluntarist runs into trouble when they appeal to **AUTHORITY VOL$_{God}$** as an explanation – put simply, it contains an ought which is covered by **VOL**, and this leads to the incoherence we saw above. However, Schroeder points out that **AUTHORITY VOL$_{God}$** is ambiguous between two possible readings:

**CONDITIONAL VOL**: Necessarily, for every *x* and every *a* (God has commanded *x* to do *a* → *x* ought to do *a*)

Or:

**CATEGORICAL VOL**: Necessarily, for every *x* (*x* stands in the *ought to* relation to the action-type: *doing whatever God commands*)

It seems that holding **CONDITIONAL VOL** for God's commands and rejecting it for Jimmy Savile's commands is adequate to describe the difference between God's commands (which lead to obligations) and Jimmy Savile's (which do not). However, **CONDITIONAL VOL** on its own does not fall prey to the anti-voluntarist argument. That is because is not committed to there actually being any obligations – it could be the case that there is nothing God has commanded you to do. We ran into trouble with **AUTHORITY VOL$_{God}$** because it contained an obligation that our voluntarist theory had to explain. So on this reading, we can avoid the anti-voluntarist argument.

Schroeder asks us to compare this with **CATEGORICAL VOL** – this states that there is an obligation that everyone has – doing whatever God commands. This reading would allow the anti-voluntarist space to make their point – how do we explain that obligation?

But if **CONDITIONAL VOL** is enough to explain the difference between God's commands and Jimmy Savile's, why should we also accept **CATEGORICAL VOL**, and leave ourselves open to the anti-voluntarist argument? One reason might be that we think that we need **CATEGORICAL VOL** to complete our explanation of the authority of God's commands. Schroeder offers us an intuitively plausible model of moral explanations under which something like **CATEGORICAL VOL** would be needed to complete the explanation of God's authority, which Schroeder calls the standard model (**SM**):

> **SM**: The explanation that *X* ought to do *A* because *P* follows the *Standard Model* just in case it works because there is (1) some further action *B* such that *X* ought to do *B* and (2) not just because *P* and (3) *P* explains why doing *A* is a *way* for *X* to do *B*.

And this gets us to the anti-voluntarist argument:

> The voluntarist believes that whenever you ought to do something, that is because God has commanded it. For this explanation to follow the Standard Model, it must appeal to some further thing that you ought to do, and not just because God has commanded this thing. And since that further thing that you ought to do falls under the scope of the theory, the explanation of why you ought to do it must appeal to the same thing – namely, itself. But that makes the explanation circular (10).[47]

And this generates the anti-voluntarist conclusion. So if all normative explanations follow the **SM** then we need **CATEGORICAL VOL** as well as **CONDITIONAL VOL** to complete our explanation of God's authority, and thus we are open to the anti-voluntarist argument. But do all normative explanations work this way? In other words, should we accept the Standard Model Theory (**SMT**)

> **SMT**: For all *x*, *a* and *p*, if *x* ought to do *a* because *p,* that explanation must follow the Standard Model

It seems as if we do have an alternative type of normative explanation, one offered by a reductive or constitutive explanation. For example, the voluntarist could offer the following:

> **CONSTITUTIVE VOL**: 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*.

This is a reductive thesis about oughts because it analyses oughts in terms of something else – God's commands. If this type of reductive explanation is viable, then it gives us a way

[47] In support of this diagnosis, Schroeder points out that the original purveyors of the anti-voluntarist argument were also the first proponents of the standard model of normative explanations. See Price (1948, 52-53), Cudworth (1996, 20).

around the anti-voluntarist argument. We agree with them that some normative explanations follow the STANDARD MODEL, but at some point these will be grounded in the type of constitutive explanation given above. Recall that the generalised anti-voluntarist argument aimed to show the inadequacy of a general explanatory moral theory. Can those other theories use this same strategy? It seems clear that they can. For example, a consequentialist might want to endorse something like:

> **CONSTITUTIVE CON**: For it to be the case that *X* ought to do *A* is just for it to be the case that doing *A* will bring about the most good.

If this type of explanation is legitimate then, we have a way in which both the voluntarist or any general explanatory moral theory can avoid the generalised anti-voluntarist argument. For the purposes of this thesis, it would mean that Korsgaard puts the anti-voluntarist argument to ill use. She means to use it as part of a strategy to dismiss all types of substantive moral realism. As we've seen so far though, it appears as if using the above type of explanation provides space for all reductive theories to avoid the anti-voluntarist argument.

Are there any reasons why we shouldn't expect explanations of this kind to be legitimate? Well, Schroeder argues that the original proponents of the anti-voluntaristic argument offered it as one half of a dilemma for the voluntarist – either they fall prey to the argument or they have to give an explanation like the ones given above. However, if they do that, then they fall prey to pre-cursor of the open question argument. For example, Price says:

> 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 [...] 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, 16-17, from Schroeder's 2005)

So if it being the case that you ought to do φ just was for φ-ing to be commanded by God, then it would betray some kind of conceptual confusion to ask 'should I do what is commanded by God?'. It seems clear that Price thinks that this will always be an open question, so being obligated to do φ can't just be φ-ing's being commanded by God.

So what does all this mean? Well, we have seen that there are two possible interpretations of what is going on in Korsgaard's rejection of moral realism. First, she could be getting at an issue to do with moral motivation. Second, she could be using the normative question to generalise the Cudworth problem faced by voluntarism to moral realism. We have also seen that there are strategies to escape these problems. To escape the first problem, we need to abandon internalism and give an account of a plausible form of externalist moral realism. To escape the second, we can first note that generalising the Cudworth problem is a bad manoeuvre for Korsgaard – it would threaten the kind of theory of moral obligation she is eventually hoping to offer. We have also seen that there is an alternative conception of explanation that allows reductivist moral realism to escape the Cudworth problem. However, this kind of moral realism becomes vulnerable to an open question argument style objection.

To fully complete a defence of moral realism against Korsgaard's attack we need two things. To deal with the first problem we need a sketch of a viable form of externalist moral realism. To avoid the second we need a sketch of a viable form of reductivist moral realism that can avoid the open question argument. It's to these aims which we now turn.

## 3.3 Externalist Moral Realism

There are, of course, a number of moral realisms available. What we are looking for in response to Korsgaard is a form of moral realism that at the very least embraces externalism about moral motivation (in order to circumvent my revisionary reading of the normative question) and, if possible, explores a reductionist strategy which we can exploit to escape the generalised anti-voluntarist argument above.[48] We can distinguish between *analytic* and *synthetic* versions of externalist realism. I will briefly consider a version of the former, before looking at one version of the latter in more depth – the so-called 'Cornell realism' of Brink, Boyd and Sturgeon. I will not be able here to examine each in much detail, but I hope to show that both Cornell realism and Finlay's analytic reductivism can at least resist the main arguments directed against them, leaving them viable alternatives to Korsgaard's neo-Kantian constructivism.

[48] As most forms of realist *non*-naturalism (e.g. Moore 1903, McDowell 1998, Wedgwood 2007) are advocated, in part, on the grounds that they can sustain a commitment to *internalism* these views will not be relevant for our purposes. This is fortunate as the metaphysical and epistemological commitments of non-naturalism are troubling. According to Stephen Finlay (forthcoming) it is even typical for non-naturalists to acknowledge this discomfort but to provide transcendental arguments to the conclusion that we must embrace non-naturalist moral properties. The idea being that our moral practices demand such properties so we must accept their existence despite the unwelcome consequences of adding them to our ontology. In investigating *naturalist* and externalist moral realism we will be indirectly testing this claim – if a naturalistic realism is plausible, then we don't need to take up non-naturalism.

## 3.31 Analytic Naturalism

Analytic naturalisms attempt to find analytic connections between moral and naturalistic predicates, thereby reducing moral facts to natural facts. One version of this view is the analytic functionalism associated with Frank Jackson and Philip Pettit.[49] A more recent, externalist, version of this type of view is Stephen Finlay's analytic reductivism. Here I will briefly outline Finlay's end-relational theory of moral semantics and then examine how well it does. I will argue that although Finlay's account has major benefits it also faces a methodological worry and may be vulnerable to a modified open question argument.

So, what does the account look like? Finlay's overarching concern is to offer a semantics for all normative terms as they are used in nearly all contexts rather than offering an account limited to just the moral uses of those terms. He points out that we use a word like 'good' in both moral contexts – 'Jane is a good woman' – and non-moral contexts – 'This is a good knife.' The same goes for terms like 'ought', 'should', 'must' and 'may': we say both things like 'Governments should not torture people' and 'You should run for the bus.' Finlay argues that we should prefer an account of these normative terms that is unified in the sense of offering the same general analysis of moral and non-moral uses. This is preferable on at least two grounds: 1. Without a unified account we would be forced to claim that 'good' (for example) in moral contexts possessed a different sense from 'good' in non-moral contexts. We would then need an explanation of how or why these two distinct meanings came to be associated with the same words. This concern is made particularly

[49] See Jackson 1992, Jackson and Pettit 1998 and Jackson 1998. For problems with this 'network analysis' of moral predicates see Smith 1994 and 1998 and McFarland and Miller 1998. This sort of view is also internalist, so not suitable for my purposes.

pressing by the fact that this feature of normative vocabulary (that it can be used in moral and non-moral utterances) is widely cross-linguistic – we could just about accept that this feature of English is pure coincidence, but this would not be anywhere near satisfying for accounting for a feature found in a large number of languages. 2. We hope that our account of moral language will be conducive to an explanation of how language users so easily make use of normative terms. If the semantics we offer for normative terms is unified and thus simpler the task of explaining how language users manage to learn and use these terms competently becomes easier.[50]

Finlay's *method* is to seek *semantic analyses* of normative terms based on linguistic data. The analytic reductivist hopes to show that normative concepts are composed out of simpler, non-normative concepts. To offer a semantic analysis of a complex concept is to explain which simpler concepts it is composed of – e.g. we might analyse the concept BACHELOR as being composed out of the simpler concepts of UNMARRIED, MALE, ADULT. These analyses are supposed to be connected to meaning in some way, so that the truths they express will be analytic - true in virtue of meaning. Finlay holds that as concepts are mental entities the study of their structure is an *a priori*, armchair pursuit; the main data relevant being our linguistic intuitions about the appropriateness/grammaticality of various constructions involving the suggested analysis.

[50] Finlay offers an interesting reason for preferring simpler to more complicated accounts of moral semantics in general. It might be quite easy to mould a complex theory to the linguistic data; however, it is unlikely that a simple theory that predicts all the relevant data is correct by accident. Therefore a simple theory that explains all the available data is more likely to be correct. See his (forthcoming) Ch. 1.

Finlay's suggested analysis holds that normative utterances are *end-relational*. The basic idea is that to say that something is good (for example) is to say that it raises the probability of some end obtaining. The idea is easiest to see in a non-moral case. Suppose I tell you whilst you are chopping bread that: 'This is a good knife.' The semantic content of that utterance is (very roughly) 'Using this knife raises the probability of the bread being cut.' That is, I am telling you that using the knife raises the probability of obtaining some end[51]. Although the relevant end is not always explicitly mentioned it is supplied by the context of the utterance – if I say a knife is good whilst we are making sandwiches I say that it is good for (raises the probability of us succeeding in) sandwich making. If you and I are violent thugs I might use the same sentence to say that the knife is good for stabbing people. Similar considerations hold for other normative terms like 'should': we can explicate the content of my utterance of 'You should run for the bus' as 'Running for the bus raises the probability of you catching it.' Finlay repeats this same sort of analysis for the most important normative terms ('must', 'may', 'should', 'ought', 'good', 'reason', and others) using a variety of linguistic data to explain the relationship between the target everyday normative utterances and the reductive analyses of their semantic content and how and why the everyday utterances can plausibly be accounted for as elliptical for the more obviously end-relational analyses.

The first thing to note about this theory is that it is reductionist. We explain normative notions in terms of something non-normative – the concept of raising the probability of an end obtaining.[52] This means that the analytic reductivist can easily adopt the type of

[51] Things are slightly complicated by normative utterances sometimes being relativised to multiple ends, but these details won't be important here

[52] One historical irony is that (at least) Blackburn's *expressivist* semantic programme is inspired by proto-expressivist approaches to *probability.* Like moral judgements, probability judgements also have some

constitutive explanations that avoid the generalised anti-voluntarist argument – Finlay tells us that the semantic analyses he offers tell us about the *essences* of normative properties: in other words, what kind of things that they are. It will then be legitimate for the analytic reductivist to say that goodness, say, *just is* the property of raising the probability of some end obtaining.

Finlay also gives us an account of how his end-relational theory allows us to actually *simplify* the standard possible world semantics for modal auxiliary verbs like 'must' and 'may' (see his Ch. 4).

Another advantage of the analytic reductivist account is that it offers an explanation of the appeal of judgement internalism without committing to an implausibly strong form of internalism. The reductivist aims to account for the practicality of moral judgements by attending closely to their *pragmatic* content. Finlay distinguishes between semantic content (the proposition/s conventionally expressed by a particular combination of symbols) and pragmatic content (propositions that are communicated [intentionally or non-intentionally] by the utterance of a sentence in a particular context)[53]. One feature of normative

---

puzzling features, leading some to try to account for them as being expressions of our confidence in some particular event happening (see Blackburn, unpublished). Finlay, on the other hand, wishes to reduce normative judgements to a class of probability judgements in the service of a *cognitivist* semantics. The expressivist proposal about probability judgements, however, is not uncontested (and the metaphysical and epistemological issues might be very different in the two domains), so Finlay's account may well be perfectly reasonable. It is just interesting that the same type of judgement can both inspire expressivist accounts of morality, and also be seen as a suitable reduction base for a realist cognitivism.

[53] It is far beyond the scope of this thesis to reach a judgement about whether this way of carving the semantics/pragmatics distinction is correct, or even whether such a distinction is philosophically fruitful. For my purposes I will just take on board Finlay's taxonomy to see whether his explanation of the practicality of moral judgements is plausible on its own terms.

judgements is that the end that they are relative to is sometimes *suppressed* – I can say that something is a good knife without adding what it is good for, or that an action is obligatory without explicitly saying what end it is necessary for. This is one of the reasons we have to do some semantic analysis to reach the end-relational theory – if the end was always mentioned, the semantics Finlay offers would just be obvious. It is these two factors (that moral utterances, like others, can communicate pragmatic content within contexts; and that ends are sometimes suppressed) that yields an explanation of the practicality of moral judgements. The situations within which it is appropriate to suppress an end when offering a moral judgement are those when the end the judgement is relativised to is salient for the speaker and their audience. So, say I am speaking at my local conservative association and say 'George Osborne's economic policies are good.' This can be perfectly appropriate because the end that George Osborne's policies raise the probability of (redistributing wealth away from the poor to the rich and providing cover for a scaling back of the welfare state, say) is one which is sufficiently salient for us in that we all share it and know that it is a concern we are in the business of promoting. Thus, it should not be surprising that *these* judgements (where the relevant end is suppressed) are tightly connected to motivation to act – it is appropriate to utter them in this form only when the end they are relativised for is one that the speaker and the audience are committed to. So, by missing out a semantic component of the moral proposition I am expressing with my utterance (the end it is relativised to) I pragmatically implicate that it involves an end that I or we have (Finlay, forthcoming, Ch. 6).[54]

[54] One might be worried that people have ends when the desire that some state of affairs obtains. Then Finlay's view will have trouble explaining the seeming categoricity of moral requirements – that they apply to us regardless of our desires. Finlay makes two moves to deflate this worry: first, he argues that something being an end does not require a desire or even a desirer. Second, he tries to generate an explanation of the *seeming* categoricity of moral judgements in terms of pragmatic factors.

This explanation of the practicality in terms of pragmatics rather than semantics relieves internalism of some of its troubling properties. The problem with strong forms of internalism is that they say that being appropriately motivated is a constraint on making a judgement with a particular semantic content. This then forces us to either build a desire-like motivational element into the semantic content (I will examine these sorts of accounts in ch. 4) or revise our theory of moral psychology in a non-Humean direction. The analytic reductivist isn't forced to do this – instead they simply hold that speakers can pragmatically imply that they are motivated in the direction of some moral judgement by choosing to utter it in a particular way (suppressing the relevant end). True, this sort of account does not accommodate a strong form of internalism – it's not conceptually impossible for there to be an amoralist. But we saw above that such forms of internalism are poorly motivated. What the analytic reductivist offers us, then, could be all the internalism we need.[5556]

However, it's possible that Finlay secures this sop to the internalist at the cost of not taking the possibility of an amoralist seriously enough. Consider what, on the analytic reductivist account, it is that an amoralist is doing. Say they utter 'Meat-eating is wrong' whilst lacking any motivation to refrain from eating meat. By suppressing the end that refraining from eating meat secures (preventing animal suffering, perhaps) the amoralist pragmatically implicates that they have that end. Now, an amoralist is not supposed to be stupid – they

[55] What's important to note here is that this appeal to pragmatic content avoids the worry that sometimes philosophical appeals to the semantics/pragmatics distinction are ad hoc. This application of the distinction only yields a plausible story because of the details of the end-relational theory that it is combined with – it is the fact that we can miss out ends when they are shared with our audience which explains the practicality of moral judgements, and we would expect to find this form of ellipsis based on considerations about how pragmatic content works in general: we don't need to tailor our account of pragmatics specifically to this case.

[56] Finlay remains an externalist though, for he agrees that it is possible for a semantically competent agent to make a moral judgement without feeling appropriately motivated. To express that judgement would be pragmatically improper (in lots of contexts) unless the relevant implicatures were cancelled.

should be just as good at working out the pragmatic content implied by their remarks within particular contexts, and they also know that their utterance is semantically completed by the particular end that they are implying they have. So, if Finlay is right, then the amoralist is doing something that they know is at least conversationally improper. In fact, what they seem to be doing is a variation of what the inverted commas theorist has them doing. Remember, for the inverted commas theorist when an amoralist says that something is good, they are only using 'good' in inverted commas – they really mean that the thing is judged to be good by others, for example. Now, on Finlay's account what the amoralist does is similar – by suppressing the end their normative utterance is relativised to they are acknowledging that that end is one that is sufficiently salient for their audience – that a large proportion of the audience have that end and will be moved by it. So, although they are not using the symbol "good" semantically improperly (as the inverted commas theorist has it) they are using it improperly at the level of pragmatics by implying that they have an end that they don't. But, suppressing the end shows that they DO know that the end is shared by their audience; so, in effect, they are performing a speech act that amounts to them acknowledging that most other people would have the relevant end. At this point, the friend of the amoralist could argue that this strategy is sufficiently similar to the original inverted-commas strategy for us to reject it as a characterisation of what the amoralist is doing.

The analytic reductivist could explore two avenues of response at this point. First, they could point out that speakers can implicate information that they don't intend to communicate (at least on Finlay's way of drawing the semantics/pragmatics distinction. See his Ch. 6). Therefore, we could just claim that the amoralist, although semantically competent, just isn't very good at working out the pragmatic content of their utterances. The

analytic reductivist could then point out that the considerations I brought out against the earlier treatments of the amoralist simply don't apply in this case – I argued that reliable application of a term like 'good' to the right extension of objects is *prima facie* evidence of semantic competence. The analytic reductivist agrees that the amoralist is semantically competent – they just say that they are a bit slipshod when it comes to pragmatic content.

However, this sort of response may not go far enough – in order to understand that they can miss out the end their normative judgement is relativised to the amoralist has to understand that it is an end that their audience shares, even if they do not have it themselves. If they did not know this, then their utterance would be semantically incomplete – it would fail to communicate any determinate proposition.[57] So, it looks like the amoralist knows enough about how normative terms work to know that you can suppress an end when your audience has that end, but they seem to have a very particular blind spot when it comes to the relationship between those terms and their own ends. We might worry then that Finlay's treatment of the amoralist is not as principled as he would like it to be.

The second avenue the analytic reductivist could explore is a particular feature of pragmatic content – its *cancellability*. Because pragmatic content is only conversationally rather than conventionally implicated by a certain collection of linguistic tokens, it's possible to cancel the pragmatic content implied by the use of an utterance within a context – to indicate that you don't wish to communicate that content. For example, suppose we were in a restaurant together and I asked you what you thought of it. If you said 'Well, the waiters are polite'

[57]Or, more accurately, if they did not know this information it would be something of a mystery how they manage to know when to suppress the relevant end.

and left it at that, you literally say that the waiters are polite; but you might also conversationally implicate the proposition that the food isn't up to much. However, you can supplement your utterance in order to cancel that particular implication – you might say 'Well, the waiters are polite. But the food isn't bad either.' Now, the analytic reductivist could claim that it is always appropriate for amoralists to cancel the problematic pragmatic content of their utterances. That it is appropriate for them to say things like 'Meat-eating is wrong. But I don't mean to suggest by that that I intend at all to stop eating meat.' However, the amoralist might reject this as a characterisation of their practice – they might claim that that is not what they are thinking when they think about meat-eating – they think just that meat-eating is wrong. This would be problematic as it's typically thought that conversational implication of pragmatic content is something that happens at the level of public conversation – to put it very simply, that the laws of pragmatics do not apply at the level of thought. If, then, the amoralist rejects this characterisation of what they are trying to communicate, then this would be some evidence against the reductivist proposal. Finlay himself believes that pragmatic considerations apply even at the level of thought, so he is unlikely to be moved by this worry. Whether this sort of view (that pragmatic considerations apply at the level of thought as well as conversation) is plausible will depend, in turn, on what we think of thought. If we think of thought as something like the *internal monologue* that we hear in our heads, then the Finlay type of view might seem plausible – when talking to ourselves we might employ things like ellipsis to save ourselves time, etc. However, if one thinks of thought in a more Fregean and determinate way – as that for which truth or falsity is an issue – then the idea of pragmatic considerations applying to thought might look incredibly strange. It's beyond the scope of what I am trying to do here to investigate this particular issue beyond this *extremely* rough sketch. All I want to conclude from this is that Finlay's pragmatic treatment of practicality is not entirely

unquestionable, but could be made to work depending upon our ancillary assumptions about the general nature of thought and language.

We might be concerned about the analytic reductivist's general framework – particularly their search for analytic connections between complex normative concepts and simpler non-normative concepts. Forty years ago this concern would be easy to express – you could just mutter something about *Two Dogmas of Empiricism* or *Truth by Convention* (Quine, 1951; 1935) and hope your audience shared your distrust of the idea of analyticity. Nowadays, following the work of people like Paul Boghossian and particularly Gillian Russell (Boghossian 1996, 1997, Russell 2008) to rehabilitate analyticity such a mere expression of ethos is not enough to satisfy anyone. However, there is still a worry here for Finlay about whether the defence of analyticity he relies upon fits with his methodological programme. He cites Gillian Russell's recent work (2008) as a defence of the notion of analyticity he wants to make use of, so it should be fruitful to very briefly consider how Russell's account works.

Russell's insight is to point out that traditional characterisations of analyticity as truth in virtue of meaning are particularly murky. This is because we can distinguish between four elements of language that we could call 'meaning':

> **Character:** the thing speakers must know to count as understanding an expression
>
> **Content:** what a word contributes to what a sentence containing it says
>
> **Reference Determiner:** a condition which an object must meet in order to be the referent of, or fall in the extension of, an expression.

**Referent/Extension:** the (set of) object(s) to which the term applies.

(Russell, 2008, 45-46)

Russell argues that the lesson we should take from 20th century philosophy of language is that these different notions can come apart in various ways. For example, the work of externalists like Burge (1979) and Putnam (1975) shows that we can count as understanding an expression with a merely deferential grasp of that expression – we might count as understanding the meaning of 'beech', say, without knowing its reference determiner and thus being unable to distinguish beeches from elms. Similar lessons can be drawn from Saul Kripke's work on natural kind terms and proper names (1980) and David Kaplan's on indexicals (1989).

The account of analyticity that Russell defends revolves around reference determiners. Sentences turn out to be analytically true when the right sort of relations of containment and exclusion hold between the reference determiners of the (logical) subject and (logical) predicate of that sentence.[58] To flesh this out a little – the condition that an object has to meet for being a bachelor contains the condition of being a male, for example, so the sentence 'All bachelors are male' comes out as analytic. In contrast, the condition for being a renate (having kidneys) does not sustain the right connections to the condition for being a cordate (having a heart) so 'All renates are cordates' would not be analytic, even if the extensions of 'renates' and 'cordates' overlapped. What drove a lot of the Quinean concerns about analyticity, Russell contends, was a failure to recognise that reference determiners were the best source of an account of analyticity. Suppose you ran all of the elements of

[58] This characterisation leaves out a lot of detail, including what Russell means by the metaphorical sounding 'containment' and 'exclusion' and what it is to be the 'logical' subject or predicate in a sentence. The full details are in her Ch. 3, but they should not be needed here.

language that Russell teases apart together. Then the notion of analyticity that was revealed could look quite puzzling – to count as competent with an expression you'd have to know how its reference is determined, which things it applied to, etc. Then you could, merely in virtue of being competent with a concept know various truths about it – e.g. that a fortnight is a length of time that lasts 14 days. However, it shouldn't be surprising that analyticity within *this* framework looks shaky – we have learned that, for example, merely being partially competent with the concept FORTNIGHT does not suffice for knowing that a fortnight is 14 days. But, this does not mean that the notion of analyticity in general is suspect. Instead, we merely need to amend our conception of analyticity to respect the lessons that philosophy of language has taught us. Mere conceptual competence does not mean you will be able to recognise the analytic truths that involve that concept, when they are presented to you, but that's just because mere conceptual competence does not guarantee that you know the reference determiner of that concept which is what accounts for the analytic truths involving that concept.[59]

Of course, there is room to still be sceptical about this version of analyticity (see for example Boghossian's review of Russell (2010) or Williamson's (2007) where he makes cases against the metaphysical account of analyticity Russell defends AND the epistemological conception Boghossian prefers), but we can circumvent a lot of this debate if we instead consider whether the account Russell offers supports the use Finlay wishes to make of it.

[59] You might worry that on a Fregean picture sense (like a reference-determiner) determines reference, so that grasp of sense does imply grasp of reference-determiner, so in virtue of grasping the sense of an expression you should be able to work out the analytic connections that it sustains. At this point Russell would simply deny the traditional Fregean picture – if sense is what determines reference, then it is not something that is always grasped by a speaker, in that being able to use a term is compatible with not knowing its sense, in this sense.

One possible source of tension is that Finlay conceives of semantic analysis as basically an *a priori* armchair pursuit (Finlay, forthcoming, Ch. 1). This is what allows him to derive semantic analyses of normative concepts from the data he has available to him (our linguistic intuitions). This, however, is problematic given the distinctions Russell makes between different elements of language. Analytic truths are those, remember, that are true in virtue of reference determiners. And, the lesson of recent philosophy of language is that mere competence with a concept (grasp of its character) sometimes does not suffice for grasp of that concept's reference determiner. Thus, there is space for a view that says, roughly, that we can grasp the character of moral terms without knowing very much at all about their reference determiners. If this is the case, then working out any analytic truths to do with those concepts will require more than *a priori* reflection – Russell's view is explicitly committed to the possibility of both contingent analytic truths (like 'I am here now') and *a posteriori* ones. We shall look at a view of moral terms which *does* allow grasp of moral concepts to be insufficient for knowledge of reference determiners (and thus knowledge of any nearby analytic truths, if there are any) in the next section, and Finlay's response to that sort of view. Whether the analytic reductivist's methodology is appropriate depends upon whether Finlay can rule out this sort of view; this is problematic because it's the sort of view that Russell herself uses (in part) to construct her account of analyticity – the one Finlay adverts to.

A more particular concern with the 'analytic' in 'analytic reductivism' is not to do with analyticity in general, but with its application to the metaethical case. It's been thought that Moore's open question argument militates against metaethical analyses in particular. The open question argument, basically, claims that for any suggested naturalistic (or metaphysical) analysis N of a moral term D the question 'I see that x is N, but is it D?' is

always open (in the sense that sincerely asking it does not betray any conceptual or semantic incompetence). If the analysis were correct this question would not be open, so the naturalistic analysis is not correct. Now, Moore then goes on to push this argument further than it will go – he moves from there being no correct naturalistic *analysis* of moral terms to the impossibility of a naturalistic *reduction* of moral terms. However, many have accepted the first step – that the open question argument shows that we cannot have analyses of moral terms in non-normative vocabulary. How does Finlay deal with this?

Finlay makes three points. First, as the open question argument is usually presented, it is far too quick. We are usually given one or two putative analyses of normative terms into non-normative vocabulary, presented with the claim that these leave the question open, then we are expected to accept that this shows that ALL suggested analyses must be false, even in advance of hearing them. This seems like a rash overgeneralisation.[60] The anti-analyticist could bolster their case by appealing to their intuitions about all possible analyses as when Wittgenstein says: "I at once see clearly, as it were in a flash of light, not only that no description that I can think of would do to describe what I mean... but that I would reject every significant description that anybody could possibly suggest, *ab initio*" (1930, as found in Finlay, forthcoming Ch.2). Against this Finlay simply points out that true analyses can sometimes be surprising and informative – it seems hopelessly overconfident to suggest that no analysis of moral terms could work just because you would be surprised if it did: far better to actually look at the suggested analyses and see what we think then. Third, and most importantly, Finlay suggests we can deal with the seeming openness of the open question argument by again attending to the distinction between pragmatics and semantics. How does this work?

[60] This is one criticism that could be extracted from William Frankenna's (1939).

Well, on Finlay's account a word like 'good' expresses an incomplete predicate which requires relativisation to an end before sentences containing it express propositions. What this means is that 'good' can be used to predicate any number of different properties of an object (having the property of raising the probability of the bread being cut, or of animals' rights being respected, or of catching the bus, etc) depending on the context involved. This is supposed to explain how someone can always ask 'I grant that x is N, but is x good?', as there are a number of different properties N can refer to; suppose that the analysis N1 correctly defines one property we can refer to with 'good', there will still be other properties that we can use 'good' to refer to. So, if someone says to us 'I see that x is N1, but is it good?' we have to understand them as acknowledging that that x falls under the predicate N1, but asking whether x falls under some other predicate we can express with the term 'good'. This is given a pragmatic explanation – we very rarely waste people's time with questions if we can avoid it, so if someone asks the open question they must be using a different completion of 'good' in that question. We can draw a useful comparison with other incomplete predicates. It's sometimes held that expressions like 'tall' need to be relativised to comparison classes – what is tall for an ordinary adult male is not what is tall for basketball players. So, supposed you analysed 'tall' for ordinary adult males as 'over 6ft4'. Now, if someone asked 'I see that Jimmy is over 6ft4, but is he tall?' you don't have to take the openness of the question as a refutation of you proposed analysis. Instead you can adopt the pragmatic explanation Finlay advances and conclude that the questioner is asking whether Jimmy is tall *in some other sense*. Thus, the supposed openness of the open question is a consequence of the fact that we can relativise 'good' to a wide variety of ends depending on context, together with the pragmatics of asking questions involving incomplete predicates. In effect, what Finlay is doing is explaining away our intuition that the relevant question *is* open by giving a debunking analysis of what the question is really

asking for. Thus, if Finlay's response is adequate it has force against updated versions of the open question argument, like the one explored in §2.14.

At this point the anti-reductivist could put their worry a different way. Although there are many ways of completing the predicate 'good' and thus expressing many different propositions using sentences containing that predicate, it should be possible, in principle at least, to specify all the permissible completions and thus the propositions that can be expressed using 'good'. The anti-reductivist can then claim that they can ask the open question argument of that complicated mass of predicates and propositions. The trouble with this strategy is that although the anti-reductivist may *claim* that they would be able to ask the open question of that complicated analysis, they have very little evidence that they would – they are back to relying upon the Wittgensteinian intuition mentioned above: that they would reject ANY analysis of normative terms out of hand. And Finlay already has an answer to that sort of objection – it's just far too quick.

So, we have seen that analytic reductivism offers a number of benefits – it gives us a unified semantic picture of normative terms, it avoids commitment to non-natural properties, it can go some way to explaining the appearance of practicality connected to moral judgements and as it is reductivist it avoids the generalised anti-voluntarist argument we drew out of Korsgaard. However, its explanation of the practicality of moral judgements is not entirely problem-free, and we might still be sceptical of the appeal to analytic connections between normative and non-normative predicates (although perhaps not because of any worry about moral terms in particular, but just because the account of analyticity that Finlay relies upon

may not support the analytic reductivist's methodology, a question we will be indirectly tackling in §3.32-5).

## 3.32 Cornell Realism

Cornell realists (e.g. Sturgeon 1985; 1986; 1988; 1992) offer us a non-reductionist externalist moral realism. They combine two main claims: 1. we can model the semantics of moral terms on the framework for natural kind terms given by Saul Kripke and Hilary Putnam; and 2. that moral properties do not have to be reduced to natural properties as they earn their ontological keep by featuring in our best explanations of natural phenomena. Both of these claims have been challenged and I will attempt to elucidate the Cornell realist view by looking at each of them. Turning to the second claim first.

## 3.33 Moral Properties and Ontological Commitment

Cornell realism claims that we cannot analyse moral vocabulary in terms of non-moral vocabulary (like Finlay's analytic reductivism above); neither can we reduce moral properties to non-moral properties as a matter of synthetic fact. However, moral properties are still perfectly naturalistically respectable according to the Cornell realist. This is because there is a useful analogy to be had between moral terms and the 'natural kind' terms found in biology, chemistry and the social sciences.

The basic idea is this: there may not be any reductive analysis of terms like 'gene', 'species', 'acid', 'culture' or 'catalyst' available to the biologist, sociologist or chemist in physical terms[61]. There may be many ways to realise the type of acid, say. Nevertheless, these are theoretical terms in good standing which refer to properties which deserve to be included within our ontology because these properties earn their ontological keep by being explanatorily relevant. There are biological explanations of naturalistic phenomena, like the distribution of certain morphological features amongst a population of animals within a certain area, that would be lost if we omitted the concepts of GENE or SPECIES from our best scientific theories. And this is true even though gene-hood and species-hood resist reduction to the terms of physics. So, because these biological, sociological or chemical kinds feature in our best explanations of natural phenomena we should accept them into our ontology despite their resistance to reduction.

This way of looking at things reflects a broadly Quinean conception of ontological commitment. Quine argued (1948) that we should base our ontological commitments on which entities are ranged over in the statements of our best scientific theories in some canonical notation (typically first-order logic with identity). To be, then (in the famous phrase), is to be the value of a bound variable. One area where this has played out is with regards to the existence of numbers and other mathematical objects. Indispensability arguments for the existence of mathematical objects start by claiming that there are some scientific facts that resist being captured without quantifying over numbers or other

[61] In fact, the relevant reduction bases here would be different for each case – the reductionist chemist would aspire to reduce chemistry to physics, the reductionist biologist biology to chemistry, the reductionist sociologist to some combination of economics and psychology (which in turn would be reduced to other bases).

mathematical objects in our statement of them.[62] Combined with a Quinean account of ontological commitment this yields the result that we should believe in the existence of numbers, or sets, or whatever mathematical object can play the relevant role. Opponents of these arguments attempt to show how you can embrace the Quinean claim about ontological commitment but avoid commitment to numbers by showing how to nominalise away references to mathematical objects in our statement of our best scientific theories.[63]

The fan of biological, chemical or sociological kinds makes a similar move to the proponent of the indispensability argument in the philosophy of mathematics. They will claim that there is some perfectly natural phenomenon that we cannot explain without invoking the biological/natural/sociological kind.[64]If that claim is combined with the Quinean thesis of ontological commitment then we will have reason to accept the existence of the relevant biological/chemical/sociological properties.

The Cornell realist exploits a similar move when it comes to moral properties. They argue that there are some naturalistic phenomena that we would not be able to fully explain

[62] There has been a debate about the need for the concept of an 'average number' for the statement of some facts for example – see Melia (1995)

[63] For an attempt to show how science could be given a nominalistic basis see Field (1980). See also Colyvan (2011) for a good summary of these debates and how they have developed.

[64] For an example of an application of this sort of manoeuvre see Beebee and Sabbarton-Leary's (2010) application of Boyd's permissive conception of a natural kind (2010) to psychiatric kinds.

without positing moral properties[65]. For example, the Cornell realist could claim that we need to posit moral properties to explain why people have the moral beliefs that they do.[66]

However, this sort of strategy faces a challenge from Gilbert Harman[67]. Harman (1977) accepts that properties earn their ontological keep by featuring ineliminably in our best explanations of experience. However, he thinks there is a crucial disanalogy between the properties posited by, for example, our best physics, and the moral properties of interest to the Cornell realist. He asks us to compare two cases – a physicist who believes (on the basis of some visual disturbance) that there is a proton in a cloud chamber; and an onlooker watching a gang setting fire to a cat who forms the belief that what the gang is doing is morally wrong. Whilst the best explanation of the physicist's belief that there is a proton in the cloud chamber is that there *is* a proton in the cloud chamber, when it comes to the onlooker's belief that burning the cat is wrong:

[65] For Quine himself, who takes the right canonical notation to be first-order logic, this would be anathema. We would, if taking this Cornell realist line, have to embrace a second-order logic where we quantify over properties.

[66] Given that their strategy relies implicitly on a roughly Quinean view of ontological commitment it could be challenged by someone who believes another account of ontological commitment gives us a more perspicuous treatment of the nature of the world. Thomas Scanlon's (2009) Locke lectures can be considered as an attempt to develop a moral realism within a neo-Carnapian rather than Quinean framework. All parties to the debate I survey accept something like the Quinean framework (at least for the sake of argument) however. In addition, I suspect that Scanlon's account makes moral properties causally inefficacious, which might raise suspicions about how it can avoid the unwelcome consequences of non-naturalism whilst retaining the right to be called a form of moral realism. How would things look using a truth-making account of ontological commitment (of the type suggested by Armstrong 2004)? On that account we should believe in only the entities required to make true the truths we accept. I suspect that because the Cornell realist rejects analytic connections between moral predicates and naturalistic predicates (which may have given them truthmakers for moral truths as an 'ontological free-lunch') the terms of the debate will remain substantially the same.

[67] The outline of this discussion depends heavily on Miller (2003)

> [Y]ou do not seem to need to make assumptions about any moral facts to explain the occurrence of the so-called moral observations I have been talking about. In the moral case, it would seem that you need only make assumptions about the psychology or moral sensibility of the person making the moral observation. (6).

This disanalogy means that although we have reason to commit ourselves to the presence of a proton in the cloud chamber we don't have to believe in the existence of distinctly moral properties.

Sturgeon (1985) argues that there are some phenomena which resist the debunking explanations offered by someone like Harman. For example, he claims that we can explain why Hitler did the things he did by using the fact that he was a morally depraved person; or the increase in opposition to slavery in 19$^{th}$ century America by the fact the slavery at that time was a particularly oppressive institution. In addition, he claims to identify a misstep in Harman's challenge to the moral realist. Harman claims that in the cat example the wrongness of setting fire to a live cat plays no explanatory role in accounting for your belief that it is wrong. One way of testing explanatory relevance is to look at relations of counterfactual dependence. If a factor is explanatorily relevant to some outcome that means, roughly, that in a world where that factor was absent the outcome would not have come about. This is what licenses Harman's claim that the presence of a proton in the cloud chamber is relevant to the physicists belief that there is such a proton – if the proton had not been present then the disturbance in the cloud chamber would not have occurred, meaning that the physicist would not have formed the belief that there is a proton in the cloud chamber. However, Harman can be read as claiming, that this relationship of counterfactual dependence does not hold in the moral case – in a world without moral wrongness, the onlooker would still have judged that burning the cat was wrong. This, though,

misunderstands the Cornell realist's position. They accept that the distribution of moral properties supervenes upon (or is constituted by/multiply realised by) the distribution of non-moral properties. Thus, it is because burning a cat is an act of inflicting unnecessary suffering (a characterisation we could give of the act using non-moral vocabulary) that it is a wrong act. In order for the act of burning a cat to lack the property of wrongness it would also have to lack the non-moral property of being an act of causing unnecessary suffering (and the other properties in the supervenience base of 'being wrong') upon which the moral property supervenes. But, if it did lack *that* property then it's implausible to think that we would judge the act to be morally wrong. So, there is a relationship of counterfactual dependence between the moral belief and the relevant moral property, which means that Harman cannot get his disanalogy.

Harman (1986) responds to this reply by arguing that mere counterfactual dependence is not enough to yield explanatory relevance (and hence ontological rights). This is because a view which claimed that moral properties were counterfactually dependent on non-moral properties, and also claimed that moral properties were causally inert (*moral epiphenomenalism*) underwrites the counterfactual dependency of moral judgements on moral properties but does not make moral properties explanatorily relevant. Sturgeon (1986) replies by arguing that this is only a problem if moral epiphenomenalism were an independently plausible view, which it is not. He seems to claim that the best way to sustain it would be to argue that higher-order properties in general are explanatorily irrelevant, which would have the implausible consequence that all biological, chemical and sociological properties are explanatorily irrelevant.

However Alex Miller (2003) effectively undermines this response to Harman. He points out that Sturgeon began by making the claim that counterfactual dependence of certain phenomena (like moral beliefs or Hitler's actions) on putative moral properties is *sufficient* for the assigning explanatory relevance to those properties. Harman's objection undermines this sufficiency claim because if moral epiphenomenalism is not incoherent then there is a way to get counterfactual dependence without explanatory relevance. Such a view does not appear obviously *incoherent* – if we compare it to a similar view in the philosophy of mind where mental properties are causally linked to physical properties in one direction only[68] we can see this. This view does not get attacked for being incoherent – instead, the main objection to it is that it makes our beliefs and desires causally irrelevant to our actions, which seems strongly counterintuitive, but not incoherent. So:

> It follows that Sturgeon's reply to Harman is implausible as it stands: Sturgeon needs to *add* something to mere counterfactual dependence to get the conclusion that moral properties are genuinely explanatorily relevant. (Miller, 2003, 149)

## 3.34 Program Explanation

What could Sturgeon add to mere counterfactual dependency to bolster the Cornell realist's ontological claims? Miller considers whether the notion of *program explanation* could help. Jackson and Pettit (1990) have argued that if we think that the only way for a factor to be

[68] See type-E dualism in David Chalmers's 2002

causally *relevant* to a phenomenon is for it to be causally *efficacious* with regards to that phenomenon (where causally efficacious properties are those "in virtue of whose instantiation the phenomenon occurs" 108) then it will turn out that many properties that we think are causally relevant are not. Consider the example of a glass vessel cracking when we boil water in it:

> Why did it cack? First answer: because of the temperature of the water. Second answer, in simplified form: because of the momentum of such and such a molecule (group of molecules) in striking such and such a molecular bond in the container surface. The temperature property was efficacious only if the momentum property was efficacious... But the temperature of the water – an aggregate statistic – did not help to produce the momentum of the molecule in the way in which it, if efficacious, helped to produce the cracking... And neither did the temperature combine with the momentum to help in the same sense to produce the cracking: one could have predicted the cracking just from full information about the molecule and the relevant laws. (110)

So, if the only way for a property to be causally relevant was for it to be causally efficacious, then the temperature of the water is causally irrelevant to the breaking of the glass vessel. But surely something has gone wrong here – we want to say that the water's boiling *was* causally relevant when it comes to the container's breaking. This suggests to Jackson and Pettit that there must be another way for a property to be causally relevant for a phenomenon. How does this work?

The basic idea is that a higher-level property (like the water's temperature) can program for the existence of a lower-level property that is causally efficacious (like the momentum of the relevant molecule) in such a way that the higher-level property is still causally relevant:

> Although not efficacious itself, the temperature property was such that its realization ensured that there was an efficacious property in the offing: the property we may presume, involving such and such molecules. The realization of the higher order property did not produce the cracking in the manner of the lower order. But it meant that there would be a suitably efficacious property available, perhaps that involving such and such particular molecules, perhaps one involving others. And so the temperature was causally relevant to the cracking of the glass, under a perfectly relevant sense of relevance, though it was not efficacious. It did not do any work in producing the cracking of the glass – it was perfectly inert – but it had the relevance of ensuring that there would be some property there to exercise the efficacy required... A useful metaphor for describing the role of the property is to say that its realization programs for the appearance of the productive property and, under a certain description, for the event produced. The analogy is with a computer program, which ensures that certain things happen – things satisfying certain descriptions – though all the work of producing those things goes on at a lower, mechanical level. (this is a composition of Jackson and Pettit's (1990, 114) taken from Miller's (2003, 151-2)).

Thus certain higher-properties can be causally relevant[69] because although they themselves are not causally efficacious, they program for the instantiation of lower-level properties

[69] The claim which the Cornell realist then uses to earn ontological rights for moral properties.

which are causally efficacious. What the program explanation does is offer us information that mere *process explanation* (explanation in terms of the causally efficacious properties) does not – in this case that the water is at boiling temperature, and that at relevantly similar possible worlds (ones where the glass is broken by a different molecule to the one mentioned in the process explanation) the glass still breaks. Is there anything stopping the Cornell realist from exploiting this sort of *program explanation* to earn ontological rights for higher-level moral properties which program for lower-level causally efficacious non-moral properties?

Brian Leiter (2001) argues that the type of examples offered by the Cornell realists are either shallow (in relation to the example that cites Hitler's moral depravity to explain his behaviour: "I would take such an answer to be a bit of a joke: a repetition of the datum rather than an explanation." (94)) or are not plausibly our best explanations of the phenomenon in question – in relation to opposition to apartheid in South Africa "we have to turn precisely to the particular lower-order social, economic, and political facts to really explain why social protest arose against racial oppression at the time it actually did." (97).

An interesting point to consider in relation to this is whether commitment to something like Karl Marx's theory of history creates problems for the Cornell realist on this score. According to this theory (at least on Jerry Cohen's (1978) influential reading) cultural institutions arise and succeed when they do because they bring about the development of the productive forces. On this account a cultural, legal 'superstructure' is determined by the functional role it can play in developing the productive forces. This might look like the strongest type of claim that would endanger the Cornell realist's ambitions – we could

explain all (or at least most) social phenomena by adverting to an economic rather than moral explanation. Thus, to return to the example of opposition to slavery, we would explain the abandonment of traditional forms of slavery by explaining how slavery no longer served the expansion of the productive forces and instead in some way 'fettered' them. This sort of framework could be what Leiter has in mind when he claims that we can explain opposition to apartheid in social, economic and political terms. In any event, it looks like the sort of theory in which distinctly moral factors turn out to be causally irrelevant. However, Cohen argues that this sort of deterministic reading of Marx ignores the relationship between Marx's theory of history and his philosophical anthropology. We can ask 'why do the productive forces drive human development (including the development of political, legal institutions, etc)?'. Cohen's answer is that it is part of Marx's anthropology that human beings are, by and large, at least somewhat rational. This is one of the reasons why we develop institutions that serve the expansion of the productive forces – we see that expanding those forces satiates human needs, or even if we do not consciously see that this is the case, we rationally respond to this kind of consideration. This interpretation of Marx is certainly contested, but it does look *prima facie* plausible. Without *something* like the anthropological claim it might look mysterious as to why the cultural superstructure is determined by the economic facts unless we make out Marx to be massively deterministic. If, then, the cultural superstructure emerges when it does and takes its particular form because that form serves expansion of the productive forces and we care about expanding the productive forces because they serve to meet human needs then there is space for an advocate of a Marxist theory of history to make use of explanations using moral properties. Thus, the tension between Cornell realism and a Marxist theory of history may be only *prima facie*. In effect, the Marxist can agree with Miller's contention that by pointing towards legal, economic and political factors as the *best* explanation of a particular

phenomenon Leiter is, in effect, letting program explanation in through the back door (see Miller, 2003 172-3).[70]

A more pressing worry is offered by Miller. That is that program explanations (at least in these contexts) are only the *best* explanations on offer because of our parochial epistemic limitations. What program explanation secures us is information about how things go in relevantly similar possible worlds – we think the *boiling* (a higher-level property) caused the cracking of the class because in close by possible worlds where the particular molecule which is causally efficacious in breaking the glass is missing, the glass still breaks (some other molecule would hit some other molecular bond in the side of the container). Miller's suggests that this modal information (about how things go in similar possible worlds) is only a theoretical boon to creatures who share our epistemic limits. If we imagine a creature without these limits – an omniscient God – we see that such a creature would not suffer explanatory impoverishment if they lacked program explanations[71]. God would have as rich an explanatory theory as us if She just relied on process explanations involving causally efficacious properties and her knowledge of the relevant modal information that program information gives us.

[70] For an exploration of structural explanations of this sort as applied to social theory see Jackson and Pettit's (1993).

[71] Here, this 'God' is being used as merely a heuristic device to militate against using program explanation to earn ontological rights for moral properties. The device is supposed to demonstrate the strangeness of using program explanation to earn ontological rights when the utility of program explanation is based merely on our limited knowledge (of lower-level facts/properties and process explanations).

You might argue that making such a move (stripping properties of their ontological rights because they only feature in *our* parochial best explanations of phenomena) is unwarranted – isn't the point of the Quinean account of ontological commitment to link ontological commitment to precisely *our* best theories? However such an argument would be off-beam. As Miller puts it: "when we are in the business of asking about what properties earn their ontological rights, we should be concerned with what properties would figure in the world as seen from the viewpoint in which all such epistemic limitations [ones due to our particular position] were transcended." (2003, 173). Why? Well, one reason is given by Joe Melia (1995) with reference to indispensability arguments for mathematical objects. A proponent of these arguments could argue from the fact that we need average numbers to represent certain empirical discoveries that we need to admit numbers to our ontology. However, Melia argues that, at least in some of these cases, although our best theories ineliminably refer to average numbers, we can imagine a better theory that eliminates them in favour of lists of objects which encode that information. These lists might be too long for us to grasp. Nevertheless, this averageless theory would be better (more parsimonious) than *our* best theory, so we ought to be able to avoid commitment to average numbers. If this is true, then Miller's point stands as well motivated – a creature not limited by our epistemic position would not need program explanation, so we cannot use it to earn ontological rights for moral properties.

Mark Nelson (2006) accepts Miller's amendment to the realist's task (showing that a creature without our epistemic limits would need to make use of program explanations to explain certain phenomena in order to earn moral properties their ontological rights through this strategy) but reckons it is still one that can be met. Remember, the key point from Miller was that God knows all the modal information (the relevant true counterfactuals) that

program encodes for us. Nelson concedes this point, but argues that even in this position there is something that God lacks: She may know *that* the relevant counterfactuals are true, but She will not know *why* they are true. To return to the breaking glass example, God may know that if a particular molecule (the one that actually did break the glass) did not hit the particular bond at the particular time that it did, the glass would have still broken (another molecule would have hit another bond at a slightly different time); but She does not know why this particular counterfactual is true. In particular, just knowing all the microphysical facts up to the point of time where the glass broke will not tell you why the counterfactual is true, and the relevant causal laws cannot explain it either.

Miller responds (2009) by arguing that higher-level properties will not feature in an account of God's explanation of the truth of the relevant counterfactuals. He does this by first noting that the proposition expressed by a counterfactual sentence is context dependent – it depends both on context of utterance and the speaker's intentions. He borrows the following example from Jonathan Lowe:

> Suppose we are together in a room which we both know to contain a considerable amount of highly inflammable gas owing to a gas leak, and we both observe the presence there of a third person, Brown, concerning whom we know the following facts: first, that he has in his hand a box of dry and perfectly sound matches, and second that he is an extremely cautious individual who is exceptionally sensitive to the presence of gas and strongly averse to risking its ignition by a naked flame (Lowe, 1995, 53)

Now take two counterfactual sentences:

(I) ‘If Brown had struck one of those matches just now, there would have been and explosion’, as uttered by A

(II) ‘If Brown had struck one of those matches just now, there would not have been an explosion’, as uttered by B

Lowe’s point is that as (I) and (II) express different propositions, both can be true at the same time. If A intends to convey something about the relationship between lit matches and gas explosions, or finds herself in a conversation where that relationship is salient, then what she says is true. If B intends to convey something about Brown’s general dispositions about danger, or finds himself in a conversation where that topic is salient, then his utterance is true. This is because of a fact about the evaluation of counterfactuals that Miller draws to our attention – on the Lewis-Stalnaker account of counterfactual conditionals a counterfactual conditional is true if in the closest possible world (or worlds) to this world in which the antecedent is true the consequent is also true. And, which worlds are closest displays a measure of context sensitivity. To evaluate A’s utterance we hold the facts about distribution of gas molecules in the room fixed and vary the facts about Brown’s psychology – we consider worlds where Brown’s psychology is different but the room is still full of gas to be closest. To evaluate B’s utterance we do things the other way round – hold the facts about Brown’s psychology fixed, but suppose that he would not strike the match unless he’d done something to prevent the gas explosion occurring – opening all the windows perhaps. What this means is that which proposition a particular utterance of a counterfactual expresses depends in part upon contextual factors like a speaker’s intentions or conversational salience. (Miller, 2009, 339)

Miller then examines Nelson's argument in light of this consideration and concludes that it misfires. Nelson claims that without program explanations of the form 'The glass cracked because the water was boiling' God will not know why counterfactuals of the form 'If that particular molecule hadn't hit that particular bond at that particular time the glass would have still cracked' are true. In order to know this God would need to know which possible world (or sets of equally close possible worlds) is (are) closest to the actual world. However, Miller contends that God does know precisely this – God knows which proposition any particular utterance of the relevant counterfactual expresses (She is omniscient, after all!). She does not need to invoke higher-level properties to do this. The upshot of all this is that Miller's original argument is left untouched by Nelson's reply – yes, God will need to know why particular counterfactuals are true, but this is something She can do without invoking higher-level properties. Thus Nelson's reply offers no succour to the Cornell realist who hopes to exploit program explanation.

However, I suspect there may be hope for the Cornell realist yet. Miller's argument relies upon the following insight – if God knows which proposition a particular utterance of a counterfactual expresses then She knows why that counterfactual is true (as She knows which is the relevant closest world). However, we can still ask on behalf of Nelson and the Cornell realist – how does God know this? Presumably Miller's thought is that as which worlds are closest is a matter of a speaker's intentions, and God has access to those, She will thus know the relevant closeness ordering of possible worlds. This, though, may be a bit hasty. The defender of program explanation could claim that although there is a role for speakers' intentions and other contextual elements in the closeness ordering of possible worlds that these elements do not totally determine what the relevant similarity measure is. Another way to put the point is that sometimes speakers' intentions will not determine a

uniquely close possible world or set of possible worlds. In light of this, Lewis offers a 'default' measure of similarity of worlds for standard contexts:

1. It is of first importance to avoid big, widespread, diverse violations of [natural] law.
2. It is of the second importance to maximize the spatio-temporal region throughout which perfect match of particular fact prevails.
3. It is of the third importance to avoid even small, localized, simple violations of law.
4. It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly. (Lewis 1979: 47–8)

The glimmer of salvation for the program explanationist comes with criteria 1 and 3. If they can plausibly claim that speakers intentions do not fully determine which similarity measure is relevant and hence which proposition is expressed by a particular utterance of a counterfactual, and that to determine criteria 1 and 3 even God will need information about higher-level properties, then there is something that God cannot know without program explanation – which proposition is expressed by a given counterfactual utterance. This is because what the appropriate similarity measure will be for evaluation of that counterfactual depends in part (according to the program explanationist) on program explanations.

If this is right, then there is hope yet for program explanation – even God needs to use it to access the modal information program explanation encodes for us. Therefore program explanations are no mere artefacts of *our* best theories. They may be found in the best

theories *simpliciter*, in which case Cornell realists can exploit program explanation to earn ontological rights for moral properties.

## 3.35 Cornell Realism's Semantic Programme

So we have seen that there are worries for the Cornell realist about their strategy for earning ontological rights for irreducible, natural, moral properties; although it's not obvious that that strategy fails. Cornell realists also advance a distinctive semantic claim – that we can model the semantics for moral terms on the semantics for natural kind terms originally advanced by Saul Kripke and Hilary Putnam. Here I will examine that semantic claim. I will not be able to settle the question of whether the Kripke-Putnam framework is the right one for natural kinds, all I intend to do is examine whether the Cornell realist can make an analogy between natural kind terms (understood in the Kripke-Putnam manner) and moral terms. This will at most license a conditional claim: if the Kripke-Putnam semantics works for natural kind terms then the Cornell realist can model their semantics of moral terms on it. So, this section will examine whether there is the similarity between the two types of terms that the Cornell realist can exploit. First though, a (very) brief history of the development of the Kripke-Putnam semantics for natural kind terms.

According to a Millian view of meaning the meaning of some term, like a name, just is its referent (or extension in the case of other types of term). So, the meaning of the name 'Venus' is the actual planet Venus. Gottlob Frege opposed this view based, in part, on the following data: it seems like identity statements involving co-referring expressions are cognitively significant. For example, learning that the Morning Star is the Evening Star is a

genuine cognitive advance, even though both names refer to the same object (the planet Venus). One way to tell this is that the identity of the Morning Star with the Evening Star was an empirical discovery of astronomy. Mere competence with the meanings of 'Morning Star' and 'Evening Star' is not enough to know the identity. On the Millian picture this is deeply puzzling – the referent of 'Morning Star' is, like the referent of 'Evening Star', just the planet Venus. So, if I know the meaning of 'Morning Star', which on the Millian picture is to know the referent, then I should be able to work out the identity just through reflection on that meaning.

This observation drove Frege to distinguish between *sense* and *reference*. 'Morning Star' and 'Evening Star' may share the same *referent* (the planet Venus) but they have different *senses*. We then give an account of meaning which gives a central role to sense. On this sort of picture, then, grasp of the meaning of the term 'Morning Star' will not be enough to know the truth of the identity between the Morning Star and the Evening Star, and thus we will have a ready explanation of the cognitive significance of 'The Morning Star is identical to the Evening Star'.

However, what this view needs is an explication of the notion of *sense*. One way to do this is to offer a *descriptivist* semantics. On these views the meaning of an expression is given by a cluster of descriptions associated with that expression. This is easiest to see in the case of names. The proposal is that a name like 'Aristotle' means something like 'the pupil of Plato who taught Alexander the Great and wrote the Nicomachean Ethics...', suitably filled in with the descriptions that speakers associate with the name 'Aristotle'. The meaning of 'Morning Star' could be given as 'the object that is visible in the morning sky at such-and-

such a time in such-and-such a position' whereas the meaning of 'Evening Star' is given as 'the object that is visible in the evening sky at such-and-such and time in such-and-such a position'. So, when you learn that the Morning Star is the Evening Star you learn that the object that is visible in a certain place in the sky in the morning is the very same object that is visible in the sky in the evening – a genuinely substantive discovery. Thus we can use descriptivism to give an explication of the notion of sense which leaves us with a explanation of the cognitive significance of identity statements involving co-referring expressions, seemingly a genuine advance on the Millian picture.

But, the descriptivist does not have it all their own way. Saul Kripke (1980) launches a battery of objections against the descriptivist project, a few of which I will brief outline (my summary of this material will be extremely quick, as I'm just trying to give an indication of what the general framework looks like before seeing how it applies to moral terms. For a more detailed account see Miller 2007 or Soames 2002, 2010). Kripke's modal objection first: if the meaning of 'Aristotle' is given by the description 'the pupil of Plato who taught Alexander the Great and wrote the Nicomachean Ethics...' then 'Aristotle was the pupil of Plato who taught Alexander the Great and wrote the Nicomachean Ethics...' will turn out to true purely in virtue of meaning and thus necessarily true[72]. However, that Aristotle was taught by Plato and taught Alexander is surely a contingent fact – there's certainly nothing incoherent about saying that Alexander could have been taught by someone else, and that there are possible worlds where such a state of affairs obtains. Kripke's diagnosis of this is that proper names are *rigid designators* – they refer to the same object in every possible world in which that object exists. So, the name 'Aristotle' still stands for Aristotle even in

[72] This ignores the possibility, raised earlier in discussion of Gillian Russell's work, of contingent analytic truths. However, that account provides space for contingent analytic truths in very narrow circumstances involving thoroughly indexical terms, unlikely to be applicable here.

worlds where Aristotle does not do the things that we associate with him – so it's no surprise that the *meaning* of Aristotle is not given by a definite description. In addition, as a matter of empirical fact people can very well grasp the meaning of a name whilst associating either incomplete (descriptions that fail to uniquely pick out one object) or false descriptions with that name; and the descriptions each speaker associates with a name can vary. This causes particular problems for a Fregean descriptivist, for the Fregean would like to preserve the claims that sense determines reference, and that sense is also *objective* – something shared by all users of a term.

However, the descriptivist could weaken their claim – instead of saying that the description gives the meaning of the name, they could instead claim that it merely fixes the referent of that name[73]. The idea is that although 'Aristotle' doesn't mean 'the pupil of Plato who taught Alexander the Great and wrote the Nicomachean Ethics...', the reference of the expression 'Aristotle' is fixed by that description. At this point Kripke presents a number of cases where our intuitions seem to indicate that this weaker descriptivist thesis is still false. Consider the name 'Gödel'. The only thing that many speakers know about Gödel (and thus the descriptive content that they associate with that name) is that he proved the two incompleteness theorems. But, Kripke asks, suppose that Gödel *didn't* prove those theorems – instead, he stole the proofs from another mathematician, Schmidt. Now, when you use the name 'Gödel', who are you referring to? Most of us[74] have the intuition that we would not

[73] We could explicate the meaning/reference-fixing distinction modally. The descriptivist who claims the relevant description fixes the referent of then name will accept a clause like 'X is the referent of Aristotle iff X is D' (where D is an abbreviation of the relevant description). They just will reject appending a necessity operator to that claim, as then the clause would have the wrong modal profile – it would make it impossible for Aristotle to have failed to do the things in the description.

[74] At least, if the experimental philosophers are to be believed, most western philosophy undergraduates. See Machery et al (2004) and (2009) for putative evidence that East-Asian philosophy undergraduates have intuitions broadly consistent with a descriptivist story about reference. However,

be referring to Schmidt – the one who actually proved the incompleteness theorems – but instead to Gödel. This shows that 'the prover of the incompleteness theorems' does not even play a mere reference-fixing role for 'Gödel'.[75]

What's Kripke's alternative picture? Instead of descriptions fixing references he thinks this job is done by baptisms and facts about causal chains linking our use of a word back to its baptism. To put it as loosely as possible, we point at some object and agree to use a certain term to stand for that object. Competence with that term is then gained by being a part of a causal-historical chain of overlapping intentions that lead back to the initial baptism. So, in the case of Aristotle: he was named by someone and at that time the linguistic community agreed to use 'Aristotle' to stand for that person. Other speakers get to refer to Aristotle because they intend to use the term 'Aristotle' to refer to whoever was the referent of the initial baptism of that term[76]: they in a way piggyback on the referential intentions of the person from whom they learned the term, all the way (hopefully) back to Aristotle himself.

---

there has been some scepticism about this type of evidence expressed in Ichikawa et al (2011) and Kauppinen (2007). Even if this data is right, however, the Kripke framework might survive unscathed – Kripke is willing to admit that we could have spoken a descriptivist language and even within English as we speak it some terms may be open to a descriptivist analysis. Kripke claims he is not advancing a new general theory of his own; instead he is casting doubt on the descriptivist's ambition to give a universal treatment of language by showing how it fails in many cases.

[75] Although see Lewis (1984) for the suggestion that all these cases show is that we need to refine the descriptive content we offer as a reference-fixer. However, Lewis agrees with the Kripkean claim that a *general* descriptivism is untenable, although for different reasons and his distrust of general descriptivism is much more limited – he still thinks it's viable to give a descriptivist treatment of almost all language, a conclusion Kripke would demur from.

[76] This is an extremely simplified account, and cannot be right as it stands – it faces problems with cases where the referent of a term changes over time like in the case of 'Madagascar' (see Evans 1973) and in other situations. However, the fan of the Kripke-Putnam framework hopes that they will be able to finesse the account to deal with these putative counterexamples. In addition, some argue that if we adopt this Kripkean account we are forced back into offering a Millian account of meaning, one which still struggles with the puzzles of cognitive significance that inspired Frege, although see Salmon (1986)

So much for names, what about natural kind terms like 'water', 'gold', 'tiger', 'acid' and 'catalyst'? Well, we can adduce similar types of worries about descriptivism with these types of terms too. Hilary Putnam (1975) famously offers us the example of Twin-Earth. Twin-Earth is a distant planet, exactly like Earth except in one respect: the clear, odourless, potable stuff that falls from the sky and fills the lakes, taps and rivers does not have the chemical composition $H_2O$, but some other chemical composition – call it XyZ. If the descriptions we associate with water (that it is clear, odourless, potable stuff that fills lakes, taps and rivers etc) gave the meaning of, or fixed the reference of our term 'water' then our term 'water' would apply to XyZ just as well as it does to $H_2O$. However, Putnam contends that this is unintuitive – our term 'water' does not refer to XyZ, just as in the Twin-Earthers' mouths their term 'Water' does not refer to H2O. This is because, like proper names, the reference of our word 'water' is fixed by some initial baptism that incorporates an intention to refer with 'water' to the stuff that shares an intrinsic nature with the stuff that is dominant causally responsible for our water perceptions – the ones that form what we can call water's 'nominal essence' – that water is wet, potable, falls from the sky, fills lakes, etc. This explains why, as Kripke points out, the term 'Gold' does not refer to fools good – although fool's gold satisfies the superficial descriptions we associate with gold (it's yellow, soft, etc) it differs in atomic chemistry from gold – it does not have atomic number 79. When we introduced the term 'gold' we intended to use it to refer to the stuff that shares an intrinsic nature with the stuff we were baptising, and so anything that lacks that intrinsic nature fails to count as gold. Similar examples abound – orange and black stripey cats are not tigers if internally they are actually robots instead of being biological organisms that

---

and Soames (2002) for attempts to deal with this worry. One move open to the Kripkean who takes up the Millian line is to claim that 'Morning Star' and 'Evening Star' do have the same meaning (as they share the same referent), but that mere competence with applying the term does not mean you grasp that meaning, which explains why you can falsely believe (as a matter of the meaning of the terms involved!) that the Morning Star is distinct from the Evening Star. This is, of course, complete anathema to the traditional Fregean picture where meanings are what competent speakers grasp.

share the morphological, physiological and genetic properties characteristic of actual tigers. One consequence of this view is that if a Twin-Earther chemist got into an argument with an Earther chemist about whether a sample of XyZ was 'water' they would be talking at cross purposes – the term has a different meaning in each of their mouths (even though they might associate all the same *descriptions* with 'water').

What we are left with is a picture where for both names and natural kind terms reference is fixed and maintained via initial baptisms in concert with overlapping chains of referential intentions. What use is this to the Cornell realist?

Well, Richard Boyd argues that we need to be more permissive than an orthodox Kripkean in our account of natural kinds[77]. Natural kinds are, for him, those terms which track the properties which inform our explanatory inferences. What does this mean? Well the idea, briefly, is that there are causal structures in the world which natural kind terms latch on to. These causal structures allow us to explain and predict various events. In some cases these will involve the sharing on one particular property among all samples of a given kind: as in Kripke's gold case, where all gold has the atomic number 79[78], and having atomic number 79 thus characterises gold's intrinsic nature and informs us about its essence. We can then use that essence in our scientific practices – in explanations of why gold is how it is, and how it interacts with other substances. However, things are not this simple for other natural

[77] Boyd (1991, 2010); for problems with the pure Kripkean account of natural kinds see also Beebee and Sabbarton-Leary (2010); for an account similar to Boyd's see Dupré (1993).

[78] In actual fact, the case of chemical kinds is considerably more complicated than this, but let's not get too worried about that – all we need for present purposes is a contrast with cases where it is even more complicated

kind terms in good standing, like biological or social kinds. For instance, there is no one genetic, morphological or physiological property that all tigers share in common. However the category 'tiger' does underwrite our explanatory practices, and maps on to genuine causal structures in the world – it's just that these causal structures are much looser than in the case of chemical kinds. What we have instead are what Boyd calls 'homeostatic property clusters'. These are a number of properties that tend to cluster around one another in instances of the kind in question – so most tigers have four legs, orange stripes, certain portions of genetic code, etc, and this is enough to group tigers as a kind, even though there is no one property that all tigers share in common, nor even one single mechanism which is responsible for tigers sharing some of the features that are common for their type. 'Tiger' is still a natural kind term because the clustering of this group of properties is enough to underwrite our classificatory and explanatory inferences – we can still use 'tiger' to license certain (*ceteris paribus*) generalisations about tigers and so on. We are still broadly within the Kripke-Putnam framework, however, because it is these properties which are causally responsible for our tiger perceptions – the homeostatic property cluster basically, in a loose way, gives us the intrinsic nature of tigerhood that our initial baptism of 'tiger' attempted to latch on to. This explains why the Cornell realist was so keen to find a role for moral terms in our best explanatory practices – it is featuring in those practices which give a term natural kind status, which allows the Cornell realist to pursue their analogy between moral terms and natural kind terms.

The Cornell realist's suggestion is that moral terms are like natural kind terms in this permissive sense. They are, as in Boyd's (1991) homeostatic property clusters that underwrite our moral practices. So, for example, 'good' refers to that cluster of properties that is dominantly causally responsible for various aspects of our moral practices – it plays a

central role in how we regulate our affairs, we typically respond to judgements that goodness is instantiated in some state of affairs by being motivated to pursue that state of affairs, and so on. What's the upshot of all this? Well, natural kind terms, on this sort of view, do not have to be reductively defined in terms of other vocabulary to earn their ontological keep. This means the Cornell realist does not fall prey to the open question argument – they never claim that 'good' or 'right' can be analytically defined in other terms. Nevertheless, goodness is still perfectly real, and perfectly natural. We are not forced down the road of offering an anti-realist or non-naturalist account of morality. The Cornell realist is also an externalist about moral motivation – they can quite coherently claim that we only typically respond to judgements about goodness by being motivated towards that goodness, for example. This means that if we read Korsgaard's attack on moral realism as an attack centring on the motivational effects of moral judgements, then her attack simply fails to get a grip on the Cornell realist, who has the added advantage of offering us a semantics seemingly commensurate with the discoveries of late 20th century philosophy of language.

In addition, this semantic thesis, if tenable, has implications for Finlay's analytic reductivism. Remember, Finlay wanted to give reductive analyses of normative terms in non-normative vocabulary, and he supported this practice of analysis by adverting to Gillian Russell's defence of analyticity. However, on this Kripke-Putnam picture moral terms would either not be suitable components of the kind of analytic truths Finlay wants to generate, or, if they were they would not be the kind of truths that would be accessible through the armchair speculation Finlay takes up as his method. This is because, to put things in Russell's terminology, natural kind terms have reference-determiners that are sensitive to context of introduction – roughly, if we had been pointing at XyZ rather than

$H_2O$ when we baptised 'water', our word 'water' would have referred to XyZ, not $H_2O$. However, the term '$H_2O$' is not sensitive in the same way. This is what makes it the case that 'Water is $H_2O$' is not analytic, though necessary, because it is impossible for the right relations of inclusion, exclusion or identity to hold between the reference-determiners of 'water' and '$H_2O$' given their different sensitivity to contexts of introduction.

Suppose though that we, rather implausibly in my view, dropped this aspect of Russell's view of natural kind terms and allowed that they could have reference-determiners insensitive to contexts of introduction (this would give bizarre results, but it would mean that there could be the analytic connections between normative and non-normative terms Finlay wants). Even then, this would not license Finlay's *method*. This is because it is part of the more general Kripke-Putnam framework that one can be competent with a term without knowing much, if anything, about its reference determiner – for hundreds of years we knew the meaning of 'water' without knowing anything about its chemical composition; in actual fact whilst falsely believing it was a simple substance. So, regardless of the details of the account of analyticity that Finlay adverts to if anything like the Cornell realist's semantic thesis (that moral kind terms are analogous in their function to Kripke-Putnam style natural kind terms) is right then the project of analytic reductivism is sunk.

So, is the Cornell realist's semantic claim plausible? Finlay himself makes two broad moves against this picture of moral language – first he challenges the general framework on a number of grounds; and also its application to the moral case. I will just briefly outline this second move before moving on to another objection that has played a significant role in recent literature. Finlay argues that, because we have *a priori* access to significant moral

truths, normative concepts must be 'metaphysically thin' and therefore 'transparent' to us. What he means by this is that normative properties are not concrete in the way that tigers and gold are. This is supported by the claim that normative properties can not only be predicated of concrete objects (like tools, people, actions) but also abstract things (like plans and ideas). So, there are two bits of evidence that moral concepts and the terms that stand for them are unlike natural kind terms – they can be reasonably applied to abstract objects and we know truths involving them *a priori*. The Cornell realist is unlikely to be moved by the second sort of consideration – they will simply deny that these putative *a priori* truths are genuinely known *a priori*. After all, for them, to even know whether a term like 'good' refers to a property at all you need to see whether it has an ineliminable role in our best explanations of various phenomena - a thoroughly empirical pursuit. The other worry may be more troubling but again the Cornell realist has avenues to explore – they could deny that things like plans and ideas are abstract, instead identifying them with certain types of concrete objects (perhaps via identifying them with types of mental states which are themselves concrete objects), or they could hold that the application of moral terms to abstract objects is either false or merely metaphorical.

A more particular argument against the Cornell realist's semantic programme is found in Terry Horgan and Mark Timmons attempt to launch a revised open question argument against it[79]. The Cornell realist eschews the analytic definitions of moral terms advanced by an analytic naturalist like Finlay, but they do hope to offer synthetic definitions which characterise the intrinsic nature (or 'real essence') of moral properties. What might this property be like? In order to generate their argument against Cornell realism Horgan and

[79] Horgan and Timmons 1991, 1992, 1996, 2000; and see 2009 for a structurally similar attack on analytic functionalism.

Timmons combine Boyd's claim that moral terms are, like natural kind terms, rigid-designators that refer to natural properties with David Brink's view that the properties in question are functional properties whose intrinsic nature is revealed by our best normative theory. This yields a thesis they label 'Causal Semantic Functionalism' (**CSF**)

> **CSF:** Each moral term *m* is causally regulated by a unique functional property, and rigidly designates that property.[80] (Horgan and Timmons 1996, 12)

It's this claim that allows Horgan and Timmons to generate problems for the Cornell realist using a modified Twin-Earth scenario.

Suppose, as a matter of empirical fact, that the mature normative theory that Earthers converge upon is consequentialist in nature, and thus (with **CSF** and Brink and Boyd's other semantic commitments) our word 'good' rigidly designates this consequentialist property. This property is what our term 'good' rigidly designates. Now consider Moral Twin-Earth. On Moral Twin-Earth the mature normative theory that Twin-Earthers converge on is deontological in nature – the property that their word orthographically similar word 'good' rigidly designates in deontological. However, this property is connected up with Twin-Earthers lives in the way goodness is connected up with ours – Twin-Earthers are typically motivated by judgements involving goodness, judgements of goodness are thought to have special bearing on Twin-Earthers' well-being, etc. Importantly, the similarities between the practices connected with their use of the word

[80] We get the causal regulation part from the fact that natural kind like terms refer to whatever is dominantly causally responsible for our perceptions of the superficial properties caused by the properties denoted by those terms. Note that this formulation talks about a unique property, whereas Boyd's mature view talks of homeostatic property *clusters*. To explicate Horgan and Timmons's argument it should be harmless to leave this complication out of the picture.

‘good’ and our use of our word ‘good’ are strong enough that they would feel inclined to translate their word with our word.

Now, Horgan and Timmons claim that our intuitions about Moral Twin-Earth diverge wildly from our intuitions about the Twin-Earth scenarios offered in favour of the Kripke-Putnam semantic framework. Remember, in the original ‘water’ case we feel as if the following sort of dispute involves people talking at crossed purposes: a Twin-Earth chemist and an Earth chemist pointing at a sample of XyZ and arguing about whether it is a sample of ‘water’. The reason why this dispute is pointless is that the term ‘water’ has a different reference in each of their mouths – for the Twin-Earth chemist it refers to stuff iff it is XyZ, but for the Earther it refers to H2O. We might even characterise their dispute as merely verbal – if they knew more about the semantic practices of their respective languages they would give up the dispute.

Now imagine a parallel disagreement about a moral term like ‘good’ – suppose some state of affairs possesses the deontological property that causally regulates a Twin-Earther’s word ‘good’ but fails to possess the consequentialist property that performs the same role for the Earther. One ostends the state of affairs and says ‘This is good’ whilst the other denies it. Do we feel that this dispute is, again, really a waste of time? We could certainly offer that interpretation of what’s going on here – that in their mouths ‘good’ simply means/has its reference fixed by different things, so really their dispute is merely verbal. However, Horgan and Timmons claim that ‘the far more natural and plausible mode of description’ of this dispute is that ‘moral and twin-moral terms do not differ in meaning or reference, and hence any apparent moral disagreements that might arise between Earthers

and Twin Earthers would be *genuine* disagreements – i.e. disagreements in moral belief and in normative moral theory, rather than differences in meaning' (2009, 8).

We can explicate this another way to show up the similarities to the old open question argument.[81] The Cornell realist accepts something like **CSF** for above, and so for each moral term they will offer a (synthetic) semantic analysis like this (for good):

> **GOOD:** x is good iff it instantiates the property that causally regulates our moral talk.

This, according to Horgan and Timmons, tells us something about the meaning of good. Thus, the following question: 'x instantiates the property that causally regulates our moral talk, but is it really good?' should be closed. Reflection on the Moral Twin-Earth thought experiment is supposed to show that it is not. This is because we can imagine ourselves in a genuine disagreement with the Twin-Earthers about this.

As presented so far this revised open-question argument might appear to beg the question against the naturalist just like the old. As Miller puts it "the...argument works only if our conviction that the question is open is well-grounded, or, equivalently, that our intuitions about the moral Twin-Earth case are correct. But to make either of these assumptions is already to presuppose the falsity of the idea that the semantics of 'good' is given by the likes of **GOOD**." (2003, 167). Miller suggests that Horgan and Timmons should instead model their argument on the revised open question argument that we saw from Darwall, Gibbard and Railton (1992) in §2.14. There, we don't start from the datum that our intuition

[81] They also offer a revised argument from queerness (1992, 248).

*is* well-grounded. Instead we start from the less controversial claim that we have the intuition in question, then proceed to argue that the best explanation for our intuition that the question is open, or that our dispute with the Twin-Earther is substantive, is that the intuition is well-grounded. Horgan and Timmons explicitly endorse this strategy in their 2009 (5).

They also have some motivation for making the inference from 'we have this intuition' to 'this intuition is well-grounded', for two reasons – the seeming strength of the intuition, and the dialectical environment they offer it in. On the first, it seems like this intuition is one that might be strongly held. One way to flesh this out is to think how things would go if we leapt into our spaceship and went off to meet the Twin-Earthers. It would be of fairly large importance to us which definition of goodness governed those interactions, and we would try to argue the Twin-Earthers around to our normative theory, in contrast to the 'water' case where we would take a far more relaxed attitude. On the second reason, remember that the Cornell realist (or anyone who models moral semantics on natural kind semantics in the Kripke-Putnam mould) heavily relies upon intuitions about Twin-Earth cases to motivate their semantic programme – it is the intuition that the dispute with the Twin-Earther about 'water' is trivial that the picture of natural kind terms as rigid-designators that have their reference fixed by causal-historical chains back to particular type of baptisms is supposed to offer a non-debunking explanation of. So, if it is not only acceptable but a vital part of the Cornell realist's general semantic commitments that our intuitions about Twin-Earth cases deserve respect, then it seems improper for them to dismiss this intuition about Twin-Earth cases out of hand – there seems to be a positive presumption in favour of Horgan and Timmons's claim that the Cornell realist needs to undermine.

Horgan and Timmons's argument has come in for a large amount of criticism.[82] The point I wish to make is that Horgan and Timmons's presentation of their argument is misleading, and that when this misleadingness is noted it loses its power. Remember that what is crucial for Horgan and Timmons's new open question argument is the difference between how we respond to the water Twin-Earth case and the Moral Twin-Earth case. In the case of 'water' the question 'This sample is made up of the stuff, $H_2O$, that causally regulates our use of 'water', but is it really water?' (**Qwater**) is closed, whereas the question 'This state of affairs instantiates the property, the consequentialist one, that causally regulates our use of 'good', but is it really good?' (**Qgood)** is open. This is supposed to show that the natural kind semantics offered by the Kripke-Putnam framework is adequate for water, but cannot account for moral terms.

However, once we examine more closely the idea of a question being 'closed' this problem dissolves. A question is closed when the answer is available to someone based merely on their competence with the meaning of the expressions involved. This is why the original open question argument militates against analytic naturalism – if there were analytic definitions of moral terms in non-normative vocabulary available these definitions would be true in virtue of meaning, and mere grasp of the relevant meanings would be enough to know they are true. Mere understanding of the meaning of the terms does not suffice to

[82] See, for example: Gert (2006) who argues that the Horgan and Timmons's style of argument, even if it has force against the synthetic definitions of 'good' offered by the Cornell realist, has no force against synthetic definitions of more basic normative notions, like harm, out of which we could construct better synthetic definitions of goodness; Sayre-McCord (1997), who claims that the Twin-Earther's word 'good' will be regulated by whichever normative moral theory is correct, regardless of what property the Twin-Earthers or best scientific theory *thinks* causally regulates their moral talk; van Roojen (2006) argues that Horgan and Timmons's argument has some force against Boyd's account, but this evaporates when we modify it slightly; Copp (2000) attempts to offer an alternative explanation of our readiness to translate Twin-Earthers' 'good' as our 'good', though see Horgan and Timmons (2000) for a response.

know that these definitions are true however, says the Moorean, so the definitions are not true. The crucial point is that if this is what it means for a question to be closed *neither* **Qwater** or **Qgood** should be closed on the Kripke-Putnam framework. This is because mere competence with the terms involved does not suffice to know the definitions that they are bringing into question. Why is this? Well even if we understand all the relevant terms we will not know, *a priori*, that, for example, that the stuff that causally regulates our 'water' talk is water. This is because the Kripke-Putnam framework that the Cornell realist exploits is supposed to be a substantive, *a posteriori* discovery. Mere reflection on our concept of 'meaning' or 'reference-fixing' is not enough to know its truth – instead we offer the framework as the best explanation of the various intuitions we have about the cases.[83] This means that it should be perfectly possible for someone to doubt whether a sample of the substance that is made up of stuff that causally regulates our 'water' talk is really water – because they don't know, merely on the basis of their semantic competence, that 'water' refers to whatever causally regulates our water talk. What this means is that both **Qwater** and **Qgood** should be open, in the relevant sense. This means that Horgan and Timmons's claim that the semantics of moral terms cannot be given a Kripke-Putnam treatment whilst natural kinds can be collapses. If the disanalogy they point to between **Qwater** and **Qgood** does exist then their conclusion should be that 'good' can be given the Kripke-Putnam treatment, but 'water' cannot, because to give a Kripke-Putnam treatment the relevant question should be open, and it is only in the case of goodness. Clearly something has gone wrong here.

[83] Another way to put the point is that according to Horgan and Timmons the semantic analysis **Good** above is, on the Kripke-Putnam framework, analytic, thus moving the problematic analyticity to a different place, and giving Horgan and Timmons space to run a revised open question argument. However, this misunderstands the Cornell realist's semantic framework. **Good** above is not analytic. Instead, it is a theoretical postulate in an empirical theory of meaning. The justification for believing it is not that it follows from mere semantic competence with "good", instead positing it helps in the construction of the best explanation for the intuitions speakers have about particular cases.

The mistake Horgan and Timmons make is to mischaracterise the evidence that Moral Twin-Earth opens up to us. Instead of characterising our response to **Qwater** or **Qgood** in terms of their openness (because if the Kripke-Putnam framework is right for 'water' neither should be closed) we should instead talk about their obviousness[84]. The answer to **Qwater** does seem obvious, and this might lead us to conclude that the question is closed. But, it could just be obvious because of our extensive knowledge of the chemical composition of water – we know that it is $H_2O$. However, in the case of goodness, what the right normative theory is is not at all obvious. So to be told that some state of affairs instantiates some consequentialist property that causally regulates our 'good' talk does not make the answer to **Qgood** obvious – because we are not sure that the consequentialist story is true.[85] If we are told that it is the consequentialist property that causally regulates our 'good' talk, our uncertainty about that consequentialist theory will threaten our acceptance of a semantic theory that links the meaning or reference of 'good' to the stuff that causally regulates our 'good' talk. Thus the answer to the question will not seem obvious, but not because it is open (in the sense of not being decided by semantic competence) rather than closed. True, it is open in this sense (because the Kripke-Putnam framework is a substantial *a posteriori* discovery) but then so is **Qwater**. Regardless of whether this *is* the correct explanation of our different intuitions what is clear is that Horgan and Timmons's position

[84] What I am in effect claiming here is that consideration of 'openness' when glossed in the typical way will have to be given a different treatment on a Kripke-Putnam framework. On that framework, precisely because something like Putnam's slogan 'meanings aint in the head' is apposite, lots of true explications of meaning will not leave the relevant questions closed. Instead, what the Kripke-Putnam theorist needs to do is to give an alternative explanation of our intuitions that certain questions are closed and others open. Talking in terms of obviousness is one way of doing this.

[85] This factor also explains why there is a difference in intuitions about when we are talking at cross-purposes. In the water case we feel as if we are talking at cross purposes, but not in the good case as set up by Horgan and Timmons. This is because what the right moral theory is is still up for grabs, and we'd treat the twin-earthers, until the point at which we've exhausted all possible avenues of debate, as interlocutors in the discussion of which moral theory is correct, rather than assume at the outset that they mean something different by 'good'.

is untenable. They aim to draw a contrast between the original Twin-Earth case (where the Kripke-Putnam framework is appropriate) and Moral Twin-Earth (where it is not) and this cannot be done in the way they adopt.

To sum up, we began looking at Cornell realism to see whether there was a version of moral realism that could evade Korsgaard's attack on moral realism (when this is construed as being an issue to do with moral realism). Cornell realism can resist this attack whilst evading the original open-question argument and offering the possibility of a plausible semantics of moral terms. We have seen how the Cornell realist's strategy for earning moral properties their ontological keep is threatened by an argument due to Harman, but that there might be hope for this aspect of the Cornell realist if they make certain moves invoking program explanation. We've also seen how the Cornell realist's semantic programme can resist the problems presented by Moral Twin-Earth. This means that we have a position that can both avoid the neo-Kantian attack on realism (read one way) and is independently plausible. Recall also that the fate of Finlay's analytic reductivist project rested, in part, on the demise of the attempt to model moral semantics on the Kripke-Putnam semantics for natural kind terms. We've seen that perhaps the Cornell realist is not in as much trouble as is sometimes thought.

However, in terms of the engagement with the neo-Kantian's attack on realism, this issue is less important. If there is some compelling reason to reject a Kripke-Putnam treatment of moral terms, then this means that Finlay's analytic reductivism becomes more motivated. If moral terms are given a natural kind style semantics, then the Cornell realist looks in better shape. But *both* positions avoid the neo-Kantian objection. To make the objection stick, the

neo-Kantian needs an argument that rules out both of these positions as untenable, which we have seen they are not. The neo-Kantian's case for a new metaethics instead rests on whether their own account offers some other advantages over these realist accounts, an investigation we will get to in chapter five. First we turn to the neo-Kantian's engagement with non-cognitivism.

# CHAPTER FOUR: THE NEO-KANTIAN AND EXPRESSIVISM

We have seen in the last chapter how Korsgaard advances an argument against moral realism, and how the moral realist can avoid that argument in a number of ways. The Neo-Kantian also presents themselves as offering a distinct metaethical position from that offered by expressivists. In this chapter I will first briefly outline expressivism (§4.1); then present Korsgaard's objection to expressivism (along with some scepticism about whether there is a genuine objection here owing to Hussain and Shah, §4.2); I will link this objection up with one of the important challenges facing expressivism – the Frege-Geach problem - (§4.3) and consider traditional expressivist solutions to this problem (§4.31-4); these attempted solutions are unsatisfactory, I will argue, and motivate an examination of so-called hybrid metaethical theories – accounts that combine elements of cognitive and non-cognitive semantics to exploit each position's strengths (§4.4-5); I will argue that these hybrid accounts fail on their own terms – they cannot provide a solution to the problems plaguing expressivism whilst leaving the rest of the terms of metaethical debate standing; finally I will look at what we can say about the Neo-Kantian's reaction to expressivism based on these considerations (§4.6).

## 4.1 Expressivism

Expressivism is a non-descriptive semantic project[86]. It holds that "to make a normative judgement is to express a non-cognitive attitude" (Gibbard, 1990, 84)[87]. This contrasts with cognitivism which holds that normative judgements are entirely descriptive, and that to make a normative judgement is to express a belief. Moral realism is a subset of cognitivism – the realist holds that normative judgements express beliefs, evaluable in terms of truth or falsity, and – sometimes at least – those beliefs are true: there are normative facts 'out there' in the world for our normative practices to attempt to latch on to. It is also possible to be a cognitivist non-realist: you could hold that moral judgements express beliefs, but that those beliefs are systematically false, a la the error theorist[88]; or, you could claim that moral judgements express beliefs, which are true when their truth conditions are met, but give an anti-realist account of those truth conditions (a position I will offer as part of the effort to get clear on the Neo-Kantian's own metaethical position in the next chapter). What's important here is the distinctive expressivist claim – that we can give an adequate semantics for moral terms using non-cognitive, desire-like, attitudes.

[86] Non-descriptive in the sense that it doesn't use beliefs or truth-conditions to characterise the meaning of moral terms. In another sense, expressivism is entirely descriptive – it attempts to give a descriptive rather than revisionary characterisation of every day moral thought and talk.

[87] This is one way of carving up the terrain. Instead, one could reserve 'expressivism' for the view that normative judgements are to be given an analysis in terms of the mental states that they express. This has the side-effect of meaning that for anyone who embraces a broadly Lockean picture of semantics (where the meaning of terms is given by the mental states they express, even in the case of declarative sentences where the relevant mental states are beliefs) even a realist, who holds that normative judgements express beliefs which are in turn straightforwardly truth-evaluable, will turn out to be an expressivist. This is not how the term 'expressivist' has typically been used, and I will use it to mean a view which combines something like the Lockean picture with a non-cognitive characterisation of the content of normative terms.

[88] Mackie (1977), Joyce (2001), and fictionalist analyses will fall into this position if they are revolutionary rather than hermeneutic, see Hussain's (2010) for the distinction.

What are the motivations for this project? Expressivists are typically seen as following in the footsteps of the emotivist analysis of moral terms offered by A.J. Ayer (1936).[89] Ayer in effect took up Moore's open-question argument against the naturalist, but (in part because of his verificationist commitments) also rejected Moore's non-naturalism. The non-naturalist involves herself in the sort of metaphysical pronouncements Ayer rejected as meaningless verbiage.[90] But, if both naturalism and non-naturalism are ruled out what do we have left? Ayer claims that moral language is not literally significant at all. Instead, when we utter a sentence like 'You acted wrongly in stealing that money' we have not literally asserted anything beyond 'You stole that money'. What the moral term does is express our disapproval of stealing. Thus, we should translate utterances like 'Murder is wrong' as being akin to 'Boo to murder!' This latter utterance does not assert anything about murder (that it instantiates the natural or non-natural property of wrongness as the naturalist and non-naturalist realist have it), in the same way that commands, cheers, and so on do not describe the world. Thus it is important to distinguish between emotivism (where moral judgements express approval or disapproval) and subjectivism (where moral judgements report that the subject concerned has the attitude of approval or disapproval in question). The difference is that the latter view makes moral judgement a matter of (a particularly self-centred) belief, straightforwardly apt to be evaluated in terms of truth and falsity. The emotivist, in contrast, holds that people don't *report* their non-cognitive attitudes of approval/disapproval, they *express* them.[91]

[89] Though see Suikannen (2007) and comments for details on earlier progenitors of this type of view

[90] To put the point less polemically, for the non-naturalist moral judgements are synthetic but not empirically verifiable, and thus they violate the positivist's criterion of literal significance.

[91] Although see Jackson and Pettit (1998) for an attempt collapse this distinction.

Modern day expressivists typically lack the verificationist commitments that forced Ayer to reject non-naturalist realism. However, they have other motivations for not taking it up: non-naturalism incurs heavyweight metaphysical and epistemological commitments – we need to add strange, non-natural, properties to our ontology; and once we do that we will need some epistemology for how we find out about the instantiation of those properties (if we are to avoid full-blown moral scepticism). If we can give an expressivist treatment of moral discourse, then we will have traded some semantic complications for some metaphysical and epistemological solvency – the expressivist requires an ontology of only natural properties; and we might think that because moral judgements are accounted for as expressions of mental states we will not need a special epistemology for them. As far as naturalism is concerned an expressivist is particularly moved by something like the revised open-question argument offered by Darwall, Gibbard and Railton (1992). The expressivist places great weight on the practical role that moral judgements have in our deliberation – in effect, they endorse some form of judgement internalism. Thus, they have a quick argument against most forms of realism:

(1) Moral judgements are inherently motivating (Internalism)

(2) Beliefs are not inherently motivating, and have no necessary connections to desire-like motivational states. (Humeanism about motivation)

(3) Therefore, moral judgements do not express beliefs. Instead they express non-cognitive desire-like states.

The conclusion **3** is incompatible with the cognitivist semantic commitments of moral realism. We can say two things about the non-naturalist: either we can see them as falling prey to the above argument (this is the sense in which even Moorean non-naturalism falls prey to the revised open question argument from Darwall, Gibbard and Railton, which uses judgement internalism in a slightly subtler way than the quick argument above), or we can see their metaphysical position as offering a way to deny **2 –** they could claim that the properties that moral judgements are beliefs about are so special that they violate the usual Humean restrictions. Just judging that they are instantiated is enough to move someone to action. But this will mean that the expressivist's earlier charge that the non-naturalist incurs significant metaphysical costs will start to have even more bite – we don't just have non-natural properties on the scene, but non-natural properties with special connections to motivation. The expressivist then offers the realist a dilemma: either they account for moral judgement entirely in terms of naturalistically respectable properties, in which case they fail to explain the special motivational effects of moral judgements; or they start to invoke heavyweight metaphysical costs. This is what prompts the thought that it is really non-cognitivism that benefits from the open question argument.

Modern day expressivists differ from their emotivist ancestors in another respect. They have noticed that moral judgements behave a lot like ordinary declarative sentences. We not only say that 'murder is wrong' we also say things like 'it is true that murder is wrong', or 'murder is *really* wrong' or 'it's a fact that murder is wrong' or 'I believe that murder is wrong'. All of these utterances seem perfectly appropriate but they all seem in tension with the simplistic emotivist analysis of moral terms. Expressivists like Blackburn (1998) and Gibbard (1990, 2003) therefore engage themselves in the project of 'quasi-realism' – the idea is to capture all the realist sounding things that we want to say using moral terms, but

starting from an expressivist starting point. It is Blackburn's hope that we will get to the point where we can perfectly sensibly talk about moral *truth* or moral *belief*, even though we start with the resources available to an expressivist – in short, we can construct those realist sounding notions out of the expressivist starting point plus a revision in what we (philosophers) think about the relevant notion of truth, or belief, for example. Thus the contemporary expressivist has ambitions to save the realist seeming appearances of moral discourse.[92]

## 4.2 The Neo-Kantian Rejection of Expressivism

So, what is Korsgaard's problem with expressivism? Pinning this down seems to be as hard as getting clear on her problem with realism. In her *Realism and Constructivism in Twentieth-Century Moral Philosophy* she presents a barrage of complaints: expressivism does not leave a suitable place for Kantian moral philosophy; we gain nothing by introducing a non-cognitivist semantics; the cognitivist/non-cognitivist distinction depends upon a misleading picture of moral concepts and knowledge (I borrow this list from Hussain and Shah's 2006b). The verdict Korsgaard offers is also quite mixed – she claims that, in a way, even if expressivism were true, it would be 'boring'. So, we have two difficulties: first getting clear on what precisely Korsgaard thinks is wrong with expressivism and, second, getting clear on the strength of her claim – does she think that expressivism is false, or true but unilluminating? What I am going to do here is first try to

[92] Thus modern-day expressivism is *conservative*, in the sense explicated by Schroeder's (2011) and Finlay's (forthcoming).

extract from Korsgaard a clear complaint against the expressivist before going on to see how the expressivist might try to respond.

Korsgaard argues that what lies behind the traditional distinction between non-cognitivism and cognitivism is a mistaken view about what our concepts are for – "that their cognitive job, so to speak, is to describe reality" (Korsgaard, 105). So a non-cognitivist agrees with the cognitivist that our cognitive concepts are for 'describing reality', but says that our moral thought and judgement is not like that. Korsgaard wants to rid us of this assumption – instead, our normative concepts possess some cognitive job, but are not in the business of describing reality. Thus both the cognitivist and non-cognitivist are both right and wrong. The cognitivist is right that our normative concepts have a cognitive job, but they are wrong about what that job is. In contrast, the non-cognitivist is right to believe that our normative concepts are not cognitive *in the sense used by both the cognitivist and the non-cognitivist*, however, they are mistaken in assuming that this means that they have no cognitive job at all.

So then, Korsgaard wants to give our normative concepts a cognitive job, although not one that is involved in describing reality. However, Hussain and Shah find it hard to make sense of this position. Cognitivism, they claim, must have something to do with knowledge:

> The point of calling a theory of normative concepts non-cognitive is that the theory rejects the assumption that the role of normative concepts is primarily that of helping with knowledge. (2006b, 34)

This deserves a little qualification – modern expressivists will want to claim that they can give an account of moral knowledge. They might, for example, argue that there is a suitable

deflationary notion of truth, and minimal notion of belief, according to which holding a particular type of non-cognitive attitude counts as knowing something. However, they *will* argue that this sort of moral knowledge is built out of materials that are practical in nature – we don't start our explanation using the concepts of moral beliefs and moral knowledge, but we end up being able to talk of those things. This is one of the reasons why someone like Simon Blackburn is uncomfortable describing himself as a non-cognitivist (even though he, by standard definitions, obviously is one). But what Hussain and Shah are intending to flag up is essentially correct – non-cognitivists at least start by seeing normative concepts as essentially tied to practical deliberation and action, rather than being in the business of describing reality; and cognitivists are comfortable with starting their enquiries into normative concepts assuming that they *are* primarily in the business of describing reality. What all this means is that it is hard to see how we could give an account of the cognitive role of a normative concept that is not in terms of its ability to describe reality – to think of a cognitive concept is to think of one that is in the business of representing the world as being a certain way. This is what underlies our (philosophers) typical practice of explicating the content of beliefs in terms of their truth conditions. If you have this sort of view, Hussain and Shah point out, what Korsgaard is trying to do will look confused.

Hussain and Shah trace this confusion to a misunderstanding about what non-cognitivism involves. Korsgaard says that "A conclusion of practical reasoning is not obviously a description of a fact about the world, but it hardly seems like some sort of emotional expletive either." (Korsgaard 2003, 105). Hussain and Shah suggest that the term 'emotional expletive' does not cover the quite complicated views about what moral judgements are, held by non-cognitivists. To them, it appears as if Korsgaard is confused between cognition and cogitation:

> A non-cognitivist theory is not a non-cogitative theory. All the major non-cognitivists – Stevenson, Hare, Blackburn, Gibbard- have complex theories about how practical reasoning proceeds in its complexity. (Hussain and Shah 2006b, 35)

On this Hussain and Shah seem to be right on the money – if we look at the details of the accounts put forward by expressivists like Blackburn and Gibbard we see that the non-cognitive attitudes that are used to characterise moral judgements are fairly sophisticated, and play a role in structures of attitudes that are extremely complicated. You can imagine them complaining to Korsgaard that although they do, ultimately, explain moral thought and talk in terms of non-cognitive states, the expressions of these states is nothing like an 'emotional expletive.' For example, as Neil Sinclair puts it "Modern expressivists eschew the idea that this state [ the one the expressivist uses to characterise a moral judgement] has a distinctive phenomenological hue." (2009, 137). What he means is that to the modern non-cognitivist the state that moral judgements express does not feel just like a paradigmatic desire or emotive state. Instead the non-cognitivist is more likely to talk about plans, intentions, or practical stances.

In addition, they demonstrate sophistication about what the targets of these non-cognitive attitudes are. Mark Schroeder, for example, has argued in *Being For* that the expressivist is well-advised to use attitudes of being for praising or blaming a particular action, rather than attitudes directed towards that action itself. For Gibbard (1990) the relevant attitudes are directed at feelings of guilt or resentment. These positions show how the non-cognitivist can give an account that is *emotionally ascended* (Blackburn 1998). What this means is that the attitudes in question are not much like bare emotional expletives (a characterisation more fitting of the earlier emotivist non-cognitivists). On these grounds Hussain and Shah dismiss Korsgaard's complaints against expressivism as confused.

However, things are not quite plain-sailing for the expressivist if they go down this quasi-realist route to try to capture our intuitions about the seriousness of moral discourse. There is a problem labelled *creeping minimalism* by Jamie Dreier (2004).[93] The basic problem is that if the expressivist succeeds in capturing the cognitivist appearances of moral discourse then they end up finding it hard to differentiate their position from cognitivism. For example, it looks as though people say things like not only 'Murder is wrong' but also 'It's *true* that murder is wrong.' As the quasi-realist brand of expressivism does not want to convict moral discourse of being in bad faith or in radical error they cannot simply agree with early non-cognitivists that moral utterances, because they lack literal significance, are not truth apt. Instead they are more likely to give a lightweight characterisation of the truth predicate.[94] What these views have in common is a rejection of a robust conception of truth, where truth is thought of in terms of correspondence to reality, or an epistemic characterisation in terms of what we have evidence for. Instead, these views give a central role to disquotational schema in their explication of truth: to understand the truth predicate (at least in part) is to understand that ''*p*' is true' iff *p*'[95] for each instance of p.[96] This means that the expressivist can quite easily recover a notion of moral truth – grasp of the truth predicate allows you to see that 'Murder is wrong' and ''Murder is wrong' is true' are

[93] But see also the earlier debate on a similar issue between Michael Smith, and Alex Miller and John Divers (Smith 1994b,c, Divers and Miller 1994, 1995) and for an exploration of 'disciplined syntacticism' see Lenman's (2003).

[94] There are a number of views of truth that are 'lightweight' in this sense – deflationism, minimalism, redundancy theories and more. Another distinction to bear in mind is between minimalism about truth aptitude and minimalism about truth. In order for the expressivist to use the strategy suggested in the text they will have to claim that moral utterances are truth apt (available to be evaluated for truth or falsity) which they can then combine with a lightweight characterisation of truth.

[95] Or, alternatively, the equivalence schema: S is true iff *p*. The differences between the two types of view do not matter here.

[96] The views then diverge on what other roles the truth predicate has – for example the minimalist argues that it also has a role as a generalising device. What they agree on is that truth is transparent.

equivalent – a competent speaker's familiarity with the relevant disquotational schema will allow them to disquote the second to derive the first, and travel in the opposite direction if necessary.

But now, if moral judgements can be true even on an expressivist account, there seems to be a problem. We could give a rough characterisation of a belief as something like 'holding that a particular content is true'. Well, on expressivism plus a lightweight characterisation of truth, it turns out that when people sincerely express moral judgements they can be thought of as holding some content to be true – to do that is just to have the relevant attitude and to understand the truth predicate. Hence we can define a minimal notion of belief on which moral judgements express beliefs. We can then use this to explicate related notions of moral facts, moral properties and moral knowledge. The problem is that the expressivist intended to characterise moral judgements as being un-belief-like, and now they seem to have lost that contrast. To put it extremely schematically, it looks as if there is some tension in the quasi-realist project – if the expressivist ends up explaining the cognitivist-looking features of moral discourse too well then they threaten their right to claim expressivism as a distinctive position. However, the quasi-realist project was well-motivated in the first place – if they don't try to save those features of moral discourse then they will have to argue that moral discourse is in radical error: whenever we say things like 'It is true that murder is wrong', 'Murder is *really* [stamp-foot] wrong', 'I know that murder is wrong'. 'You should believe that murder is wrong' we are just making mistakes. They also open themselves up to something like Korsgaard's initial charge: that on the expressivist account moral judgements are like mere emotional expletives.

The first problem is particularly severe because it threatens to force the expressivist to dispense with plausible principles of charitable interpretation. The expressivist's semantic

programme is (in part) driven by extra-semantic considerations – they, being thoroughly naturalistic, cannot countenance admitting robust moral properties to their ontology; and they think that moral judgements are inherently motivating. But the first is a motivation shared by an error-theorist who takes the step of saying that moral discourse *is* in radical error. There are two problems with the error theory though – first, because it is cognitivist it has trouble explaining the intrinsic motivational effects of moral judgements[97]; but secondly it is extremely *uncharitable* – on the error theorist's account we make gross errors a lot of the time. But why should we, the expressivist can claim, saddle ordinary speakers with the charge that they are massively confused if we can instead give a slightly different theoretical analysis of what they are saying where it turns out that they are *not* in massive error? This approach is vastly more charitable, and on that ground has much to commend it. However, if the problem of creeping minimalism threatens quasi-realist expressivism, then not only does Korsgaard's charge start to bite, but expressivism loses the advantage it had over something like error theory.

Expressivists do have some ways to try to respond to the problem of creeping minimalism though (Sinclair, 2009 gives a short survey). One approach is to try to retain a contrast between *minimal* and *robust* senses of beliefs – the expressivist can then claim that moral judgements are beliefs in the minimal sense whereas non-moral beliefs are beliefs in a robust sense. They could then co-opt something like Wright's pluralism about truth – where

[97] This is a problem for them from the expressivist's perspective. I've argued earlier that there is no pressing reason to accept internalism, so the error theorist gets a clean bill of health from me at least on this count. In fact, the problem with the error theorist in my view is that they are too much under the spell of internalism in that they think that moral discourse aims to ascribe properties that *are* intrinsically-motivating, before moving on to argue that such properties are, from the standpoint of naturalism, irredeemably queer. I argued that if we reject internalism properly, we don't have to view moral discourse as purporting to be internalist, but failing to find the right sorts of properties to latch on to – instead moral discourse is externalist, and we explain the reliable connection between moral judgement and motivation in terms of contingent psychological connections in normal human beings.

we distinguish between minimalist truth (where all that matters is satisfying certain platitudes and displaying syntactical discipline) and robust truth (where we are talking about something more like genuine correspondence) and tie each type of belief to the relevant notion of truth. However, this sort of strategy runs into problems with so-called *mixed-inferences* – arguments involving both moral and non-moral sentences look to be valid, but if we are using two notions of truth then these arguments actually turn on something like a form of equivocation and are thus invalid (see Christine Tappolet's 1997). This problem though can be resisted if we adopt a functionalist account of truth, where truth is indentified by its functional role in certain sorts of inferences, a role which can be realised in multiple ways (minimally or robustly).

Another approach is to start from an inferentialist framework (where the meaning of a term is tied to the inferential role it has) and to distinguish between *theoretical* and *practical* inferential roles. Moral concepts would be those that have a practical inferential role where "a practical inference is one whose premises provide practical support for a conclusion that can constitute practical knowledge about how to live"; non-moral concepts in contrast possessing a theoretical influential role: "a theoretical inference is one whose premises provide evidential support for a conclusion that can constitute theoretical knowledge about the world" (Sinclair 2009, 141). This is the view take up by Matthew Chrisman's (2008). Sinclair commends the view for the fact that it takes up a distinction between practical and theoretical reason already acknowledged by competent users of moral discourse. However, this position is not so much a way of fleshing out the distinction between expressivism and descriptivism as a way of changing the terms of the debate. We no longer would offer a characterisation of non-moral beliefs in terms of truth-conditions (except derivatively), so we would again lose the relevant contrast. Instead we have a method for distinguishing between moral and non-moral utterances within an inferentialist framework. This is not to

argue against the view though – instead I am merely suggesting it is not going to help the typical expressivist, even if Chrisman's own view is interesting and illuminating.

The final approach to the problem is to differentiate moral from non-moral beliefs in terms of the type of explanation we offer of them – whether they are explicated in terms of tracking worldly properties or not. This is the approach offered by Allan Gibbard (2003). As you can see, I have not been able to give anything like a comprehensive survey of these approaches to the problem of creeping minimalism. However, that is unrelated to the point I wish to make: Hussain and Shah dismiss Korsgaard as confused on the grounds that expressivist analyses of moral concepts are actually fairly sophisticated – however, when we look at the detail of these analyses it turns out to be very hard for the expressivist to secure that sophistication.

This is linked to the main thrust of what I'd like to say: Korsgaard draws our attention to the fact that moral judgements are not like 'emotional expletives'. One of the most fundamental problems with expressivism is that moral utterances are not like other expressions of non-cognitive states – they can be embedded in various ways (in conditionals, within the scope of modal operators, negated and so on) and this causes a problem for expressivist analyses. Korsgaard is drawing our attention to differences between moral utterances and mere expressions of non-cognitive states, and one of the differences we find is that moral utterances obviously embed in various ways that causes problems for the expressivist – a set of problems that travels under the banner of 'the Frege-Geach problem' which I now turn to.

## 4.3 The Frege-Geach Problem

Consider this argument:

(1) Bullying is wrong.

(2) If bullying is wrong, turning a blind eye to bullying is wrong.

So

(3) Turning a blind eye to bullying is wrong.

This argument looks perfectly valid – the truth of its premises guarantee the truth of its conclusion. It's also the sort of argument which is a pervasive feature of moral reasoning – one way to get someone to come to accept that something is wrong (or right) is to get them to accept that it's being wrong (or right) follows from some moral claim that they already accept. You can imagine someone who, although they accept that bullying is wrong, doesn't accept that turning a blind eye to bullying is wrong. If you get them to accept that if bullying is wrong, then turning a blind eye to bullying is also wrong then they face two options – either they can come to accept that turning a blind eye to bullying is wrong; or they can change their mind about the wrongness of bullying. This looks like a perfectly rational case of moral reasoning, underwritten by the validity of the above argument. What's not on for them to do is to say anything like 'Well, I accept that bullying is wrong, and I accept that if bullying is wrong then turning a blind eye to bullying is also wrong, but I think turning a blind eye to bullying is perfectly fine.' Such a person would be guilty of some sort of inconsistency.

What does this have to do with expressivism? Well, the expressivist has trouble explaining the validity of the above argument. This is because they give an account of the meaning of

'Bullying is wrong' as it features in **1** in terms of some non-cognitive sentiment – something like, very roughly, 'Boo to bullying!'.[98] However, it doesn't look like they can characterise the meaning of 'Bullying is wrong' as it features in the antecedent of the conditional in **2** in the same way. Why not? This is because **2** can be accepted by someone who does not disapprove of bullying. We can see this in a number of ways – we can imagine an alien anthropologist who is interested in investigating the structure of humanity's moral thought and talk but who themselves does not make any moral judgements. Or, we can make our example more mundane and simply consider someone who does not disapprove of bullying at all – a consistent bully perhaps – they might agree that if bullying is wrong then turning a blind eye to it is also wrong, and so assert **2** with perfect propriety, whilst lacking a sentiment of disapproval towards bullying.

This is the point most famously made by Geach (1958) when he says

> There arises here a difficulty for what may be called performatory theories of the predicates "good" and "true" – that to predicate "good" of an action is to commend it, and to predicate "true" of a statement is to conform or concede it. For such predications may occur within "if" clauses; the predicates "good" and "true" do not then lose their force, any more than other predicates used in "if" clauses do; but "if S is true" is not an act of conforming S, nor "if X is good" an act of commending X'.

[98] For the sake of exposition I will take the relevant sentiment to be disapproval of the action involved. Different expressivists characterise the attitude expressed by moral judgements in different ways, but, for now, these differences should not matter

(Geach 1958, 54 fn.)[99]

(The point remains unchanged if we substitute an attitude of disapproval for one of commendation in the above).

The reason why this is a problem is that now the argument from **1** and **2** to **3** turns on equivocation – 'bullying is wrong' as it features in **1** means something different from what it means in the antecedent of **2.** This means the valid-looking argument we started with comes out as invalid, and we end up attributing massive error to everyday users of moral discourse – every time they offer an argument like **1-3** they are committing a fallacy of equivocation.

This is the Frege-Geach problem as traditionally posed against simple expressivist semantics – moral judgements retain their inferential and logical properties even when not used with the same force: 'bullying is wrong' is unasserted in **2** and yet **2** still licenses the move from **1** to **3.** The point is far more general than merely explaining the validity of moral modus ponens arguments however. We encounter the same problem when moral terms are used in other complex constructions. Take the question 'Is x wrong?' This question does not typically express disapproval of x. However, an utterance of 'x is wrong' answers the question, which the expressivist claims *does* express disapproval of x. But how can this be the case, if 'wrong' in the question means something different to 'wrong' in the

[99] The point is developed further in Geach (1960) and (1965) and a similar point is made in Searle's (1962). The broad outline of my brief exploration of the history and point of the Frege-Geach problem is indebted to Schroeder's (2008a) and (2008b) and discussion with Jussi Suikkanen. See Suikkanen's (2007) and comments for a good discussion of the history of precursors to the Frege-Geach problem, and also James Urmson's (1968). Daniel Boisvert's (2004a), (2004b) and (2004c) give a more comprehensive survey of the issue than I have time for here.

answer? On this sort of expressivist account it looks like 'x is wrong' is an irrelevant response to 'is x wrong?', and this cannot be correct. The same point too holds for other complex constructions such as negation. In 'bullying is not wrong', 'wrong' is not used to disapprove of bullying. In fact, if anything the opposite attitude is being expressed. However, again this will mean that 'wrong' as it appears in 'bullying is not wrong' means something different to what it means in 'bullying is wrong'. This then makes it hard to see how 'bullying is wrong' contradicts 'bullying is not wrong', if the words involved mean different things.

The lesson to be drawn is that the Frege-Geach problem illustrates a more general problem with non-cognitivist semantics. In order to underwrite the validity of modus ponens arguments and explain how negations contradict unnegated statements and so on, the expressivist has to offer us an account of how the meaning of complex utterances is composed out of the meaning of simpler utterances. In order for **2** to license the move from **1** to **3** above 'wrong' must have the same meaning throughout, and the meaning of **2** as a whole must be built out of its simpler parts. We can understand Geach and Searle as arguing that the resources available to the expressivist are unsuitable for the job.

The centrality of compositionality to the Frege-Geach problem is acknowledged by modern attempts to give an expressivist solution to it. R. M. Hare (1970) argues that explaining compositionality is a genuine task even for the cognitivist who analyses meaning not in terms of the mental states expressed by utterances, but in terms of the truth conditions of those mental states. 'Bullying is wrong' does not have the same truth conditions as 'If bullying is wrong, then turning a blind eye to bullying is wrong'. So the cognitivist is forced

to tell a story about how the truth conditions of complex utterances is built up out of the truth conditions of their simpler components. The expressivist is simply faced with a similar task, except that they have at their disposal different elements – their job is to explain how the mental state expressed by **2** (remember, the expressivist gives the meaning of an utterance in terms of the mental state it expresses) is a function of its simpler constituents. Modern day expressivists thus view the Frege-Geach problem as presenting a challenge – to offer a compositional non-cognitivist semantics that underwrites the semantic or inferential properties of utterances containing moral terms. If they can do this then they have an answer to Geach and Searle's charge that on the expressivist account 'wrong' has different meanings in the different constructions it features in.

I will now briefly survey three approaches to this task before we move on to see what the hybrid-expressivist has to offer.

## 4.31 Higher Order Attitudes

Simon Blackburn (1973, 1984) suggests that we explain the inconsistency involved with accepting **1** and **2** and not accepting **3** by having **2** express a higher-order attitude. **2** on this account expresses the attitude of disapproving of a certain combination of other attitudes – namely disapproving of bullying and not disapproving of turning a blind eye to bullying. Thus, if you accept **1** and **2** but fail to accept **3** then you are guilty of inconsistency – in virtue of accepting **1** you disapprove of bullying; in virtue of accepting **2** you disapprove of

disapproving of bullying and not disapproving of turning a blind eye to bullying; in virtue of rejecting **3** you do not disapprove of turning a blind eye to bullying. This is inconsistent because given you accept **1** and **2** you should also disapprove of turning a blind eye to bullying, and yet your rejection of **3** shows you do not disapprove of turning a blind eye to bullying.

The chief worry with this approach is that it over-generalises validity[100]. This problem is most clearly stated by Mark van Roojen's (1996)[101]. Compare:

(1) Bullying is wrong.

(2) If bullying is wrong, turning a blind eye to bullying is wrong.

So

(3) Turning a blind eye to bullying is wrong.

To:

(4) Bullying is wrong.

(5) It is wrong to disapprove of bullying and not disapprove of turning a blind eye to bullying.

(6) Turning a blind eye to bullying is wrong.

According to Blackburn, **5** expresses the same attitude as **2.** We are looking for an account that underwrites the validity of the argument **1-3**, so if this account does that, and **5**

[100] Other difficulties are found in Bob Hale's (1986), (1993a, 1993b); Nicholas Zangwill's (1992)

[101] Here I use a modified version of Schroeder's (2008a, 709-10) presentation of the problem.

expresses the same attitude as **2** then the argument **4-6** should also be valid. However **4-6** is not valid. The lesson we can draw from the failure of this higher-order approach is, as Schroeder puts it:

> If expressivists are to be able to explain validity, they are going to need to appeal to a kind of incoherence among attitudes that is of a more specific type than the broad kind of incoherence to which Blackburn initially appealed. They are going to have to appeal to incoherence among attitudes that is of the very same type as the incoherence involved in both believing that *p* and also believing that ~*p*. (2008a, 710).

We will see this desideratum reflected in the hybrid expressivist's respect for the inconsistency constraint I present below.

## 4.32 Inconsistency in Content

The most obvious way to try to respect this desideratum is to find a way in which the mental states deployed by the expressivists in their semantics have the right kind of inconsistency, that the attitudes they invoke are, as Schroeder puts it 'inconsistency transmitting', where for an attitude to be inconsistency transmitting is for it to be inconsistent to bear that attitude towards inconsistent contents (2008b, 577). Belief seems like a paradigm case of an inconsistency transmitting attitude (it's inconsistent to believe *p* and also believe *not-p*). Supposing, or wondering might be non inconsistency transmitting (wondering whether *p* is not inconsistent with wondering whether *not-p*). The expressivist can now claim that something like disapproval, or intention *is* inconsistency transmitting,

and use that attitude to construct their semantics.[102] The picture would look something like this – we explain the contradiction between 'bullying is wrong' and 'bullying is not wrong' by claiming that these two utterances express attitudes with inconsistent contents: on the one hand disapproval of bullying and on the other disapproval of not bullying.

However, this approach fails as it is because it fails to provide enough attitudes to characterise all the meaningful utterances we can make using moral terms. This is seen in the discussions of the 'negation problem' sparked by the work of Nick Unwin (1999, 2001). The problem is that the attitude of disapproval of not bullying is not the one expressed by 'bullying is not wrong', instead it is expressed by 'not bullying is wrong'. And now we have no attitude remaining to characterise the meaning of 'bullying is not wrong'. To see the lack of a suitable attitude, consider the following set of utterances from Schroeder:

| | |
|---|---|
| (7) Bullying is wrong | (disapproval of bullying) |
| (8) Bullying is not wrong | (disapproval of x) |
| (9) Not bullying is wrong | (disapproval of not bullying) |
| (10)Not bullying is not wrong | (disapproval of y) |

**7** and **8** are inconsistent, and this inconsistency should be explained, on this suggested approach, by inconsistency in the content of those attitudes. Similarly, **9** and **10** are inconsistent. So, **8** has to express disapproval of something inconsistent with bullying. But **10** has to express disapproval of something inconsistent with not bullying. Thus, on this approach **8** and **10** will end up being assigned inconsistent contents, and thus **8** and **10** will contradict one another. However, **8** and **10** are not inconsistent – we can see this most

[102] This is the approach taken by Gibbard's 2003, although see footnote 103 below on the issue of whether this mischaracterises Gibbard's position

clearly see this if we take a less morally loaded example: clasping your hands three times every time you wake up is neither morally required nor impermissible, so it can be the case that clasping your hands three times in the morning is not wrong and not doing it is also not wrong. Explaining the compositional properties of moral terms in this way threatens to eliminate the category of the merely permissible.

### 4.33 Hierarchy of Attitudes

This happens when we try to characterise the inconsistency between an utterance and its negation in terms of the contents of the attitude expressed – using the same attitude (disapproval) but which takes a different content. Another approach is to posit a distinct attitude for each utterance. Unwin (1999, 2001) illustrates the pressure in this direction by considering the case of reports of the relevant attitudes. Take a construction like 'James thinks that bullying is wrong'. In such a sentence there are three places to place the negation operator: 'James doesn't think that bullying is wrong'; 'James thinks that bullying is not wrong'; 'James thinks that not bullying is wrong'. The expressivist has the resources, when restricted to using the same attitude (say, disapproval), for explaining the meaning of at most two of these utterances (they have a lack of disapproval of bullying, and disapproval of not bullying at their disposal). In order to provide the meaning for the missing negation they will have to posit a distinct attitude.

Unwin's and Schroeder's diagnosis of this problem is that the expressivist's account lacks the right kind of structure to find a place for the negation operator. But, if this is the case, then the problem can be extended – similar problems arise for modal operators, conjunction and disjunction, tenses, and other complicated constructions. This means that given the

complexity of moral sentences we can construct which are still meaningful, the expressivist is forced to posit an increasingly large number of distinct attitudes for each complex utterance. It could be that the expressivist can provide the right attitudes with the right inferential and semantic properties to complete this project. However, the problem is now that the expressivist semantics looks fairly *ad hoc* – they end up characterising the relevant attitudes in terms or the inferential relations they have to sustain, which looks to be equivalent to saying "that complex sentences express that state of mind, whatever it is, that would ensure that they have the right semantic properties" (Schroeder, 2008a p. 714, see also his 2008c and 2008d).

### 4.34 Adding Structure to the Attitude

So it looks like an expressivist cannot rely on inconsistency in the content of attitudes to explain the semantic properties of moral terms, nor posit distinct attitudes to play the right sort of role. Another way forward might be to follow Schroeder in introducing an attitude with the right amount of structure to play the role demanded from it in explaining compositionality. The problem for the expressivist is that they assume that for each moral predicate (right, wrong, etc) there is a distinct attitude that is expressed by sentences predicating that term of some object – for 'wrong' there is the attitude of disapproval: thinking that bullying is wrong is to have an attitude of disapproval towards bullying. What would happen if we introduced such a constraint into a cognitivist semantics, and posited a different attitude of belief for each descriptive predicate? Schroeder, using the example of 'believes-green' as the attitude towards grass expressed by 'Jon thinks that grass is green', argues that the negation problem would emerge for ordinary descriptive discourse:

**G** Jon thinks that grass is green.

**N1** Jon does not think that grass is green.

**N2** Jon thinks that grass is not green.

**G*** Jon believes-green grass.

**N1*** Jon does not believe-green grass.

**N2*** ???

Again, we have run out of places to put the negation operator. The reason why this problem does not emerge for the cognitivist is that no cognitivist takes 'believes-green' to be a distinct propositional attitude. Instead, they can characterise this sort of belief as holding a more general attitude which combines with a particular property.

The lesson for the expressivist is to use an attitude with a similar level of structure, so that they end up analysing moral utterances in terms of "a more general non-cognitive attitude and a descriptive property or relation" (2008b, 589). The details of which attitude we use aren't of much concern, but Schroeder suggests 'being for'. The idea is that we then analyse an attitude like disapproval as something like 'being for blaming for'[103]. So, 'bullying is wrong' expresses the non-cognitive attitude of being for blaming people for bullying. What this account does is introduce the right level of structure to find an additional place to place

[103] Schroeder characterises Allan Gibbard as offering an approach like that surveyed under 'hierarchy of attitudes'. I suspect the similarity between Schroeder's analysis of disapproval in terms of being for blaming (a view inspired by Gibbard's 1990) is what prompts Ralph Wedgewood (2010) to argue that Gibbard's position is more like Schroeder's own suggestion, and thus that Schroeder's suggested semantic program for expressivism is not as novel as Schroeder claims. See Schroeder (2010) for a response. Whichever side is right in this debate, what I say about this sort of attempt is unaffected.

a negation operator. We could then offer an analysis of each way of negating 'James thinks that bullying is wrong.':

(11) James thinks that bullying is wrong – James is for blaming for bullying

(12) James does not think that bullying is wrong – James is not for blaming for bullying

(13) James thinks that bullying is not wrong – James is for not for blaming for bullying

(14) James thinks that not bullying is wrong – James is for blaming for not bullying

Schroeder then goes on to show how to build up a notion of inconsistency and an account of the logical connectives out of these building blocks (2008b, 2008d).

However, although this approach looks to have gone the farthest in explaining the compositionality of moral language, it does face a problem (freely admitted by Schroeder). The notion of inconsistency and the account of the logical connectives it yields are distinct from the characterisations we give of those in the case of purely descriptive language. This yields a problem when we consider that we do not speak two distinct languages – one descriptivist and one evaluative. We make inferences across the two types of language, for example:

(15) Fish can feel pain

(16) If fish can feel pain it is wrong to go angling

(17) It is wrong to go angling

Now, someone might accuse anyone uttering **16** of committing the naturalistic fallacy – of inferring a moral judgement from a statement about how things are naturalistically speaking

– but that does not demonstrate that the above argument is invalid. Someone who wields the naturalistic fallacy in this blunt way is not impugning the validity of the argument, they just claim that statements like **16** are never true. But this does not mean that **17** does not follow from **15** or **16**, nor, even more pressingly, that we can do without an account of the meaning of mixed utterances like **16** – even if **16** is always false, it is still the sort of thing that people can think, and surely there is *something* they think and assert using **16**. We have to be able to give an account of the contents of these mixed utterances. And of ones of an even simpler form like:

(18) Bullying is wrong and snow is white

Which is inconsistent with both 'bullying is not wrong' and 'snow is not white'. But now we face a dilemma: if we account for the meaning of something like **18** in terms of belief we can readily explain how it is inconsistent with 'snow is not white' – but we will have trouble explaining its inconsistency with 'bullying is not wrong', where, remember, we use a completely different notion of inconsistency. We'd have another problem explaining **18** in terms of the relevant non-cognitive attitude.

This leads Schroeder to conclude that:

> [T]he only way to apply the advantages of the account that I've sketched here, on which we can reduce the explanation of the inconsistency of arbitrary sentences to the inconsistency of the contents of the attitudes that they express, is to allow that all sentences express the same general kind of attitude. (2008b, p. 597).

What this means is that Schroeder, if right about the implications of his view, has demonstrated the position outlined by Max Kölbel's (1997). Kölbel there argues that,

contrary to what has been thought, expressivism does not provide a way of giving a non-cognitivist treatment of a part of language where that part of language interacts with the part we give a descriptivist treatment to. Expressivism is thus only viable for discourses that do not interact with descriptive discourse in any significant way. The problem for expressivism in ethics, then, is that ethical discourse does significantly interact with non-ethical discourse.

If this is right, then the Frege-Geach problem in part illustrates the difficulties inherent in explaining the interface between the part of language we give a non-cognitivist treatment of, and the part we treat as cognitivist. Schroeder's suggestion on behalf of the expressivist, that they end up characterising *all* sentences as expressions of non-cognitive attitudes, is thus one way out of this bind – we can avoid the problems that come from explaining how moral and non-moral utterances interact by *globalising* our expressivism. This is, in fact, the line taken by a few expressivists.[104] However, it is worth noting that taking up this line is not without its costs. Expressivism is typically inspired by the thought that there is something especially problematic with ethical discourse. Moral judgements are inherently motivating, while descriptive judgements are not, for example. But now if we globalise expressivism we lose at least this motivation – for there will be no difference between ethical and non-ethical judgements at this level. In fact, now the different motivational import that ethical and non-ethical judgements have (which the expressivist hoped to explain by invoking attitudes with an intrinsically motivational element) is actually an embarrassment to the globalised expressivist – at the very least a piece of data that far from supporting their position, needs to be explained away.

[104] See Stephen Barker's (Ms), for example, and Huw Price looks set to explore similar themes in his (forthcoming). We can also see some of the work of Matthew Christman (2008 and 2010) as attempting to occupy this logical space

Another advantage usually offered on behalf of expressivism is its metaphysical and epistemological solvency – it does away with the need to invoke strange moral properties or a special epistemology for their study. But if we globalise, we may be worried that our view does too well in deflating our metaphysical commitments – we end up being anti-realist about not just moral properties, but also entities in the physical universe.

Of course what I say here is very broad, and it could be that some expressivists are happy to globalise their expressivism. I can end this discussion with only two further points. First, it is a result of the work of people like Schroeder, Unwin and Finlay that now means that it is not enough to simply gesture at an alternative semantic picture anymore – we need to see the details worked out, and it is up to the globalised expressivist to give us a detailed account of their programme for evaluation. Second, as I have tried to briefly show, globalised expressivism is unlikely to attract the sort of expressivist who was motivated by seeing problematic and distinctive features in moral discourse.

Where does this leave us? I've canvassed three historically important approaches to the Frege-Geach problem, and have tried to show that they all face problems. This provides us with motivation to see if hybrid-expressivism fares any better at tackling the problem. To talk schematically, hybrid-expressivism attempts to combine elements of cognitivist and non-cognitivist semantics to solve the problems with both of these accounts individually. The impetus to look to hybrid expressivism for a solution to the Frege-Geach problem is two-fold: it may give us a way to make sense of Korsgaard's contention that both realism and expressivism are true, in a sense; and we have seen above that reflection on the Frege-Geach problem leads one to suspect it arises as a result of the interaction between ethical and non-ethical discourse, and perhaps hybrid-expressivism, which uses elements of the two

distinct type of semantics that have been offered for each of these holds some promise of a solution.

## 4.4 Hybrid Expressivism

So-called hybrid meta-ethical theories attempt to combine elements of non-cognitivist and cognitivist semantics to solve the problems that attend each account individually. What I aim to do here is to look in detail at one version of a hybrid theory – Michael Ridge's ecumenical expressivism – to see what lessons can be learnt from it. I shall argue that Ridge's ecumenical expressivism fails to offer us a viable, distinctive, solution to the Frege-Geach problem. In the course of this I shall also investigate another hybrid position Ridge delineates – ecumenical cognitivism, and argue that it too fails to match up to the hybrid theorists ambitions. I will then briefly look at two other types of hybrid theory: Copp's realist-expressivism, and Bar-On and Christman's neo-expressivism. Overall I will argue that the example of hybrid metaethical theories illuminates a constraint on metaethical theorising that has, up until now, tended to be respected – that accounts of moral semantics and moral psychology should be integrated in the right kind of way. Hybrid metaethical theories either do not meet this constraint, or are not properly classified as fully hybrid. I shall also then reconnect the material about hybrid expressivism back to Korsgaard's concerns. I will argue that hybrid expressivism does not give us a position that the neo-Kantian would accept.

One thing to note before I get into a characterisation of Ridge's views, is that hybrid metaethical theories are usually construed as attractive because they are minimally revisionary (this is most explicit in the case of Ridge). What I mean is that they attempt to

provide a solution to the problems bedevilling cognitivism and non-cognitivism that leaves most of the terms of metaethical debate standing. This is why Ridge, for example, characterises his solution to the Frege-Geach problem as 'cheap'. In particular, most hybrid theorists are attempting to keep in place the Humean theory of motivation – that beliefs and desires are distinct existences, and that there are no necessary connections between them. This desideratum is particularly pressing for the expressivist, as we saw above that they can use their commitment to the Humean theory of motivation to batter realism. After all, if mental states could have representational *and* motivational contents, judgement internalism would be less of a problem for the realist. This is also important for my criticisms, as I will mainly be arguing that Ridge's positions fail because they don't secure judgement internalism and the Humean theory together.

## 4.41 Ecumenical Cognitivism vs. Ecumenical Expressivism [105]

Ecumenical views claim that moral judgements express *both* beliefs *and* non-cognitive, desire-like attitudes. At first glance, it might seem hard to see how we could distinguish different types of ecumenicism using the cognitivist/expressivist dichotomy. As stated, the view sounds like a synthesis of *both* views, so why should we expect there to be both a cognitivist and an expressivist version of ecumenicism? Ridge begins by offering us the following way of distinguishing between cognitivism and expressivism in general:

> **Cognitivism:** For any moral sentence M, M is conventionally used to express a belief such that M is true if and only if the belief is true.

[105] The material in sections 4.41 to 4.45 is based on joint work between myself and Alex Miller (under review). This is an entirely collaborative work, with each author making an equal contribution.

> **Expressivism:** For any moral sentence M, M is not conventionally used to express a belief such that M is true if and only if the belief is true (Ridge 2006, 307).

With this distinction in place, Ridge offers the following on how to distinguish between cognitivist and expressivist versions of ecumenicism:

> Ecumenical cognitivism allows that moral utterances express both beliefs and desires and insists that the utterances are true if and only if one of the beliefs expressed is true. Ecumenical expressivism also allows that moral utterances express both beliefs and desires but denies that a moral utterance is guaranteed to be true just in case the belief(s) it expresses is (are) true (2006, 307-8).

It is important to be clear about the role of "guarantee" here. Later in the 2006 paper Ridge says of ecumenical expressivism that "as long as the belief expressed by a moral utterance is not *semantically guaranteed* to provide the truth-condition for the utterance, the fact that the belief expressed *contingently* provides the truth-conditions for the token utterance is consistent with expressivism as characterized here" (2006, 312, emphases added); and he notes that a version of ecumenical expressivism that concedes that moral utterances are truth-apt would nevertheless "deny that their truth-conditions *necessarily* are provided by the beliefs they express" (2006, 316, emphasis added). Clearly, the mention of a *semantic* guarantee indicates that the modality involved is conceptual: the truth of the belief that Jones is an unmarried male semantically guarantees the truth of the belief that he is a bachelor, since, as a matter of conceptual necessity, all unmarried males are bachelors. It is the fact that the beliefs expressed by moral utterances are not semantically guaranteed to provide their truth-conditions that apparently allows the ecumenical expressivist to bypass the Moorean open-question arguments that challenge cognitivism (2006, 309).

With this much on board, we now have:

> **Ecumenical Cognitivism**: a moral judgement M expresses both a belief and a desire, and, as a matter of conceptual necessity, M is true iff the belief expressed is true.
>
> **Ecumenical Expressivism**: a moral judgement M expresses both a belief and a desire, but it is not conceptually necessary that M is true iff the belief expressed is true.

The difference between the two views, then, is that for the ecumenical cognitivist the truth-conditions of the belief expressed by a moral judgement give the truth-conditions of that judgement in the sense that if those conditions are met, then as a matter of conceptual necessity the moral judgement is true. The ecumenical expressivist, on the other hand, denies this: according to the ecumenical expressivist the truth-conditions of the belief expressed by a moral judgement do not give you the truth-conditions of that judgement in this way - even if the truth conditions for the relevant belief are met, it does not follow as a matter of conceptual necessity that the moral judgement is true. It is easy to see why the ecumenical expressivist should deny this conceptual link between the truth-conditions of the belief expressed by a moral judgement and the truth of that judgement – otherwise they would not be expressivists at all, ecumenical or otherwise, nor would they be able to sidestep the open-question argument.

As thus far stated ecumenical expressivism may look like a purely negative thesis: ecumenical cognitivism, without the conceptual connection between the truth of the belief expressed by a moral judgement and the truth of that judgement. However, in order to remain truly expressivist the ecumenical expressivist will have to make the additional,

positive, claim that it is the presence of the relevant desire that makes a moral utterance a moral utterance. As Ridge puts it: "for the ecumenical cognitivist, belief has a kind of priority, in that which beliefs are candidates for counting as moral is fixed by their content" whereas "ecumenical expressivism instead gives logical priority to desire" (2006, 308-9). Another way to make the same point is that in their attempt to explain the nature of moral judgement ecumenical cognitivism and ecumenical expressivism adopt the same directions of explanation as their non-ecumenical cousins. The cognitivist assumes that our moral judgements express beliefs that possess truth conditions, and attempts to explain moral judgement by first identifying specifically moral truth-conditions before moving from them to the contents of moral beliefs, which in turn give us the sought for account of moral judgement. The explanation goes from truth-conditions to judgements. In contrast the non-cognitivist, as touched on above, starts by taking our moral judgements to be expressions of non-cognitive attitudes, which in sophisticated versions of the view such as Simon Blackburn's quasi-realism *feed into* an account of moral truth and moral truth-conditions. For the expressivist we *end up* with an account of moral truth-conditions, while for the cognitivist they form the starting point of the account of moral judgement. The ecumenical expressivist and ecumenical cognitivist respect this distinction: for the ecumenical cognitivist, the content of moral judgement is given by the truth-conditions of moral belief, whereas the ecumenical expressivist sees the desire, and not the truth-conditions of the associated belief, as the determinant of the distinctive content of moral judgement.[106]

[106] See e.g. Blackburn 1993a (chapter 3) and 1993b for the distinction between the two directions of explanation.

So, ecumenical cognitivism deserves to be called a form of cognitivism because it shares with non-ecumenical cognitivism the theses that a moral judgement expresses a belief, that if the truth-conditions of the relevant belief are met then the judgement is semantically guaranteed to be true, and that it is the fact that the relevant belief is expressed that makes the judgement a distinctively moral judgement . What makes it ecumenical is the additional claim that moral judgements also express desires (Ridge 2006, 307).

Likewise, ecumenical expressivism deserves to be called a form of expressivism because it shares with non-ecumenical expressivism the theses that that moral judgements express desires, that the truth of the judgement is not semantically guaranteed by the truth of any belief expressed by the judgement, and that it is the fact that the relevant desire is expressed that makes the judgement a moral judgement (Ridge 2006, 307, 316). What makes it ecumenical is the additional claim that moral judgements also express beliefs.

The distinction between the two types of ecumenicism should then be clear: ecumenical cognitivists claim that the truth-conditions of the belief expressed by a moral judgement give you the truth-conditions of that judgement in the sense that if those conditions are met then the moral judgement is conceptually guaranteed to be true. The ecumenical expressivist denies the existence of this conceptual link between the truth of the belief expressed by a moral judgement and the truth of that judgement, and suggests that it is the associated desire-like attitude that makes the judgement a moral judgement.

With this distinction in hand I can now turn to how ecumenicists can attempt to solve problems that are the traditional bugbears of their non-ecumenical cousins.

## 4.42 Ecumenical Cognitivism and Judgement Internalism

As we've seen before judgement internalists claim that there is some sort of conceptual connection between making a moral judgement, and being motivated to act in accordance with that judgement, typically endorsing something like Smith's practicality requirement:

> **PRAC**: If an agent judges it is right to φ then she will be at least somewhat motivated to φ, unless she is practically irrational.

We have also seen how non-cognitivists can cause trouble for cognitivists by combining judgement internalism with the Humean theory of motivation – if beliefs and desires are distinct existences, sustaining no necessary connections to each other, and moral judgement is a matter of belief, then there shouldn't be the conceptual connection between moral judgement and motivation posited by the internalist. Humeanism, cognitivism and internalism appear to be inconsistent. Suppose that Jane, a practically rational agent, makes a moral judgement. Assuming cognitivism for reductio, Jane's judgement expresses a moral belief. Given motivational internalism it follows as a matter of conceptual necessity that Jane is motivated to act in accord with her moral belief. Given Humeanism, Jane, since she is motivated to act, must have a desire that meshes with the belief. The connection between that desire and her moral belief is conceptually necessary, since there is a conceptually necessary relation between the moral judgement that expresses the belief and the motivational state that contains the desire. This contradicts the Humean assertion that beliefs and desires are "distinct existences".

Ridge claims that ecumenical cognitivism has a straightforward solution to this problem that allows it to consistently retain both Humeanism and motivational internalism.

According to the ecumenical cognitivist moral judgements express both beliefs and desires. It should not then be surprising that moral judgements can motivate: they express a desire-like attitude, so there is no need to cite an *additional* desire-like attitude to explain why we are motivated to act in accord with them. So the ecumenical cognitivist can simply exploit the traditional expressivist explanation of the practicality of moral judgement – that moral judgements move us to act because they are expressions of desire-like motivational states – and in this way reconcile cognitivism with both Humeanism and motivational internalism (Ridge 2006, 309).

With this in hand it is possible to explain the difference between ecumenical cognitivism and a view that is superficially similar. "Besire" theories claim that moral judgements express "a unitary mental state which has properties of both belief and desire" (Altham 1984, 284): moral judgements are partly representational (like beliefs) whilst also partly motivational. Such a position seems bizarre since "intuitively any representation can exist without the motivation allegedly essential to that representation" (Ridge 2006, 304). Ecumenical cognitivism, on the other hand, has no need for such strange mental states. Ridge claims that unlike the besire theorist, the ecumenical cognitivist can, consistently with motivational internalism, hold on to a Humean moral psychology on which there are no unitary states with both belief-like and desire-like characteristics.

However, it is not all plain sailing for the ecumenical cognitivist. Ecumenical cognitivism remains committed to the claim that moral judgements are semantically guaranteed to be true if the beliefs they express are true. We can ask what the content is of these beliefs. One possibility is that they are about the instantiation of natural properties. However if this is the

ecumenical cognitivist's position they face Moore's open question argument: for any proposed naturalistic analysis N of moral predicate M, somebody who asked whether something which is N really is M would not betray any conceptual confusion. The question "Is something which is N really M?" seems open. The Moorean then claims that the best explanation of the fact that the question seems open is that the proposed naturalistic analysis is false.[107]

Moved by this argument, the cognitivist might argue that moral beliefs concern the instantiation of non-natural, *sui generis*, irreducible moral properties (as Moore does himself). If they take this line, then they face the task of explaining the supervenience of the moral on the natural. It does not seem possible for there to be bare differences in the instantiation of moral properties: if two things differ in some respect with regard to their moral properties, there must be some naturalistic difference between them too. If, as the non-naturalist cognitivist realist holds, moral properties are distinct from natural properties, the nature of this connection looks mysterious.[108]

Thus, Ridge contends, the possibility of ecumenical cognitivism allows us to "transform" metaethical debate. Through their ecumenicism these cognitivists gain an easy solution to

[107] Even if this argument goes through, the question then becomes whether a failure of any proposed naturalistic *analysis* of a moral terms has much force against naturalist cognitivism. It is a matter of some controversy whether there is a descendant of Moore's argument that militates against synthetic versions of naturalism. See e.g. Horgan and Timmons 1992, and section 3.3 above.

[108] In fact, there are two possible worries for the cognitivist concerning supervenience. The first is the difficulty of accounting for the a priori supervenience of the moral on the natural (see Smith 1994: 21-24), which specifically threatens the Moorean non-naturalist. However, Blackburn provides us with another worry concerning supervenience - his so called "ban on mixed-worlds" argument (1984: 182-6) - which threatens the naturalist and non-naturalist cognitivist alike.

the problem of combining cognitivism, motivational internalism and the Humean theory of motivation. They trade this problem for renewed interest in the ability of cognitivism to account for the supervenience of the moral on the natural whilst defusing the open-question argument.

### 4.43 Ecumenical Expressivism and the Frege-Geach Problem

The expressivist, on the other hand, does not have to face these particular problems. Expressivists deny that the truth of a belief expressed by a moral judgement conceptually guarantees the truth of that judgement. Therefore, asking "Is something which is N really M?" does not betray any conceptual confusion, and the open question argument simply fails to get a grip.[109] Ridge also contends that the expressivist has a straight-forward explanation of the supervenience of the moral on the natural:

> [T]he expressivist needs only to explain the sensibility of adopting a supervenience constraint. Since the point of moral discourse is to recommend options on the basis of their natural properties, it is easy to see why such a constraint is sensible (Ridge 2006: 306)[110]

The major problem that expressivists face, as we saw above is the Frege-Geach Problem. Consider again:

[109] Things are not quite this simple: it is possible to argue that even expressivists still have to answer a variant of the open question argument (see Miller 2003: 47-51, 88-94).

[110] This is the line taken by Blackburn (see e.g. Blackburn 1984: 182-6).

(1) Bullying is wrong.

(2) If bullying is wrong then turning a blind eye to bullying is wrong.

So,

(3) Turning a blind eye to bullying is wrong.

We saw earlier that a straight expressivist explanation of the validity of the move from **1** and **2** to **3** falls prey to a number of objections. As we saw before, for the reasons given by Van Roojen, any proposed solution to the Frege-Geach problem needs to meet the following constraint:

> **Inconsistency Constraint:** the account must explain why someone who accepts the premises of a valid argument involving moral terms, but who denies the conclusion, is making a *logical* mistake. This inconsistency must be logical, rather than the pragmatic inconsistency exemplified by "Moore's paradox" style sentences, e.g. "I believe that P, but not-P" (see Ridge 2006, 313).

To see how a little ecumenicism would help the expressivist out with respect to the Frege-Geach problem, we need two elements: the ecumenical expressivist's analysis of moral judgements, plus their revisionary account of validity. Take a moral judgement of the form "X is wrong". The ecumenical expressivist claims that this expresses both a desire-like attitude and a belief. But which belief and which desire-like attitude? On Ridge's account this judgement expresses (a) an attitude of disapproval towards actions insofar as they have

a certain property N and (b) a belief that X has that property.[111] Two points need to be stressed. First, the moral judgement does not inherit the truth conditions of the belief identified in (b); neither is the meaning of the utterance given by the truth conditions of the belief referred to in (b). Ridge needs to make these claims in order to differentiate ecumenical expressivism as a distinctly *expressivist* position. Second, the agent concerned is not required to know much at all about the property on the basis of which they disapprove of things. The belief identified in (b) refers to that property via "anaphoric pronominal back-reference" (Ridge 2006, 313-7).

Given that ecumenical expressivists may want to claim that moral judgements are not truth apt[112], they cannot make use of the standard notion of validity that says that an argument is valid when the truth of the premises guarantees the truth of the conclusion. Instead Ridge offers the following revisionary account that he calls a "close cousin" of the traditional definition:

> **Validity**: An argument is valid just in case any possible believer who accepts all of the premises but at one and the same time denies the conclusion would thereby be guaranteed to have inconsistent beliefs (Ridge 2006, 326).

[111] In fact, on Ridge's preferred account, the attitude of (dis)approval is "a state of [dis]approval to actions in so far as they would garner the [dis]approval of a certain sort of advisor" (Ridge 2007a: 98). For the present purposes, nothing turns on the additional complexity introduced by this, so I simply ignore it in what follows.

[112] Expressivists who adopt a deflationary theory of truth-aptness may want to claim that moral utterances are truth-apt and have truth conditions that are contingently provided by the concomitant minimal belief. They could then follow Daniel Stoljar 1993 and Huw Price 1994 in arguing for a truth-conditional treatment of moral *modus ponens* using expressivist resources. Ridge argues that it would be a bad move, dialectically, to tie the success of expressivist solutions to the Frege-Geach problem so closely to the success of deflationary theories of truth apt-ness. In addition, he contends that these deflationary solutions still run into problems respecting the various constraints on attempted solutions to the problem (Ridge 2006, 312-3). So, officially, ecumenical expressivism is supposed to be neutral with respect to deflationism about truth-aptitude.

With these elements in place, we can see how Ridge's ecumenical expressivist can attempt to solve the Frege-Geach problem. Take the modus ponens argument involving moral terms above. According to the ecumenical expressivist, in **1** "bullying is wrong" expresses (a) a non-cognitive attitude of disapproval towards things that have a certain property N; and (b) the belief that bullying has that property. **2** expresses the belief that if bullying has the property N then turning a blind eye to bullying also has N.

Now, denying **3** would involve the belief that turning a blind eye to bullying doesn't have the property N. But now it would be inconsistent to accept **1** and **2** but reject **3**: in virtue of rejecting **3** you'd believe that turning a blind eye to bullying lacks N and in virtue of accepting **1** and **2** you would believe that turning a blind eye to bullying has N . This is a straightforward inconsistency of belief, so the argument turns out to be valid in such a way that the **Inconsistency Constraint** is respected.

So, if Ridge is right the ecumenical expressivist can use cognitivist resources to obtain a relatively easy solution to the Frege-Geach problem. However, because the ecumenical expressivist denies that there is a conceptual link between the truth of the belief expressed by a moral judgement and the truth of that judgement, they also avoid the open question argument, and are not prevented from helping themselves to the standard expressivist explanation of supervenience. Again, ecumenicism helps to transform contemporary metaethics. Since ecumenical cognitivism can help itself to expressivist accounts of the practicality of moral judgements, the focus of the metaethical debate between cognitivism and expressivism moves away from the Frege-Geach problem and motivational internalism

and back towards the open-question argument and the supervenience of the moral on the natural.

## 4.44 Ecumenical Expressivism Does Not Solve the Frege-Geach Problem

The nub of the Frege-Geach problem for non-ecumenical expressivism is that it is unable, for example, to account for the appearances of moral sentences in unasserted contexts in a way that preserves the validity of simple inference patterns such as moral modus ponens. Ridge attempts to avoid the problem via the claim that the relevant moral sentences express beliefs, in such a way that an agent accepting the premises but not the conclusion of moral modus ponens would thereby be guilty of a straightforward inconsistency in belief – thereby securing the validity of the argument via the revisionary account of validity outlined above.

However, Ridge's ecumenical expressivist does not provide an adequate reply to the Frege-Geach problem. First, we should note that it is not enough merely to show that the argument from **1** and **2** to **3** comes out as valid on an adjusted conception of validity. The expressivist must also show that there is no *equivocation* involved between the appearance of "Bullying is wrong" in **1** and in the antecedent of **2**. If she cannot do this, then the alleged fact that the argument comes out as valid on the adjusted conception of validity constitutes a *reductio* of the conjunction of ecumenical expressivism with the adjusted conception of validity (since that conjunction appears to imply that an argument can simultaneously commit a fallacy of equivocation and count as valid): in which case, no plausible solution to the Frege-Geach problem will have been delivered.

So, what does the ecumenical expressivist have to do in order to speak to the concern about equivocation? He needs to show how the meaning of "Bullying is wrong" as it appears in the antecedent of **2** can be given in terms of the sentiment of disapproval it expresses in **1**. Recall that it was the inability of the non-ecumenical expressivist to do this that constituted the Frege-Geach worry in the first place: the claim that "Bullying is wrong" expresses a belief in the antecedent of **2** by itself makes no progress on the original worry since that worry was generated not so much by the absence of a role for belief in the account of the meaning of "Bullying is wrong" as it appears in the antecedent in **2** but by the absence of a role for the noncognitive sentiment of disapproval.

The ecumenical expressivist cannot speak to the worry about equivocation by invoking the belief he claims to be expressed by moral judgements, since he denies that the meaning of the judgements is given by the truth-conditions of the belief. If they were, from the fact that the truth-conditions of the belief are satisfied it would follow as a matter of conceptual necessity that the moral judgement in question is true: but denying this is a crucial part of what makes the ecumenical expressivist an *expressivist*. In other words, Ridge could only adopt this strategy – giving a role to belief in determining the distinctive content of "Bullying is wrong" – at the expense of losing the right to the distinction between cognitivist and expressivist versions of the ecumenical view.[113]

[113] What if the ecumenical expressivist protests that the belief that bullying is N can be assigned a role in determining the meaning of "Bullying is wrong" in a manner consistent with ecumenical expressivism: the belief determines the referent of "Bullying" and thereby explains why "Bullying is wrong" has a different meaning from e.g. "Torture is wrong"? This is consistent with ecumenical expressivism, since although the belief contributes to the meaning of "Bullying is wrong" it does not determine its truth-conditions. (I am grateful to Neil Sinclair for raising this point). In response to this, we can ask whether this is all the belief contributes to the meaning of "Bullying is wrong". If it is, then this is unlikely to be of much help to Ridge – for although now we will have managed to guarantee that "Bullying is wrong" differs in meaning from "Torture is wrong" (since the belief fixes the relevant action type as the referent

So how might the ecumenical expressivist find a role for the sentiment of disapproval in the story about the meaning of the antecedent in **2**? According to the ecumenical expressivist, in **1** "Bullying is wrong" expresses a non-cognitive attitude of disapproval towards things insofar as they have a certain property N. In fact, Ridge's ecumenical expressivist takes "If bullying is wrong then turning a blind eye to bullying is wrong" to be expressing the *same* attitude of disapproval in terms of which we give the meaning of **1**. Likewise for the appearance of "Bullying is wrong" in the conclusion **3**: this expresses a non-cognitive attitude of disapproval towards things insofar as they have the property N. Now this may appear to solve the worry about equivocation: there is no equivocation because the desire-like attitude in terms of which the meaning of moral utterances is to be given is the same across all three steps of the argument.

However, it would do so only at a severe price. The original worry about the expressivist account of the appearance of "Bullying is wrong" in **2** was in part that someone could with

---

of the utterance) we haven't yet ruled out equivocation, for the belief does not show how "Bullying is wrong" differs from "Bullying is right" or "Bullying is morally neutral". If, however, the belief makes more of a contribution to the meaning of "Bullying is wrong" than merely fixing the referent of "Bullying" then Ridge owes us an account of what this additional role for the belief is and how it works. It seems unlikely that he could do this without ending up as a cognitivist. In addition, note that in any event the Frege-Geach point about the absence of equivocation can be run in terms of truth-conditions themselves rather than in terms of meaning. If "bullying is wrong" as it appears in **1** has different truth-conditions from its appearance in the antecedent in **2**, then the argument from **1** and **2** to **3** is still guilty of a fallacy of equivocation: if the belief is not the determinant of truth-conditions here, it is obscure how the ecumenical expressivist can avoid the charge of equivocation. Of course, the ecumenical expressivist may try to avoid this worry by denying that "Bullying is wrong" has truth-conditions: but officially ecumenical expressivism is supposed to be consistent with expressivist views that don't deny that moral utterances are truth-apt, so this move would still involve a significant departure from Ridge's account of ecumenical expressivism. Moreover, even waiving this point and granting the involvement of belief in determining meaning (though not truth-conditions), considerations relating to compositionality still scupper the view. See footnote 116 below.

perfect propriety utter **2** without thereby taking up an attitude of disapproval towards bullying. We no longer have that problem, but now we have another problem: someone could with perfect propriety utter **2** without thereby taking up an attitude of disapproval towards things insofar as they have a certain property N (imagine a Martian anthropologist using **2**).[114]

Moreover, the ecumenical expressivist who takes this line now seems pushed to account for the *compositionality* of moral sentences. The Frege-Geach problem, after all, amounts to more than just accommodating the validity of intuitively valid inferences. In its most general form it can be formulated as follows: give an account of the meaning of moral sentences (such as "Bullying is wrong") in terms of which they contribute to the meanings of complex expressions in which they appear (such as the antecedents of conditionals) in such a way that intuitively valid inferences involving them are not impugned (by, for

[114] For my purposes here, we can take a Martian anthropologist to be an agent who is a global agnostic about 1st order moral questions but who nonetheless uses conditionals like **2** to record facts about the structure of human moral practice. In an appendix to his 2006, Ridge in fact considers this potential counterexample to ecumenical expressivism but suggests that the ecumenical expressivist can deal with it by viewing non-atomic moral judgements as in a sense multiply realizable. There is the standard way in which ordinary agents make moral judgements – for example by disapproving of things insofar as they have some property N and believing that if bullying has N then turning a blind eye to bullying has N too. But there is another way, applying to the case of global agnostics about 1st order moral questions: "it is most plausible within the framework of Ecumenical Expressivism to understand such an agent as taking a stand against the approval of certain sorts of observers – those observers who would simultaneously [disapprove of bullying] but at one and the same time [fail to disapprove of turning a blind eye to bullying]. In the Ecumenical framework, this will amount to the agent's adopting a perfectly general noncognitive attitude, here an attitude of refusal – refusal to approve of an observer unless it has certain features (once again we have a belief with anaphoric reference back to the content of a noncognitive attitude) preclude simultaneously [disapproving of bullying] while [failing to disapprove of turning a blind eye to bullying]" (Ridge 2006, 335). It is not clear that this convincingly deals with the counterexample. For one thing, what justifies Ridge in saying that these are two ways of making the same non-atomic moral judgement rather than ways of making different types of judgement? And even if Ridge can answer this first question, can't we just as easily imagine a Martian anthropologist who records facts about the structure of human morality while being globally agnostic about the moral status of combinations of attitudes as well as globally agnostic about 1st order moral questions?

instance, committing fallacies of equivocation). So, in order to have a viable solution, the proposed account of the meanings of **1**, **2** and **3** should be capable of yielding the result that the meaning of **2** is a function of the meaning of the conditional "If … then …" together with the meanings of **1** and **3**. But since in all three it is the same desire-like attitude that is preserving meaning, compositionality simply goes by the board. Compare this with the non-ecumenical expressivist attempt to solve the problem in terms of higher order attitudes: the meaning of **2** is given by B!([B!(Bullying)]; - [B!(turning a blind eye to bullying)]), which is a function of the sentiments that give the meaning of the antecedent and the consequent together with an account of the semantics of the conditional in terms of the expression of higher-order attitudes. Here we have at least an attempt at outlining a functional relationship between the meaning of the conditional **2** and the meaning of its constituents. In comparison, since there is only a *single* sentiment in play in the ecumenical expressivist story, there is simply no specification of a functional relationship of the sort that would potentially subserve an explanation of compositionality. Of course, since the beliefs involved, unlike the sentiments, have specific truth-conditions, they would be able to enter into an explanation of compositionality. But as noted above, the ecumenical expressivist cannot take the truth-conditions of the beliefs to give the meanings of the relevant moral judgements on pain of losing the distinctively expressivist component of his ecumenical view.

On reflection, there seems to be something very odd about Ridge's proposed solution. On the one hand, Ridge wants to say that it is the relevant desire-like attitude that constitutes the distinctively moral content of "Bullying is wrong". On the other hand however, as we've just seen, the desire-like attitude plays no role whatsoever in explaining how the presence of "Bullying is wrong" contributes to the meanings of complex sentences in which

it appears: it is just an extra wheel.[115] Here we might be reminded of Davidson's comment on meanings as entities:

> Paradoxically, the one thing meanings do not seem to do is oil the wheels of a theory of meaning – at least as long as we require of such a theory that it non-trivially give the meaning of every sentence in the language. My objection to meanings in the theory of meaning is not that they are abstract or that their identity conditions are obscure, but that they have no demonstrated use (Davidson 1967: 20-21).

If the ubiquitous desire-like attitude in Ridge's account has no demonstrated use in the account of how the meaning of "Bullying is wrong" contributes to the meanings of more complex sentences that contain it as a constituent, does this not suggest that it is a mistake to see that attitude as playing a role in constituting the meaning of "bullying is wrong" in the first place?[116],[117] Ridge may retreat by suggesting that although the desire-like attitude

[115] The "extra wheel" terminology comes from Mark Schroeder's discussion of Daniel Boisvert's "expressive-assertivism". See Schroeder 2010, chapter 10.

[116] This consideration is relevant to the suggestion considered in footnote 113 above. Even if we waive the objection expressed there and allow the ecumenical expressivist the idea that the belief that bullying is N plays a role in determining the meaning of "bullying is wrong" as it appears in both **1** and the antecedent of **2**, the ubiquitous desire-like attitude that is expressed along with the belief plays no role whatsoever in the account of how the meaning of **2** is determined by the meanings of **1**, **3** and the conditional operator. (Independently, Neil Sinclair raises a similar worry about the account of negation that Mark Schroeder has developed on behalf of expressivism. See Sinclair 2011). Moreover, even if Ridge could overcome this problem and find a real semantic role for the desire-like attitude in the moral case, given the implausibility of involving such a desire-like attitude in an account of the semantics of *non-moral* conditionals, ecumenical expressivism would face a further problem. This is the problem of dealing with so-called mixed-inferences (where we are dealing with arguments with both evaluative premises and premises containing no evaluative language), similar to the problem affecting non-ecumenical expressivist views identified in Hale 1986, by Schroeder above, and the problem affecting views which posit two different truth predicates for evaluative and non-evaluative discourse, identified by Christine Tappolet's 1997 (although as seen above this particular version of the worry could be misplaced). Ridge could respond to *this* worry by arguing for an expressivist characterisation of the conditional operator in general, thus pushing his own position towards a more global expressivism. However, it's arguable that this sort of view fails to mesh with the spirit driving evaluative expressivism

plays no role in compositionality, arguments in moral psychology independently license viewing it as a component of the meaning of "Bullying is wrong". This position may not be logically inconsistent, but it is nonetheless problematic. Where a view in moral psychology implies that some feature is part of the meaning of an expression even though it plays no role in accounting for its contribution to the meanings of complex sentences containing it, what we have is effectively a reason for reconsidering the relevant claim in moral psychology. The retreat in effect takes us back towards a non-ecumenical cognitivist form of motivational externalism.

It seems, then, that the mere invocation of beliefs that make "anaphoric pronominal back-reference" in the manner envisaged by Ridge yields only the superficial appearance of a solution to the Frege-Geach problem, and that the ecumenical expressivist cannot repair this problem without relinquishing the distinctively expressivist part of his position. Once we

---

– to give a metaphysically and epistemologically solvent account of moral discourse which shows that it not committed to the kind of heavy-weight entities (properties, facts and so on) that our non-evaluative discourse is. The Frege-Geach problem puts this kind of contrast under pressure by pointing out that the two types of discourse interact in such a way that makes it difficult to give evaluative discourse this special treatment. To take this as a reason to offer a global expressivism (where even seemingly obviously non-evaluative utterances are interpreted as expressing commitments) means the expressivist arguably gives up on the contrast between the evaluative and non-evaluative they started with, as we saw above. Whether this sort of view is ultimately viable is, however, beyond the concerns of this section; for making such a move is not congenial to Ridge: global expressivism is not ecumenical expressivism. In addition, the fact that the mixed-inferences problem that afflicted non-ecumenical expressivism still threatens Ridge's view shows how far the latter is from solving the Frege-Geach problem "on the cheap".

[117] It would be no use for Ridge to distinguish between sense and tone (in the way familiar from Frege) and then to see the belief that bullying is N as determining the sense of "bullying is wrong" with "wrong" merely coming in at the level of tone. This view is effectively the same as the "Realist-Expressivism" defended in Copp 2001, and Ridge categorises Copp's view as a form of ecumenical *cognitivism*. Neither can Ridge insist on including both the belief and desire-like attitude as components at the level of sense: this would be to surrender the idea that there are cognitivist and expressivist forms of ecumenicism, and also appears to collapse into a form of anti-Humeanism.

are clear about what is required to solve the Frege-Geach problem, then, we see that the ecumenical expressivist solution offered by Ridge is in fact no solution at all.

It might be worthwhile at this point to pause briefly in order to explain how this objection to Ridge's attempted solution of the Frege-Geach problem differs from an objection that has been developed by Mark Schroeder (Schroeder 2009, forthcoming).

Schroeder's objection starts out from the observation that Ridge's account of moral sentences sees them as involving a kind of sentential anaphora. "Bullying is wrong", for example, is held by Ridge to express (a) a desire-like sentiment of disapproval towards action-types insofar as they possess a certain property and (b) a belief that bullying possesses *that* property. The pronoun in (b) is anaphoric on the reference to the property in (b). Now consider the following:

(19) Superman flies.

(20) If Clark Kent flies then I'm a walrus. So,

(21) I'm a walrus.

This is truth-preserving but not logically valid: someone who isn't party to the substantive information that Superman and Clark Kent is the same man could rationally accept **19** and **20** and deny **21**. Likewise for:

(22) Superman – he flies.

(23) But Clark Kent – if he flies then I'm a walrus. So,

(24) I'm a walrus.

This is truth-preserving given the preferred interpretation of "Superman" and "Clark Kent", but for logical validity we require truth-preservingness in *any* model, not just in the preferred interpretation.

According to Schroeder the moral modus ponens argument is akin to these because seeing that the moral modus pones argument is truth-preserving on Ridge's interpretation requires knowledge of the substantive assumption that moral sentences all express the same desire-like attitude. Without that assumption there is no guarantee that the belief expressed in the first premise of the moral modus ponens is the same as that expressed in the antecedent of the conditional second premise. So Ridge has not captured the logical validity of moral modus ponens and so has failed to solve the Frege-Geach problem "on the cheap".

Schroeder's objection is subtle and deserves more careful scrutiny than can be given it here. However, it seems Schroeder's objection is somewhat narrower than that presented in some of the influential presentations of the Frege-Geach problem in its application to Blackburn's quasi-realism, such as Hale (1993a) and Wright (1988). There the objection seems to be that Blackburn cannot frame the moral modus ponens argument in a way that satisfies some expressivist surrogate of the notion of truth-preservingness. The moral modus ponens argument on Blackburn's account doesn't do this because it is no better than an argument that equivocates and which has true premises and a false conclusion – and which is

therefore *a fortiori* not truth-preserving (or possessed of a surrogate thereof). The objection raised against Ridge in the text concerns this more general worry: the moral modus ponens argument on Ridge's interpretation is not *even* truth-preserving (because of its failure to deal with the worry about equivocation) and is therefore not logically valid (since being truth-preserving is a necessary – though not sufficient – condition for logical validity).[118]

## 4.45 Ecumenical Cognitivism Does Not Capture Judgement Internalism

According to Ridge, ecumenical cognitivism can yield a form of cognitivism that simultaneously respects judgement internalism and the Humean theory of motivation. According to ecumenical cognitivism (i) moral judgements express beliefs, (ii) there is a conceptual guarantee that if the truth-conditions of the belief are met, so are the truth-conditions of the judgement, and (iii) moral judgements express desires. The position is distinct from ecumenical expressivism in virtue of (ii) – which the ecumenical expressivist denies – and in virtue of its denying that the desire that is mentioned in (iii) is what makes

[118] In a sense, then, I am suggesting that although the Frege-Geach problem suggests that the expressivist cannot capture the logical validity of e.g. the standard moral modus ponens example, it does so via suggesting that the expressivist cannot account even for the fact that such arguments are truth-preserving. Note, too, that the Frege-Geach problem doesn't appear to have anything to do with putatively deductively valid arguments. Consider the following argument. (i) It would be right to buy *The Big Issue* from Mark when one passes him. (ii) In the past, Jim has almost always done the right thing. So, (iii) Jim will buy *The Big Issue* from Mark when he passes him. This is surely a good, non-deductive argument, but the Frege point would still apply: on the face of it looks as though an expressivist would be unable to avoid the charge that there is an equivocation on "right" between (i) and (ii) and that the argument is therefore guilty of a fallacy. So the fundamental point of the Frege-Geach problem seems not to concern *formal* validity in the way envisaged in Schroeder's objection to Ridge. (This is not to say that Schroeder's objection to Ridge's account of formal validity is not a good one, just that it is not the most fundamental problem in the vicinity).

the judgement a specifically moral judgement. The position is distinct from non-ecumenical cognitivism in virtue of (iii), and it is (iii) that makes the position a form of judgement internalism. Throughout, "belief" and "desire" are Humean in character: there is no postulation of necessary connections between beliefs and desires, and no postulation of "besire" like states that simultaneously display belief-like and desire-like characteristics. I shall argue that ecumenical cognitivism along these lines is not an attractive metaethical position.

Recall that judgement internalism is the view that as a matter of conceptual necessity, an agent who makes a moral judgement will be motivated to act in accord with it. Since desires are motivational states, and since ecumenical cognitivism incorporates the claim that moral judgements express desires, it seems that ecumenical cognitivism is a form of judgement internalism: if moral judgements express motivational states, after all, it should be no surprise that there is a conceptual guarantee that an agent making a moral judgement is motivated to act accordingly.

But what is meant by "express" as it appears in (iii)? There are two possibilities. Either the claim that moral judgement J expresses a desire D requires a *conceptual* connection between J and D or, alternatively, only a less than conceptual connection is required. Suppose that a conceptual connection is required. Then, judgement internalism certainly appears to have been captured: if the presence of a motivational state is a requirement on being able to make a judgement with moral content, there will be a conceptual guarantee, of the sort required by judgement internalism, that an agent who makes a moral judgement will be motivated to act accordingly. But note that in (i) "express" must involve a conceptual

connection between the moral judgement and the relevant moral belief: the ecumenical cognitivist is distinguished from the ecumenical expressivist in virtue of the fact that the former but not the latter postulates a *conceptual guarantee* that the truth-conditions of the belief give the truth-conditions of the judgement. But now we have a moral judgement simultaneously sustaining a conceptually necessary relation to a belief, on the one hand, and a desire on the other. But if the judgement is conceptually linked to a belief, and also conceptually linked to a desire, it follows that the moral judgement expresses a besire. This is precisely what the Humean theory of motivation disallows: the existence of this relation of necessary connection would effectively yield the existence of besire-like states that simultaneously display belief-like and desire-like characteristics. [119]

[119] It might be objected that this does not necessarily constitute a rejection of Humeanism. So long as there is no relation of necessary connection between beliefs of type-B and desires of type-D, a particular judgement may sustain an internal relation to both, because tokens of type-B can exist in the absence of tokens of type-D (and vice versa), in contexts where the relevant judgement is not being made. (this point comes from Neil Sinclair). It is not obvious that there is no departure from Humeanism here. Humeanism inter alia rules out the existence of "besires", where a besire is "a unitary mental state which has properties of both belief and desire" (Altham 1984: 284). Granted, there may be no unitary psychological state with both belief-like and desire-like features, but there are still *unitary contentful entities* – moral *judgements* – that have intrinsically belief-like features and intrinsically desire-like features. Arguably, whatever problems beset Anti-Humeanism formulated in terms of psychological states will also beset Anti-Humeanism formulated in terms of judgements. For example, moral judgements would appear to have two incompatible "directions of fit" (Smith 1994: 117-118); and it would appear to be impossible to judge that X is right yet fail to be motivated to X, so that it would not be possible to accommodate the case of e.g. the depressive who is fully aware of the moral significance of a praiseworthy course of action but nevertheless lacks the motivation to pursue it (Smith 1994: 120). In addition, the argument against anti-Humeanism in chapter 1 of (Smith 1994: 7-8) – that on an Anti-Humean view desires would be derivatively truth-assessable – could be reapplied to judgements that sustain internal relations to both beliefs and desires. Call a judgement with an internal relation to a belief a belief-implicating judgement and a judgement with an internal relation to a desire a desire-implicating judgement. If some belief-implicating judgements are also desire-implicating judgements, then some desires would be derivatively assessable in terms of truth and falsehood: for we could count the desire implicated by the judgement as true whenever the belief implicated by the judgement is true. It thus seems that Ridge's position rejects the spirit – if not the letter – of Humean views of motivation.

What if, on the other hand, the claim (iii) that moral judgements express desires does not carry connotations of a conceptually necessary relation between the presence of the desire and the judgement's having specifically moral content? Then there will be no imputation of the existence of besires or a clash with Humeanism about motivation. But now we appear to have lost the crucial claim made by judgement internalism: this was not that as a matter of empirical fact an agent who makes a moral judgement will be motivated to act accordingly but that *as a matter of conceptual necessity* such an agent will be motivated to act accordingly. If the notion of expression that appears in (iii) does not generate a conceptual relation between moral judgement and desire, therefore, it is obscure how the ecumenical cognitivist view can accommodate judgement internalism.

Overall, then, the "ecumenical cognitivist" faces the following dilemma. If the notion of expression that appears in (i) and (iii) in each case generates a conceptual connection between moral judgement and the respective types of mental state, we get a relationship of conceptually necessary connection between beliefs and desires that is directly at odds with the Humean theory of motivation; if the notion of expression that figures in (iii) does not generate a conceptual relation between moral judgement and desire, we avoid a conflict with Humeanism, but are no longer well placed to embrace judgement internalism; while if the notion of expression that appears in (i) does not generate a conceptual connection between moral judgement and belief, it is no longer clear that we have a version of cognitivism. Either way, there appears to be no space for an ecumenical cognitivist view capable of combining (i), (ii) and (iii).

## 4.46 Diagnosis

According to Ridge, ecumenical expressivism is an expressivist view capable of solving the Frege-Geach problem on the cheap, and ecumenical cognitivism is a cognitivist view capable of meshing judgement internalism and the Humean theory of motivation. Ridge suggests that this requires a transformation of metaethics in which the Frege-Geach problem and issues about moral motivation are moved off-stage and in which the focus is purely on the open-question argument and the supervenience of the moral on the natural. I have argued that there are no ecumenical views – expressivist or cognitivist – capable of neutralising the Frege-Geach problem on the cheap or squaring judgement internalism with the Humean theory of motivation.

We have also seen above that Ridge's account severs a link between moral psychology and semantics – we end up with an account of the semantic properties of a moral term that does not use an element found in our account of the moral psychology. Arguably, metaethical theorising is in part the search for an *integrated* account of moral semantics and moral psychology. The position to which Ridge is imagined to retreat here would constitute an abandonment of this search for integration, and so a change in the terms of metaethical debate. It would be hard for Ridge to represent himself as joining the debate, leaving the terms of the debate in place, and showing how some of the thorny issues that arise within it can be solved "on the cheap". Note that the other main metaethical views – non-ecumenical cognitivism, non-ecumenical expressivism, and anti-Humeanism – all aspire to the kind of integrated view the retreat would give up on: the moral psychology and moral semantics do not just sit side by side, but are integrated in the sense that the semantic features postulated in their moral psychologies actually play a role in their moral semantics. (We can thus see

metaethics as in part at attempt to respond to a kind of challenge of a piece with what Peacocke calls "the Integration Challenge"(Peacocke 2000), and the point here is that a retreat by Ridge of the sort canvassed in the text would in effect amount to an admission that the Challenge cannot be met).

Drawing attention to this Challenge allows us to explain two features of metaethical theorising. First, the view most similar to hybrid views is the so-called 'besire' theory (Altham, 1984). This view claims that moral judgements express unitary mental states with both representational and motivational content. This view may seem bizarre, but we can see why someone trying to incorporate the strengths of both expressivism and cognitivism would be drawn to it – it manages to secure those by respecting the desideratum that a metaethical account should offer a unified picture of moral psychology and semantics.

Second, early metaethical views in the expressivist tradition (like Ayer's emotivism) acknowledged that something like the beliefs that Ridge claims are expressed by moral judgements along with the desire-like attitudes would be floating around in the moral judge's psychology. Ayer even claims that if we knew enough about the types of things that a particular agent judged to be bad (say that they typically thought that only green things were morally abhorrent) then you could work out at least some of their beliefs about an object from their moral judgements (if they tell you that X is evil, chances are that they believe that X is green). However, very few in the expressivist tradition sought to make use of these beliefs to solve the Frege-Geach problem until recently. It could be that Ayer and company did just overlook the fact that they had all the resources that they needed ready at hand. However, it could also be the case that they did not exploit this opportunity because they were trying to respect the desideratum that one's accounts of moral psychology and

semantics should be unified in the appropriate way. Taking this second line allows us to explain Blackburn's remark that:

> We can see that it does not matter at all if an utterance is descriptive as well as expressive, provided that its distinctive meaning is expressive. It is the *extra import* making the term evaluative as well as descriptive, which must be given an expressive role. It is only if that involves an extra truth-condition that expressivism about values is impugned (Blackburn 1984: 169-70).

Why does Blackburn not simply help himself to the relevant descriptive contents to get a cheap solution to the Frege-Geach problem in such cases? One explanation would be that he anticipated the main point that I made against Ridge: namely, that since the truth-conditions of the associated belief are not viewed by the ecumenical expressivist as giving the distinctive meaning of the moral utterance, or as making the utterance evaluative, they cannot be invoked to defuse the worry about equivocation that takes centre-stage in the proper presentation of the Frege-Geach problem.

I will now turn (briefly) to two other attempts to provide a hybrid metaethical account.

## 4.5 Realist-Expressivism and Neo-Expressivism

Recall that hybrid metaethical theorists are on the lookout for a position that will allow them to secure judgement internalism whilst having a solution to the Frege-Geach problem, all whilst keeping the rest of the terms of metaethical debate mostly untouched (in particular retaining a commitment to the Humean theory of motivation). Ridge's ecumenical expressivism attempts to do this by claiming that moral judgements express beliefs as well as desires. However, because these beliefs are not properly used to characterise the meaning of the moral judgement in question the account doesn't secure a solution to the Frege-Geach problem that leaves everything else in place. Here I will briefly consider two alternative ways of hybridising: first, we could claim that having the right sort of motivational state is a conventional implicature of making a moral judgement, as in David Copp's realist-expressivism[120]; second, we could claim it is the act of making moral claims that is expressive of a motivational attitude as in Dorit Bar-On and Matthew Chrisman's neo-expressivism.

The main point I wish to make about these theories is not so much that they are wrong (although we shall see shortly that each faces some problems), but rather that they are not properly hybrid. Both end up, in effect, giving something like a pragmatic explanation of the motivational import of moral judgements – being in the right kind of motivational state is not constitutive of making a moral judgement, instead being in the right motivational state is conversationally implicated by making a moral assertion, or expressed in the act of

[120] See also Stephen Barker's (2000). Stephen Finlay (2004, 2005) considers a similar account where the implicature is conversational rather than conventional.

making that assertion. What this means is that the motivational state does not have a role in characterising the meaning of a moral judgement. Instead it is an additional element on top of the semantic content. This means that the connection between moral judgement and motivation that realist-expressivism or neo-expressivism secures will be weaker than that secured by full-blown expressivism – because the motivational state is not a constituent of the meaning of the moral judgement, the connection will not be the strong conceptual one originally posited. This is why Copp is explicit that "realist expressivism is entirely compatible with *externalism*" (2001, 3, emphasis in original). All Copp is attempting to do is give an explanation of the intuitions that lead people to accept internalism.

Thus, this type of hybrid view faces the opposite problem of Ridge's ecumenical expressivism – because the ecumenical expressivism doesn't really use the belief they claim is expressed by a moral judgement to characterise the meaning of that judgement they cannot offer a solution to the Frege-Geach problem. Because the realist-expressivist and neo-expressivist do not use the desire to characterise the semantic content of the moral judgement, they cannot secure full-blown judgement internalism. What they can try to do instead is offer an account that explains why there is typically a link between moral judgement and motivation by turning attention to the pragmatics of moral judgement. However, we have seen how this avenue is open to someone who characterises themselves as merely realist (Finlay's analytic naturalism above §3.31).

This does not mean that Copp and Bar-On and Chrisman's views are without merit. It is one thing for the realist to claim that the motivational import of moral judgements can be given a pragmatic explanation, it's another thing to actually work out the details of that

explanation.[121] All it means is that this is what Copp and Bar-On and Chrisman's proposals amount to – a fleshing out of how realists can use the pragmatics of making a moral assertion to explain away internalist intuitions, rather than a truly hybrid account that secures judgement internalism whilst keeping the resources needed to solve the Frege-Geach problem. What examination of hybrid theories has revealed, I think, is that these theories are not truly hybrid in the sense of combining elements of cognitivist and non-cognitivist semantics. Sure enough, they do borrow elements from both of these semantics, but do not use them as elements in a hybrid *semantics*. Instead one element or other plays little role in the semantic theory, merely featuring to secure the element of the non-hybrid semantic theory that cannot be captured by the other non-hybrid semantic theory. What this shows is that if you want to secure the advantages of the non-hybrid versions of these semantic theories (judgement internalism for expressivism and a simple solution to the Frege-Geach problem for the cognitivist) you have to actually buy fully into those semantic theories.

One final point that is worth making before turning to the details of Copp and Bar-On and Chrisman's proposals is that if the above is right then the realist-leaning theories offered by Copp and Bar-On and Chrisman look to be in somewhat better shape than Ridge's ecumenical expressivism. This is because, as we saw in the discussion of van Roojen above (§2.25), most modern internalists loosen their internalism requirement enough that an explanation of internalist intuitions should be enough to satisfy most. However, the Frege-Geach problem is not something that you can 'cheat' in this way. We have seen how the Frege-Geach problem gets to the heart of how moral discourse interacts with non-moral

[121] As Finlay (forthcoming) points out, it's hand-waving towards pragmatic explanations that has given pragmatic explanations in general a bad name (Ch. 2).

discourse, and that if you do not use the truth-conditions of the belief you invoke in your hybrid theory to characterise the semantic content of moral judgements you are going to have a very hard time explaining the compositionality of moral language.

Turning now to the details of the two proposals under consideration. Copp's realist-expressivism is inspired by Frege's discussion of colouring. For Frege, "This dog howled all night" and "This cur howled all night" state the same thought – they have the same truth-conditions. The differ, however, in what they imply about a speaker's attitude towards the dog in question – using 'cur' implies I feel an attitude of contempt towards the dog, just as if I said 'dog' with a contemptuous tone of voice[122]. This difference in implication Frege calls colouring. This provides Copp with a model for moral judgements – here we have a case where an utterance 'expresses' (in the sense of conveys that a person has both) a belief and a non-cognitive attitude. Thus Copp introduces the notion of 'Frege-expression'. A use of a term Frege-expresses a state of mind when it is a matter of linguistic convention that using that term typically conveys that the user of the term is in that state of mind. Copp's proposal then is that moral judgements express a run of the mill belief (in the standard sense of 'express') as well as Frege-express a non-cognitive desire-like state.

How does this work? Copp's view, simply put, is that moral judgements express two propositions. Take the example 'Bullying is wrong'. This expresses both: that bullying is forbidden by a relevant set of standards; and that I am in the non-cognitive desire like state

[122] Copp differs from Frege in that he wants to use 'meaning' to refer to what is communicated or conveyed, as a matter of convention, by the use of a sentence; rather than restricting 'meaning' to the semantic content of the utterance. On this way of doing things colouring becomes part of the meaning of an utterance. Nothing I say here turns on this different way of carving up the terrain, although I will continue to, as above, use 'meaning' to mean semantic content.

that is acceptance of a set of standards that forbids bullying. This second proposition is conventionally implicated, rather than being part of the semantic content. However, it allows competent speakers to convey that they are in the relevant non-cognitive desire like state, so that non-cognitive desire-like state is part of the 'meaning' of the moral judgement, in the expansive sense of 'meaning' introduced by Copp.

This view has the advantage of offering a simple solution to the Frege-Geach problem – the first proposition expressed (in the standard sense of 'express') by a moral judgement has the right kind of truth-conditions to feature in an explanation of the validity of moral modus ponens arguments, etc. What does the view say about internalism? Well, the second proposition, Frege-expressed by the moral judgement, implies that you are in the non-cognitive desire like state of accepting a set of standards that forbids bullying. Thus, it is *inappropriate* to say that something is wrong when you don't accept a set of standards that forbids it. Thus, typically, when someone says that something is wrong they will be motivated to avoid that thing (as they accept a set of standards that forbids it, where accepting a set of standards is understood to mean being in a non-cognitive desire-like state).

This provides Copp with a more sophisticated characterisation of the amoralist than is available to traditional internalism. Remember that the traditional internalist says that the amoralist is not making a genuine moral judgement - they cannot grasp the meaning of the terms involved. Copp, in contrast, can say that when the amoralist judges that something is good without being motivated to pursue it they are saying something that is true, though inappropriate. It is inappropriate because it misleads their interlocutor into thinking that the

amoralist subscribes to a set of standards that mandates pursuing that object as this is what uses of the term 'good' Frege-expresses. At this point we could object to Copp by saying that his characterisation of the amoralist really just collapse back into the old response – this is because on Copp's account the non-cognitive desire-like state Frege-expressed *is* part of the meaning of the judgement. However, this criticism seems to be a mere artefact of Copp's more expansive conception of 'meaning' – what's important is that the amoralist can grasp the semantic content of the terms in question, it's just that they do something inappropriate when they put them together.

Another criticism is that on Copp's account it looks like we don't ever get genuine disagreement. Suppose that you say 'Bullying is wrong' and I say 'Bullying is not wrong'. On Copp's account the primary propositions expressed by these utterances are, respectively, <Bullying is forbidden on some set of standards> and <Bullying is not forbidden on some set of standards>. As long as there are two different set of standards available (one of which forbids bullying and another that doesn't) then what we have said is not inconsistent. Instead, if we disagree at all, then it's about which sets of standards we should adopt. But this disagreement looks like a normative one. Either we can posit ever ascending sets of standards that permit certain sets of standards and not others; or claim that some set of standards is more authoritative than another; or give something like an expressivist account of the inconsistency of sets of standards. The first option is not viable, and the last is not compatible with the motivations for realist-expressivism. The second suggestion might work, but it looks hard to see what work the notions of standards is now doing – it is inconsistency with the moral facts (about which set of standards is authoritative) which is explaining inconsistency rather than the standards themselves.

This discussion has been overly brief, but I hope this brevity will be of little concern, as we can note that this worry about Copp's position can be assuaged if we draw attention to the fact that it is a symptom merely of Copp's very particular analysis of the primary proposition expressed by the use of a moral sentence. There is nothing in the structure of Copp's account to rule out providing another analysis which does secure the right kind of disagreement.

A more serious problem with Copp's account is that raised by Allan Gibbard against what he calls the 'colouring model' (2003, 168-9). Suppose we are in a situation where you feel disgust towards a particular action – setting a cat on fire, say. Suppose then that you assert 'That was cruel'. I, being a reprehensible character, feel no such sense of disgust. Now, as our non-cognitive states do not enter into the truth-conditions of moral judgements, I should be able to (with perfect propriety) respond to your assertion with 'That's true, but I wouldn't put it that way' (thus agreeing with the semantic content of the utterance, whilst cancelling the implicature).[123] However, this looks odd. A more natural response would seem to be 'That's not cruel'. What I think this problem indicates is that the connection between moral judgements and motivation that realist-expressivism secures is too weak to satisfy the expressivist. We can see Gibbard's objection as a compressed argument for a

[123] Typically conventional implicatures are not so easily cancellable, Chrisman and Bar-On suggest against Copp. Instead, they suggest, the realist-expressivist should claim that the relevant implicature is conversational. The problem with this move is it makes judgement internalism a mere artefact of conversation (Bar-On and Chrisman, 2009, 154). Copp, however, argues that conventional implicatures are cancellable, at least in a weaker sense than the one used by Grice, although recently he has developed the term *simplicature* (2009) to denote the kind of implicature that has the features needed to secure his account. At first blush the introduction of this notion strikes me as a little *ad hoc*.

strong form of internalism – one that makes my response to you semantically incorrect. This is the sort of strength of internalism that Copp's account cannot provide.[124]

How does the neo-expressivist account fare? On this account we need to distinguish between two notions of expression. First, a linguistic product containing a moral term (like an utterance or sentence) s-expresses a proposition. Second, the act of making a moral utterance a-expresses a non-cognitive desire-like state (Bar-On and Chrisman, 2009). This distinction is inspired by Dorit Bar-On's account of avowals (Bar-On 2004). The way that this view accounts for the compositionality of moral terms is by using the truth-conditions provided by the proposition s-expressed by the product of an act of assertion (unlike Copp, Bar-On and Chrisman don't tie themselves to any particular analysis of this proposition). Internalism is respected because the act of making an assertion a-expresses a non-cognitive desire-like state. Amoralists, then, understand the meaning of the terms involved in their assertions (in the sense that the products of those acts of assertion – the sentences or utterances produced – can be true, and the thing that they wish to communicate). However, when they assert a moral judgement without feeling properly motivated the act of asserting that judgement is improper – they don't have the relevant state expressed by these acts of assertion. Thus the amoralist is semantically competent (as the externalist holds), yet they display more than a psychological or moral flaw – they do something improper at the level of assertion.

[124] Although see Tim Henning's (2011) for a response to this objection from Gibbard on behalf of a moral realist who exploits a two-dimensional semantic framework.

The first worry about this view is that it is inspired by an account of avowals. Avowals display some puzzling features – they are epistemically secure in a way that encourages something like an expressivist analysis of their contents, yet they are capable of being embedded in similar ways to moral judgements (thus raising parallels of the Frege-Geach problem). Bar-On and Chrisman's account or moral judgements explicitly exploit the structure of Bar-On's account in the case of avowals. However, moral judgements don't seem to exhibit anything like the epistemic security of avowals – why would this be if we give an analysis of the meaning of moral judgements that parallels that of avowals? Bar-On and Chrisman suggest this is accounted for by the different purposes of moral discourse and self-ascriptions of mental states. However, the worry then is that it is not this account that is actually doing the work of explaining the special features of avowals – some other factor needs to be invoked to explain their epistemic security. This is not a problem for their analysis of moral judgements per se, instead it merely indicates a difficulty with trying to underwrite that analysis with a comparison to the case of avowals.

Another worry that you might have is that internalism is not just a constraint on the appropriate utterance of a moral judgement – that there is something wrong with an amoralist who thinks to themselves 'torching a cat is wrong' without feeling any motivation to refrain from cat-torching. Bar-On and Chrisman can, however, deal with this problem by claiming that inner judgements also a-express motivational states.

A more pressing worry, again, is that this version of hybridism doesn't capture judgement internalism. One way to express the worry is as follows: Bar-On and Chrisman's account, as it stands, has nothing to say about the *direction* of motivation of the motivational state a-

expressed by a moral judgement. All we know so far is that moral judgements, to be proper, must be accompanied by some motivational state. But why should judgements of wrongness be associated with a motivation to avoid, and judgements of goodness with a motivation to pursue?[125] It seems just obvious that judgements of goodness should be associated with positive motivations, but the obviousness of the point should not blind us to the fact that it still requires explanation. The expressivist has an answer to this question – it is part of the meaning of 'bad', say, that it expresses some sort of con-attitude towards objects judged to be bad. However, the neo-expressivist (like the realist-expressivist and the ecumenical cognitivist) eschews the tight connections between the meaning of moral terms and motivation to give this sort of explanation.

What I hope to have shown above is that these formulations of hybrid theories are not precisely that. They do not take onboard features from the semantic theories of their non-hybrid cousins then *combine* them into a new hybrid theory. Instead they keep a cognitivist semantics, then add a non-cognitivist element at the level of pragmatics. This means that they cannot provide a tight enough connection between moral judgement and motivation to satiate the non-cognitivist. And once that ambition is junked, these hybrid positions are relegated to a role as merely more sophisticated fleshings out of cognitivism. Even on that level, we've seen that they face a couple of problems.

[125] Another way to put the point is – what does the neo-expressivist have to say about not the amoralist's practice, but about Joyce's agent of pure evil (one who avoids the good and pursues the bad)?

## 4.6 Hybrid Metaethics and Neo-Kantianism

Where does this leave us? We saw that Korsgaard argues that moral statements are neither like statements of fact, nor like 'emotional expletives'. In the previous chapter I explored Korsgaard's problem with the first characterisation of moral statements she rejects (that owing to the moral realist), where the problem was that moral realism fails to capture the practical significance of moral judgements. There I acknowledged that Korsgaard doesn't intend her remarks to refer merely to the motivational import of moral judgements, but argued that taking the complaint in this way allows us to get a grip on what might be wrong with moral realism. I also argued that if we take the problem of practical significance in this way, there are good reasons for the moral realist to be unconcerned. However, suppose, like someone like Blackburn (see his (unpublished)) we were particularly concerned about the problem of practical significance. Then a natural place to look for a solution would be expressivism. However, here we run into a characterisation of moral thought and talk that Korsgaard rejects. Moral judgements are not like 'emotional expletives'.

With the help of Hussain and Shah, I explained how this concern is misguided when stated in this robust way – it ignores the distinction Hussain and Shah bring out between cogitation and cognition. However, unlike Hussain and Shah, I think that Korsgaard was getting her finger on something important – that moral judgements don't seem to function in the same way as expressions of non-cognitive states. This merited discussion of the Frege-Geach problem where we saw there are good reasons to suspect that, taken this way, Korsgaard is right – we can't easily explain the compositionality of moral terms by taking moral judgements to be expressions of sentiments. This led me to an exploration of hybrid

metaethical theories. The idea was: suppose Korsgaard is right that moral judgements have features that are not accommodated by the realist (practical significance) or by the expressivist (the complicated semantic properties moral terms exhibit) then perhaps the best way forward is to have a view that combines elements of both these projects. Thus, we could offer a novel way of interpreting Korsgaard's own position (which has been hard to get a grip on), but even if the position we end up with is not one the neo-Kantian would accept, we would have at least a position which answers the concerns that might drive someone to neo-Kantianism.

However, as we have seen, the type of hybrid theory that privileges moral discourse's practical significance over its realist-seeming elements (ecumenical expressivism) doesn't have the resources to solve the problems with pure expressivism. In contrast, the type of hybrid theory that gives a greater role to truth-conditions and beliefs in characterising the meaning of moral judgements (neo-expressivism and realist-expressivism) fail to sustain a connection between moral motivation and judgement that is tight enough to please someone bothered by the issue of practical significance (in effect, hybrid views of this type are merely sophisticated versions of the moral realisms we canvassed in the last section. There I argued that these views were adequate and could avoid the problems posed by the neo-Kantian. However, if you were unmoved by that defence then the hybrid versions or realism seem to be no more advanced on this dimension).

So this is how things look: I have attempted to show that moral realists should be untroubled by Korsgaard's criticisms. Expressivists, however, are put under more pressure by worries that could be underlying Korsgaard's criticism of their views. What we have yet

to do is take a look at Korsgaard's own, positive, metaethical position. Even if Korsgaard's complaints against realism and expressivism do not mark out a distinctive position for her to occupy, and even if they can be resisted, we should still consider her own view. It could be in the statement of that that we find a clearly delineated, novel, position. In addition, even if none of Korsgaard's rivals fail for the reasons she provides, we could still find that the neo-Kantian's position is preferable. We won't have a knock-down argument that it must be true (because its rivals do not fail as spectacularly as Korsgaard supposes), but perhaps neo-Kantian constructivism offers us an account of moral thought and talk that is better that its rivals in terms of theoretical virtues like simplicity, generality, and explanatory power. It is to consideration of the neo-Kantian's own position that I now turn.

# CHAPTER FIVE: NEO-KANTIAN CONSTRUCTIVISM

When we come to Korsgaard's own metaethical position we run into difficulties characterising her view. We saw before that she talks of wanting to transcend the distinction between cognitivism and non-cognitivism, and some problems with what that proposal could even mean. When she states her constructivist[126] position most explicitly (2003) she ends up saying things that make little literal sense.[127] What we do know is that she takes her view to be inspired by John Rawls's neo-Kantian constructivism in the case of justice. However, there are well known problems with trying to use Rawls's framework to give a metaethical account (see Brink's (1989) fourth appendix). What I try to do here is offer a novel way of characterising Korsgaard's view that both secures some of the features she wants from her metaethics but which is also clear enough to evaluate. I argue that we should construe neo-Kantian constructivism as a particular form of cognitivist anti-realism, along the model for judgement-dependent qualities offered by Crispin Wright (§5.1). I then argue that if we do things this way the viability of neo-Kantian constructivism depends upon being able to give a derivation of the categorical imperative, and I turn to Korsgaard's attempt to do this in §5.2, where we see how she attempts to show that the categorical imperative is binding for us in virtue of various claims about the nature of agency. This attempt, I argue, in turn depends on another doctrine – constitutivism – which I explore in §5.3. The viability of constitutivism depends upon claims about the inescapability of agency

[126] There are other ways of characterising constructivism (for example, the work of Sharon Street (2008, 2010) Lenman (2008, 27-8) also offers a clear framework on the constructivist's behalf. In order to try to make at least some progress I have decided to concentrate on trying to get clear about Korsgaard's own, distinctive position.

[127] For example, she repeatedly states that "Concepts refer to solutions of problems" (e.g. 2003, 117).

(§5.4). Finally (§5.5) I consider whether Korsgaard has the resources for responding to a standard objection to Kantian moral theories.

## 5.1 Neo-Kantian Constructivism and Judgement-Dependence

Here I wish to suggest that we can generate a coherent view that provides some of the features that Korsgaard wants from a metaethical theory if we model neo-Kantian constructivism on the account of judgement-dependent[128] qualities offered by Crispin Wright. What I think animates Korsgaard's search for a version of procedural realism that doesn't embrace substantive moral realism is a distaste for moral realism attempting to ground the normative force of obligations in a metaphysical way – in order to explain the normative force of an obligation the realist, she thinks[129], has to cite the existence of a special kind of fact, one that is intrinsically normative[130]. What Korsgaard wants to do instead is give an account of the correctness of moral judgements that depends purely on our access to the right kind of procedure. I will now outline the Wrightian judgement-dependence framework and try to show how it can secure for Korsgaard those two ambitions, whilst giving us a clear proposal to evaluate.

Secondary qualities have been the target of a large degree of speculation in philosophy since the distinction between primary and secondary qualities was popularised by thinkers

[128] Also called 'response-dependent'.

[129] As we saw above in discussion of the anti-voluntarist argument it's not clear that the moral realist is restricted to this type of explanation alone.

[130] Which then opens up space for the generalised anti-voluntarist argument I looked at in (§3.2).

like Galileo and Locke. Secondary qualities are supposed to have some sort of intermediate metaphysical status – there is typically supposed to be a fact of the matter about whether something is red, say, even though redness as such does not belong to the objective fabric of the world. These qualities are 'mind-dependent' in a particular way – whether a particular object is red or not does not depend upon an individual perceiver's judgements, but the existence of redness in general depends upon, in some way, our perceptual responses. In addition, secondary qualities possess strange features, that might make them apt for comparison to moral qualities. Against Mackie's complaint that there cannot be objective goodness in the world, for example, because it would be too metaphysically queer due to having to-be-pursuedness built into its nature, John McDowell argues that red things have to-be-seen-as-redness built into their nature. If this sort of comparison can be made to work then we look to have a view that would appeal to the neo-Kantian constructivist – in the colour case we have correct answers to questions like 'is this object red?', but the truth of those answers is not grounded in purely mind-independent features of the world, but rather in the effects that objects in the world produce on our minds.

What I've offered above is rather vague and highly metaphorical. Such features have dogged the debate around secondary qualities since its inception. In an attempt to clear the area up Crispin Wright (1992) has suggested that we think not in terms of primary and secondary qualities but in terms of judgement-dependent and judgement-independent qualities. A quality is judgement-dependent when our best judgements about that quality play an extension-determining role for that quality. Our best judgements about judgement-independent qualities, in contrast, play an extension-tracking role. In other words, the truth about the extension of a judgement-dependent quality is constituted by our best judgements about that quality – the truth in that area cannot outstrip our best judgements. In contrast in

the case of judgement-independent qualities it is always possible for the truth of that area to go beyond our best judgements. By thinking of qualities in this way we get to see why we can view secondary qualities as somehow less real than primary qualities – the secondary qualities are judgement-dependent (and thus the truths about their instantiations cannot go beyond our best judgements), the primary judgement-independent (where the relevant truths *can* go beyond our best judgements). This could then feed into a Dummettian-inspired characterisation of the distinction between realism and anti-realism where realism is identified by the claim that truth in a particular discourse can outstrip what we have evidence for, and anti-realism by the denial of that claim.[131]

Wright goes on to provide a framework for telling when a quality is judgement-dependent or –independent. He sets up the framework by considering the qualities of colour and shape. The idea here is that any adequate methodology for telling apart judgement-dependent and judgement-independent qualities should place colour on one side and shape on the other, given that these are paradigmatic instances of secondary and primary qualities. How it works is this: we first try to discern the ideal conditions (**C**) under which a suitable subject's (**S**) judgement about the extension of some particular term are maximally credible. What we are looking for is the *best* conditions for making judgements of the relevant type. In the case of colour, for example, it's unlikely that the judgements of someone who is colour-blind looking at colour samples in a darkened room would play an extension determining role.

[131] In fact, The connections between judgement-independence and evidence transcendent truth is, according to Wright, potentially more complicated than this. I offer this loose, potentially false characterisation to give more of a flavour of the way in which judgement-dependence is anti-realist

What we then do is use those **C** conditions to generate a provisional equation of the following form:[132]

**PE$_{\text{Red}}$**: S were in C → (x is red ↔ S judges that x is red)

**PE$_{\text{Square}}$**: S were in C → (x is square ↔ S judges that x is square)

Once we have done this we check whether these provisional equations pass four tests (in addition to being true and hence extensionally adequate). In order to think that the quality in question is judgement-dependent the provisional equation must be: *a priori;* substantial; fulfil an independence condition; and an extremal condition. Looking at these in turn.

If the truth of the relevant provisional equation were only *a posteriori*, then we'd have no license for claiming that the truth of the discourse in question conceptually depends upon our best judgements regarding that discourse. The provisional equation will be merely *a posteriori* in cases where we are in such a favourable epistemic position that we get things right all the time. This could be the case even where the states of affairs we are thinking of are constituted entirely independently of our best judgements. For the truth of the provisional equation to be any evidence in favour of a judgement-dependent account of the quality in question it must be knowable *a priori*. (114-7).

[132] Wright moves away from the *basic* equations of his (unpublished) towards casting things in terms of provisional equations to avoid Robert K. Shope's conditional fallacy (see Wright 1992, 117-120).

The **C-**conditions characterised in the provisional equation must be *substantial* in the sense that they are "specified in sufficient detail to incorporate a constructive account of the epistemology of the judgements in question"(112). This is to rule out construing the conditions so that the subject has "whatever it takes" to come to the correct judgement, instead we need a "concrete conception... of what it actually does take." (112).

We also must be able to characterise the relevant **C**-conditions without using the concepts that feature in the relevant judgements (the *independence* condition).[133] This is not to avoid a kind of circularity in the proposed equation, for the equation is only designed to tell us about the dependence of the truth of the discourse in question on our best judgements. It is not a attempt to give an analysis of the associated truth conditions. However, if we did invoke the very concepts we were trying to give a judgement-dependent account of, it would be open to an opponent of the account to ask the judgement-dependence theorist to show that their invocation of those concepts is compatible with the thesis that they are trying to prove – that the extensions of those concepts is determined by our best judgements. In order to avoid this worry we should try to characterise the **C**-conditions so that we do not assume anything about the extension of the relevant concepts.[134] (120-123)

Finally, our provisional equation should respect the *extremal* condition. In effect, our proposal that truth in the relevant area is judgement-dependent should be the *best explanation* of the truth of the provisional equation. This condition, like *a prioricity* one, is

[133] In fact, for Wright at least, we must not use them in a way that presupposes facts about the details of their extensions. It's OK, however, to use them if they are used in other ways – within the scope of intensional operators for example.

[134] Although some accounts of judgement-dependence dispense with this condition: see Pettit's (1991).

needed to distinguish between judgement-dependence and infallibility. Take the example of God. Presumably God, in ideal conditions(!), only judges that something is the case when it is the case. This connection could also be *a priori* (in fact, if the concept of God is the concept of an omniscient being then it could also be analytic), respect the independence condition and the substantiality condition (God can do this because She has special epistemic powers). But, in this case, we would not want to say that the truth of everything is judgement-dependent. Instead we'd have a better explanation of the *a priori* truth of the relevant provisional equations – God's omniscience (similar considerations might hold for the case, pain, that Wright considers, 123-4).

How do things look for $\mathbf{PE_{Red}}$ and $\mathbf{PE_{Square}}$ (recall that if this framework is going to be at all plausible red should come out as judgement-dependent and square as judgement-independent)? Wright argues that we *can* characterise the **C**-conditions for $\mathbf{PE_{Red}}$ in the appropriate way. We will have to mention factors like: the subject is attending to the object in question and is free from distraction; they have a normal (in a statistical sense of 'normal') nervous system; the object is viewed at 12 noon, outside, on an overcast day in Fife[135]; etc. All of these conditions can be stated without assuming anything about the extensions of colour concepts, and they are substantial. In addition the resulting provisional equation is plausibly *a priori* true and we have no other, better, explanation of its *a priori* truth than that redness *is* judgement-dependent. Given that the $\mathbf{PE_{Red}}$ passes the relevant tests we can conclude that colours are judgement-dependent qualities.

[135] Perhaps, if we could provide the extensive argumentation needed, we should be open to the possibility that there are places in England as well that are suitable locations for making good colour judgements.

What about $\mathbf{PE_{Square}}$? Here we face a difficulty, Wright argues. It's obvious that shapes can appear to be different from different angles – a square can look like a diamond or some other quadrangle depending on the angle of viewing. So, in order to come up with plausible **C**-conditions we have to include the subject viewing the object in question from a number of angles. But, if this is going to work we have to assume that the object's shape stays constant throughout this process. Thus we have to assume facts about shape constancy to get the **PE** off the ground, and thus violate the independence condition.[136] Thus, we have to conclude that shape is a judgement-independent quality. Thus the framework yields the desired results, at least in paradigm cases of primary and secondary qualities.[137]

My positive proposal is to explicate neo-Kantian constructivism by using Wright's framework. The basic idea is that there is nothing, in principle, to stop us building a procedure into the **C**-conditions for making our best moral judgements. We have already seen that Korsgaard wants to be a mere procedural realist, in the sense that she thinks that there are right and wrong answers to moral questions, but this is only because we have a suitable procedure for answering those questions. Tying the truth of moral judgements to the outcomes of such a procedure using the judgement-dependence framework would secure this result for Korsgaard – remember that this framework gives an anti-realist gloss

[136] What if the subject in question had eyes on the ends of their fingertips, such that they could view an object from multiple angles at the same time? Then we would not run afoul of the independence condition. What this example demonstrates, however, is that in the shape case the PE is not *a priori* true.

[137] Of course a number of objections have been raised against this way of doing things. For example Wright's motivation for including the extremal condition could be challenged by McDowell's argument against the common sense view of pain (1998, however the motivation for it I provide would still stand). However, I do not have the space to consider all of these. What I want to get clear on instead is how it would work using this framework to investigate morality, in particular whether it can be used to usefully explicate neo-Kantian constructivism, although I will briefly touch on the 'missing explanation argument' below.

on judgement-dependent qualities, whilst retaining the possibility of giving truth-conditions for judgements about those qualities (it's just that the relevant notion of truth will be similarly anti-realist). What would such an account look like? Korsgaard is, as well as a constructivist, a neo-Kantian, so the relevant procedure for her is applying the categorical imperative to the maxim underlying an action. We would thus end up with a PE looking something like this:

> **$PE_{Permissible}$**: S were applying the test of the categorical imperative[138] → (x is permissible ↔ S judges that x is permissible)[139]

I have also noted that Korsgaard seems to be animated, in part, by a distaste for trying to ground the normativity of morality in metaphysics. However, we saw that when she tries to give an argument against moral realism (a position which at least seems to try to explain the normative force of moral obligations using metaphysics) the argument seems to fail. We have also seen that in her more recent work (2003) she complains not so much that moral realism is false, just that it doesn't get to the heart of what matters. Thinking about constructivism on the model of Wright-style judgement dependence gives the constructivist a way to express their antipathy towards metaphysics in a coherent way. We can see this if we consider the so-called 'missing explanation argument' (**MEA**).

[138] We would, of course, have to build in other conditions concerning factors like the agent attending to the task, being suitably informed of the non-moral facts, and so on.

[139] Being sanctioned by the categorical imperative means that an action is permissible. If an action fails the test it is forbidden. Actions will be obligatory when their omission is forbidden. The content and Korsgaard's derivation of the categorical imperative will be returned to later.

We can see this best if we consider the following consequence of the judgement-dependence framework:[140]

> **RED**: It is (non-trivially) a priori necessary that: if x is red then x is disposed to appear deep red to standard subjects under standard conditions. (Miller, 2001, 81)[141]

What the framework gives us is a relationship of *semantic dependence* between truth about a particular subject matter and our best judgements about it. That is, when read left to right the bi-conditional in the **BE** and **PE** tells us that the truth of colour judgements is conceptually tied to our best judgements. However, there seems to be a problem with this. Mark Johnston argues (1989, 1993a, 1993b, 1998)[142] that on the Wright framework a true empirical explanation goes 'missing'. In effect, Johnston points out that in addition to wanting to claim a relationship of semantic dependence between colour judgments and the truth of those judgements, we also want to be able to say that people judge things to be red *because* they are red. The because in this statement looks to be explanatory. However, on the Wright framework this empirical explanation goes missing. One way to see what's going on here is to consider the following three claims (this reconstruction of the argument is taken, with some modification, from Miller's 2001, 80-1):

(i) x is red

(ii) When an object has some colour then standard subjects under standard conditions are disposed to see it as having that colour; i.e. they are disposed to have its colour appear to them.

[140] Here I have put the claim in terms of objects *appearing* red to subjects under standard conditions. These are the terms the argument is put in by Johnston and Miller. However, the difference does not matter as, presumably, subjects judge things to be red on the basis of their appearing red.

[141] This is derived from reading the relevant **BE** right to left.

[142] See Pettit (1991), (1996), Menzies and Pettit (1993), Wright (1989), (1992), Miller (1995), (1997), (2001), Blackburn (1993), McFarland (1999) and Haukioja (2006) for responses.

(iii) Standard subjects under standard conditions are disposed to see x as red, i.e. they are disposed to have its redness appear to them.

The problem is this – with **RED** we can derive (iii) from (i), as an *a priori* and necessary matter, without using (ii). (ii) is entirely redundant. However, (ii) looks like an empirical generalisation that does have some explanatory relevance to the truth of (iii). Thus Wright's account of judgement-dependence makes a perfectly good empirical explanation go missing.

The best way to respond to this argument is to follow Miller in arguing that the contingent generalisation from which Johnston derives (ii) above is too strong. Johnston derives (ii) from:

**A.** : Standard subjects (with respect to a family of qualities had by a range of objects) have a disposition which in standard conditions issues in the appearing of an object having some of the qualities just when the object has these qualities. (Johnston, 1998, 17).

The motivation for holding **A** is that it is supposed to explicate the sense in which we think of standard subjects as responding to, or perceiving the qualities of coloured objects. However, Miller argues that we can makes sense of subjects responding to the colours of objects in the right way by merely embracing:

**A*:** Standard subjects (with respect to a family of qualities had by a range of objects) are such that: if conditions are standard, and they view an object, then

> the object will appear to have one of those qualities just when it has that quality. (Miller, 2001, 82).

By plugging in **A*** instead of **A,** instead of getting (i)-(iii) above we get:

> (i*) x is red
>
> (ii*) If S is a standard subject, conditions are standard, and S views an object, then the object will appear to have a particular colour when it has that colour.
>
> (iii*) x appears red to S (82)

In contrast to (iii), we cannot derive as an *a priori* and necessary matter (iii*) from (i*) and **RED** alone. Thus there is space for an empirical explanation of the following form:

> **EX1**: If S is a standard subject and S views x in standard conditions then (x appears red to S *because* x is red).

The explanation that does go missing is:

> **EX2**: (If S is a standard subject who views x in standard conditions then x appears red to S) *because* x is red. (Miller, 2001, 83)

But this should be of little concern – **EX2** explains why a conditional linking objects appearing red to viewings of an object in the right conditions is true in virtue of that object being red. In contrast **EX1** explains why the object appears to be red in virtue of the object's being red. It is **EX1** we want, and it is this empirical explanation that does not go missing.

What this material suggests is that on the judgement-dependence model claiming a relationship of semantic dependence between our best judgements about a quality and the truth about that quality's extension is compatible with giving an empirical explanation of why we make those particular judgements. Another way to think about it is this. Suppose that we found that all red objects have a particular primary quality in common – the same surface-reflectance properties, say. Then, on the judgement-dependence model we could, as a contingent matter of fact, claim that objects look red because they possess that particular property. What is important is what we would say in the case where we find that there is no single primary quality that all red objects possess – red objects are wildly heterogeneous, from the standpoint of primary qualities. In that case we would still be able to hold on to the claim of semantic dependence made by the judgement-dependence theorist, we would merely give up hope of being able to offer an empirical explanation of things looking red in terms of their primary qualities.[143]

Similar considerations apply in the debate between Korsgaard and the moral realist. Suppose that all right actions share in common some natural property (that they promote the greatest balance of pleasure over pain, say). Then we can, on this framework, offer the following true empirical explanation 'Actions are right because they promote the greatest

[143] Incidentally, with the compatibility of the judgement-dependence theorist's semantic claim and the right kind of empirical explanation demonstrated we can see why D J Bradley's recent (2011) recasting of judgement-dependence in functionalist terms is unmotivated. Bradley argues that we can avoid the explanation going missing if instead of claiming a relationship of dependence between the judgement-dependent quality and judgements about that quality, we claim one between the quality and being in a state that is disposed to lead to best judgements about that quality. This move is unnecessary if there are no missing explanations to be accounted for. Bradley complains that responses to Johnston's MEA like Blackburn (1993); McFarland (1999); Haukioja (2006); Pettit and Menzies (1993); Miller (1995), (1997), (2001); disappointingly "criticize non-essential details of Johnston's exposition" (Bradley, 2011, 299). If the material I exposit above is right, then it is clear that this is one of those, very many, cases where attention to the details reaps rewards.

balance of pleasure over pain'. However, this does not threaten the constructivist claim that the truth of judgements about rightness depends upon our following the correct procedure. We can see this in the case where there is no such natural property uniting right actions – in that case the constructivist can still make their claim of semantic dependence.

What this means is that the type of metaphysical positions Korsgaard tries to argue against are simply irrelevant to her concerns (if this model can be made to work). She can say to the realist – 'I simply do not care whether all right actions share some natural property in common, for whether they do or not is only relevant for our attempts to give an empirical explanation of our judgements about the rightness of actions in terms of those actions being right. Whether or not there is some property that all right actions share, my claim that the truth about rightness depends solely upon us having a correct procedure for evaluating claims about rightness still goes through'. [144]

So, if we cast neo-Kantianism constructivism in this light we can explain why Korsgaard has a distaste for the kind of metaphysical explanation of normativity given by the realist, and why her arguments against it seem to misfire. She should not be arguing that moral realism is false but, instead (as she starts to do in her later work), that it is irrelevant to her

[144] There may appear to be some tension between what I say here and in chapter 1. There I argued that if all right actions shared one or a few natural properties in common that would be good *prima facie evidence* that we can reduce rightness to that property or those properties, and if right actions were wildly heterogeneous in that respect this would block the naturalistic reduction. However, this tension is only apparent: there I freely admitted that the evidence the homogeneity of right actions would provide is only *prima facie*, and that it could be defeated with further argumentation. What this exploration of judgement-dependence tells us, I think, is that one way to show that that evidence is irrelevant is to take on board a judgement-dependence framework and demonstrate that it can be made to work with rightness.

concerns. The cognitivist anti-realism provided by the judgement-dependence framework gives us a clear way of fleshing this out.

However, this framework cannot secure all of Korsgaard's ambitions. She also wants to transcend the cognitivist/non-cognitivist division. We saw in chapter 4 how the attempt to do this looks problematic, and above that when Korsgaard is explicit on this issue that she says things that are hard to get a grip on. What the judgement-dependence framework can't do is provide an easy way for Korsgaard to express this concern. This is because the framework, as usually stated, is cognitivist. It provides a link between the *truth conditions* for particular area and our best judgements about that area. However, the truth conditions provided are anti-realist in flavour – they don't outstrip our best judgements. Perhaps Korsgaard (or other neo-Kantian constructivists) could be happy with this kind of constructivism – where moral judgements end up with truth conditions, but not ones of a substantial, realist bent. To put things very schematically Korsgaard seems to be worried that moral realism distances morality from our own standpoint – it makes being obligated a matter of facts existing out there in the world, and it's hard to see how these facts can get a grip on us if that is the case. On the cognitivist anti-realism underwritten by the judgement-dependence model we end up with moral judgements being belief-like in that they have certain truth conditions, but these truth conditions are given by *our* best judgements about morality in question. Thus there isn't the kind of distance between ourselves and the truth about morality that Korsgaard seems to be worried about. This gives us a sense of how putting moral qualities together with secondary qualities makes their special features (their normativity) look a bit less troubling, and explains the popularity of the type of companions in guilt strategy that could be built off the back of these observations.

Adopting the judgement-dependence framework also gives us a clear way of evaluating the constructivist's proposal. On my account, constructivism becomes a kind of cognitivist anti-realism with a suitable procedure built into the C-conditions for best moral judgement. If this is to work then the **PE**s we end up with (e.g. $\mathbf{PE}_{\mathbf{Permissible}}$ above) need to obey the four conditions outlined. It's truth must be knowable *a priori*, the C-Conditions must be substantive and respect the independence condition, and construing moral qualities as judgement-dependent must be the best explanation of the truth of the equation. We won't need to worry about the extremal condition – morality does not seem to be a place where we should be worried about the *a priori* truth of the equation being an artefact of our infallible access to the mind-independent moral facts. Wright himself, though, argues that moral qualities are not suitable candidates for a cognitivist anti-realist treatment because the C-conditions for best moral judgements violate the suitability and independence conditions. This is because when we state the conditions under which moral judgements are maximally credible, we cannot eliminate reference to S being a 'morally suitable subject'. This then creates a problem, which Callum Hood puts succinctly enough to be worth quoting at length:

> The moral suitability condition cannot amount to having whatever-it-takes to form correct moral judgements on pain of violating the substantiality condition. The alternative is that the satisfaction of the moral suitability condition depends upon an anterior determination of the extension of 'morally suitable'. Now an analogy can naturally be drawn between shape discourse and ethical discourse. Just as satisfaction of the stability condition in the case of shape was not logically independent of the extension of 'square', likewise it seems that the satisfaction of the moral suitability condition is not logically independent of the extension of 'morally suitable'.

(Hood, unpublished, 2010)

Wright himself puts the main point this way: "S's moral suitability, in particular, is itself, presumably, a matter for moral judgement" (Wright, 1988, 23). Thus a judgement-dependent treatment of moral qualities fails to meet the four conditions required, and thus moral qualities can't be judgement-dependent.

However, this is where the positive proposal I have made starts to have some bite. There seems to be nothing, at least in principle, blocking us from giving a characterisation of the neo-Kantian's preferred procedure for settling moral questions that does not violate the independence condition. It is, at least, a matter worth investigating. In addition, if we could characterise the C-conditions as following a certain procedure (where this procedure is applying the categorical imperative) we are unlikely to run into problems with the substantiality condition. What my proposal does do is shift focus on to the *aprioricity* condition. We have to attend closely to whether the truth of the connection between moral truth and following the categorical imperative is knowable *a priori* (and whether they *are* so connected, of course).

Fortunately, we can find in Korsgaard an engagement with just this sort of question, in her attempt to give a derivation of the categorical imperative and an explanation of its normative force over us. Constructivism, I have urged, is best thought of as a variety of Wright style cognitivist anti-realism where the relevant conditions for best moral judgement are construed as following a certain procedure. Whether Korsgaard's position works, and

provides us with a genuine metaethical alternative, depends on whether her derivation of the categorical imperative works. This is the issue to which I now turn.

## 5.2 The Derivation of the Categorical Imperative

In *The Sources of Normativity* Korsgaard attempts to show how Kant's categorical imperative is binding on us. Kant offers a number of formulations of the categorical imperative, but only two will be relevant here: that we should act only on maxims that we could, without contradiction, will to be universal laws (the Formula of Universal Law – **FU**); and that we should act only on maxims that we could will to be universal laws as an equal legislator in the Kingdom of Ends (**FKE**) (Korsgaard 1996, 98-9). The interrelations between these two formulations and the others that Kant gives, along with the best way to explicate their content, has been a vexed issue in the study of Kant's thought. I will not get into those debates here. Instead I will just explain how Korsgaard understands these formulations a little before going on to how she attempts to derive them.

In Korsgaard's terminology, she labels the **FU** the categorical imperative[145]. What the **FU** tells us, she argues, is that any maxim that we act on must have the correct form to be able to be willed as a universal law. In other words: "nothing determines what the law must be.

[145] In fact, in her later work she uses 'categorical imperative' to stand for the principle that guides our action which, in contrast to the **FU**, *does* have moral content. Throughout this thesis I have used 'categorical imperative' to stand for this principle, what Korsgaard calls in her (1996) 'the moral law'. Outside of this paragraph I will revert to using 'categorical imperative' to mean the principle with moral content, and **FU** to stand for the weaker constraint.

*All it has to be is a law*." (98, emphasis in the original). In addition, whether a maxim is suitable to be a law depends upon its *form*. Its form must be such that it can be willed to be a universal law. Now, there are some notions here that need unpacking. First, a maxim. A maxim, for Korsgaard, is the principle underlying my 'action', where an action is my act together with the purpose for which that act is performed. So, if I lie to you to get you to lend me money then the maxim of my action (which is a combination of the *act* of lying together with the *purpose* of getting money) is something like 'lie in order to get money'. What *form* does it have to have? To be suitable as the kind of thing that we can will the maxim must be universal, in that it is applicable as a guide to behaviour in all similar circumstances. Why is this? Well, Korsgaard thinks that if you adopt a principle (maxim) to guide your behaviour and you then discard it for no reason, then you will have obliterated the distinction between the person willing an act and the incentive for which that act is performed.[146](2010, ch. 4). So, when deciding whether to perform a particular action, if the **FU** is binding on us, we have to check whether the maxim that that action embodies has the right form to be universalisable in the right sort of way.

As this stands though, it does not look like the **FU** will be able to give much content to a moral theory given how weak *its* content is. It just tells us that the principles we act on have to have a suitable form to be treated as universal, in the sense that I would see them as giving reasons to act in similar ways in similar circumstances. But this test permits all kinds of immoral maxims – I can, presumably, will an action that embodies the maxim 'murder people in order to steal their possessions when you want them'. The **FU** just tells us that this maxim is only appropriate if we would treat it as providing us with reasons to act in a

[146] At the moment this sounds rather too quick, but once I get onto explaining the particular role Korsgaard posits for the constitution of agency in her theory the motivation for some of these claims should look a little clearer.

similar way in other similar circumstances. So, every time someone has something I want I murder them to get it. And this is sanctioned by the **FU**. Korsgaard acknowledges this deficiency in the **FU**, but argues that the Kantian has the resources to work towards a stronger constraint on our actions. (1996, 98-99).

In order to get any moral content we need what Korsgaard calls the 'moral law'. This is the Formula of the Kingdom of Ends (**FKE**). It tells us to only act on maxims that we could will to be universal laws as an equal legislator in the Kingdom of Ends. There is quite a lot packed into this principle, not all of which I can explicate. For the purposes of trying to get a grip on Korsgaard's derivation of this principle we can give it a quite simple gloss. What the **FKE** is getting at is that when we test our maxims we have to take other members of the Kingdom of Ends into account. What is the Kingdom of Ends? For Kant this is quite a complicated matter, but for Korsgaard it simply includes all rational agents. To be a member of the Kingdom of Ends is to be such that you must be treated as an end in yourself, rather than as a *mere* means.[147] What about the equal legislator business? What Korsgaard is getting at is that the principles that we choose to express in our behaviour must be ones we would rationally agree on in a Kingdom of Ends where everyone is taken into consideration.[148] We can see, roughly, how this formulation is more stringent than the **FU** – it's unlikely to permit the maxim of killing others to take their possessions, for example.

[147] Again, a lot of ink has been spilled over what it means to treat someone as a mere means. To get into this issue would take us far away from the task I've set myself. For the purposes of this thesis we can just try to run with the intuitive notion we get from the phrase as it stands.

[148] Here we can see affinities with Rawls's neo-Kantian constructivism in the case of justice. For Rawls we are looking for the principles we would choose to co-ordinate over from behind a 'veil of ignorance' where facts about who you are are unavailable to you. The veil of ignorance is, of course, a mere literary device. What is doing the work is the ideal that justice should be impartial, so that a system is only fair if we do not favour it because of particular facts about what kind of person we are. We can, to get our

What I have given here is in many respects inadequate – I haven't been able to fully explain what a maxim is, in what sense the **FU** tests its 'formal' properties, what it is for an action to embody a maxim, and the details of how we would get full blown moral content out of the **FKE**. In some cases these lacunae are shared by Korsgaard's own work, but aside from that all I've tried to do is give a rough and ready characterisation of the ideas involved in order to be able to evaluate Korsgaard's arguments for her claims.

Korsgaard's main argument (from *The Sources of Normativity*) for the categorical imperative is something like the following:[149]

(1) As human beings, we are faced with the necessity of acting, of making choices.

(2) Being self-reflective, we make these choices on the basis of reasons.

(3) Therefore we must have some reasons available to us to make choices.

(4) In order to have reasons, you must have some conception of yourself under which you take your life to be worth living (some practical identity).

(5) For the reasons that flow from this identity to be binding upon you, you must take that identity to be valuable.

---

heads round what Korsgaard is doing, think of Korsgaard's proposal as trying to explore that intuition in the case of morality more generally.

[149] This formulation of the argument is borrowed, in part, from William FitzPatrick (2005, 662-3).

(6) But your reason to have some practical identity is not a reason that flows from any particular practical identity.

(7) It is a reason you have in virtue of your human nature. In particular, in virtue of your capacity for rational agency.

(8) To see *this* reason (that you need some practical identity, because of your rational nature) as binding, you need to see your identity as a rational agent as valuable.

(9) Therefore, because of the necessity of acting, it is also necessary that you see your rational agency (your 'humanity') as valuable.

(10) It is not possible to value your own humanity without valuing humanity in general.

(11) Therefore, because of the necessity of acting, it is also necessary that you value humanity in general.

(12) And this just is Kant's categorical imperative.

What support is available for these claims? **1** is the claim that for beings like us agency is inescapable (I will return to this issue below). **2** is a claim about the nature of agency – that when we are faced with making choices we, if we are to be considered as agents at all, must make the choice on the basis of reasons, rather than acting randomly. I will not challenge this claim here, but see Jonathan Way's (2010, §5.1) for a way you could launch such a

challenge. We can construe **3** as a claim following from a transcendental argument – agency is possible (some of us are, at least some of the time, properly considered agents), and we know from **2** that for agency to be possible there must be reasons available to us, therefore there must be reasons available for us to make our choices.

**4** is based on Korsgaard's conception of what it is to be an agent - that the point of action is to constitute yourself as an agent with a particular identity. From this she thinks it follows that in order to make choices (given that the point of making choices is to build and reinforce an integrated identity for yourself) you need to have *some* practical identity which you aim to preserve. A practical identity is, for Korsgaard, 'a conception under which your life is worth living'. Her point seems to be that in order to think of the reasons you act upon to have some force, they must be related to something that you think is valuable (**5**). Let's take a concrete example – one practical identity is a teacher. As a teacher you are faced with making choices about how to conduct your behaviour. One way to settle these questions is to reflect on what it makes sense to do, given your practical identity as a teacher. Thus, if you treat your identity as a teacher as valuable, you will see that you have reason to stay in grading work, rather than going out on the lash. Staying in, then, looks like a choice-worthy action because it is bound up in the practical identity of being a teacher.[150]

However, that you need some practical identity in particular is not a reason that stems from any particular practical identity (**6**). It's not because you are a teacher that you need to be a teacher. Instead you need to adopt some identity or other because you need some conception of your life as worth living to make certain actions (because they are suited to the identity in question) present themselves as choiceworthy. This reason you have for

[150] What I've said here, particularly about self-constitution, is apt to sound fairly vague and perhaps paradoxical. We will return to this issue below.

valuing your practical identity is one that flows from the fact that you must have some identity or other, because you are an agent (**7**). In addition, practical identities conflict – your duties as a teacher might clash with your duties as a drinking partner. When this happens, you will need some basis for choosing among these competing practical identities, and that's where the demands inherent in merely being a rational agent might be able to help.[151]

**8** relies upon the idea that for rational agency to provide reasons for you (in this case, a reason to have *some* practical identity or other) you need to value your own rational agency. Thus we reach **9** where we are told that all rational agents must value their own rational agency (their 'humanity'). At this point Korsgaard will argue that this commits every rational agent to the **FU**. The reasoning here is not entirely clear, but the idea seems to be that rational agency requires that you cannot treat something as a reason for acting without treating it as a reason for so acting in other, similar circumstances (and this gets us the **FU**). To treat a consideration as a reason for action in a particular case only is to think that there is nothing about the consideration itself that *makes* it a reason for action, and so it's hard to see, Korsgaard argues, how we can then treat it consistently as a reason even in that particular case.[152] Treating considerations as reasons requires considering them to be universally binding.

Claim **10** is supposed to be a consequence of what Korsgaard calls 'the essential publicity' of reasons. She argues that Wittgenstein's private language argument shows that reasons

[151] *This* claim seems questionable – how would the requirements of agency help to solve this case, for example? However, we do not need to be too concerned about this as it's not essential to Korsgaard's argument.

[152] See Raymond's Geuss's (1996) for a convincing case, building on Schlegel's criticism of Kant, that this move of Korsgaard's severely underestimates the human capacity of freedom.

cannot be 'private' in the sense that they apply only to us. It then follows, she claims, that if I have to see my humanity as valuable I have to see everyone else's humanity as valuable. This yields **11** which tells us that we have to value everyone else's humanity. And, Korsgaard thinks, this is just to treat people as ends in themselves and take them into consideration when choosing principles to act upon, which is just (according to Korsgaard) the categorical imperative in its **FKE** formulation.

There are a number of problems with this argument. The first is the move I've outlined in the last paragraph. Korsgaard's reasoning for the essential publicity of reasons is extremely hard to follow (see 1996 Ch. 4). Trying to get clear on what she is doing would take a whole thesis. All I can do here is simply assert that Korsgaard seems to confuse the question of whether a reason is public with the question of whether a reason is self-directed. She may be able to show that reasons have to be public, but not that their content cannot be self-directed. If so, then it's perfectly consistent for me to value my own humanity (and thus be forced to respect the, quite weak, **FU**) without valuing anyone else's, and thus the step to the full-blown categorical imperative fails.

Another difficulty is that Korsgaard, in a number of places, relies on the claim that in order for you to have reasons you must think that these are reasons because acting in accord with them secures something that you take to be valuable. This seems to commit one to quite a strong anti-realism about reasons.[153] A realist about reasons, who claims that whether some consideration is a reason for action *does not* depend upon our taking the thing procured by that action to be valuable, but instead on whether that thing really is valuable, could block

[153] If I thought that certain ends were intrinsically valuable, independent of me, then I could think that that an act would further those ends is a good reason to do it, whether or not I regard myself as valuable – this criticism comes from FitzPatrick (2005), and Hussain and Shah (2005) repeat the claim.

the moves up to **8**. They would claim, in effect, that whether you have a reason to act in a particular way does not depend on your seeing your practical identity as a source of value, nor on you seeing your rational agency as valuable. You just have reasons, and being a rational agent involves responding to them in appropriate ways.

However, running the criticism this way seems to be off beam. If reasons claims are normative claims themselves, then we should *expect* Korsgaard to be anti-realist about them. Of course a realist would not accept this position, but if they just baldly assert a realism about reasonhood or value against Korsgaard this would simply beg the question against the constructivist framework. What it does mean is that Korsgaard cannot rely upon a realism about any of the normative claims in her derivation of the categorical imperative, but if cognitivist anti-realism is viable then this should not be a worry.

Instead what we should be worried about is whether, in doing things this way, Korsgaard violates the independence condition on the relevant provisional equations. Taking this strategy, Korsgaard does not build conditions that violate the independence condition directly into the C-conditions for best moral judgement, however she does make claims that violate the independence-condition in the course of explaining the *a priori* truth of the relevant equations. Korsgaard could then argue that she is not violating the independence condition at all – the truth of the link between the truth values of claims about the extension of 'permissible' (for example) and our best judgements of permissibility (when we apply the categorical imperative) can be stated without assuming any facts about the extension of 'permissible'. It's just that getting you to see that this link exists *does* require assuming facts about the extensions of moral concepts (that not considering yourself as valuable in virtue of your rational agency is impermissible, say).

Suppose though, that we think that this is illegitimate, and that Korsgaard does violate the independence condition. At this point the neo-Kantian constructivist can try to argue that violating this condition is not necessarily lethal to their programme. As we saw above the provisional equations are not meant to give an analysis of the truth conditions in question, so there is no direct charge of circularity available. Instead, violating the independence condition means you don't have an easy way of demonstrating the success of your account. The constructivist might be happy with this, but it would certainly be a weakening of their dialectical position – it would make it hard for them to convince anyone else of the truth of their position. What I think this means is that the criticisms from FitzPatrick (2005) and Hussain and Shah (2006) do have some bite, but not where they expect them to. Instead of the problem being that Korsgaard relies on anti-realism about value (which, if the constructivist project can get off the ground, is not a problem) the problem is that Korsgaard, in her derivation of the categorical imperative, makes normative claims and that this threatens to violate the independence condition on the provisional equations I am trying to reconstruct a constructivist position from.

Another concern that we might have is that Korsgaard tries to build up to the categorical imperative by making various claims about what is necessitated by rational agency – what you have to do to be a rational agent. What is the status of these claims? In addition we might worry that the above argument does not secure us the right result – we learn that we are bound by the categorical imperative if we are to be a rational agent, but couldn't we dodge the requirements of the categorical imperative by opting out of being a rational agent? In order to answer these questions Korsgaard has recently developed a type of *constitutivism*, the last piece of Korsgaard's view I have left to explain.

## 5.3 Constitutivism

Constitutivism tries to ground the claims about agency Korsgaard uses in her argument for the categorical imperative by arguing that they are *constitutive standards* for being an agent at all. The idea is that the 'teleological organisation' of something supports normative judgements about it. Korsgaard deploys the example of building a house. A house has a certain function – it is *for* providing stable shelter. In order for a house to provide shelter it has to meet certain standards – the walls must be solid, the roof must be above rather than under the walls etc. These standards provide guidance for the activity of house-building. If you are not at least trying to put the roof on top of the walls, and build walls that stand up, it's not the case that you are just building a house badly; you are not building a house at all. From this we get the idea of a constitutive principle of an activity. You cannot build a house without building walls that support a roof, in the same way that you are not walking unless you are putting one foot in front of another. If, unless you are performing an activity in line with these constitutive principles you are not performing that activity at all, how is it possible to perform an activity *badly*? Korsgaard argues that you must at least be *guided* by the constitutive principles in question – they must be what you take to be directing your activity. At a certain point, however, if you fall away from the constitutive principles in question badly enough we will say that you are no longer performing the activity at all. A shoddy builder is one who builds a house that doesn't stand up for very long. A child who throws a load of bricks together in such a way that they are not even trying to create a structure that stands up is not house-building. (This is a summary of Korsgaard's (2009), ch. 2).

Constitutive principles, Korsgaard claims, are able to meet sceptical challenges quite easily. Suppose you are building a house, and someone asks you 'well, why are you putting up

walls strong enough to support that roof, why don't you just build the walls out of twigs?', then you have a reasonable response to them: because if I did that then I wouldn't be building a house at all. The idea is that if you have reason to be engaging in an activity, then you must be guided by the constitutive principles that constitute that activity or you won't be doing that activity at all. This view has traces in her earlier (2003) where she argues that what constructivism tells us is the way to solve a problem when we acknowledge that it is a problem we share. For example, the problem of distributive justice is one of how we distribute goods fairly. What Rawls's principles tell us, Korsgaard claims, is which principles a system of distribution must embody to be a system of distribution *at all*.

How does this apply to action? Well, Korsgaard claims that the function of actions is to constitute ourselves as agents. It is by acting on the basis of reasons that we make ourselves into rational agents with the right level of 'psychic unity' to have our own personal identity. The principles that are constitutive of the activity of self-constitution are, Korsgaard contends, the hypothetical and categorical imperatives. So, in order to be agents at all we must be guided by the hypothetical and categorical imperatives. By taking up constructivism with this constitutivist element we have an answer to the normative sceptic we encountered in chapter 1. The normative sceptic asks why they should be moral. The constitutivist constructivist tells them that if they are to be a rational agent at all they must be guided by the categorical imperative.

There is a lot to this claim, resting, as it seems to, on the history of existentialist notions of character. I shall not go into it in detail, however it is necessary to spend some time trying to dispel the notion that Korsgaard's idea of self-constitution is paradoxical (she deals with this 'paradox of self-constitution' in her (2006) ch. 2). The puzzle is this: Korsgaard wants

to claim both that it is as a possessor of an identity that you are the author of your actions, and that by acting you create your own identity. But how can you create your identity through your actions, and choose your actions on the basis of your identity? If *you* are already there to choose your actions, why would you need to make yourself? And if you make yourself through your actions, how can the actions depend on the identity they are creating?

Korsgaard hopes to dissolve this paradoxical appearance through a comparison with living things in general. To be a living thing is to engage in the activity of constantly making yourself into yourself. For example, being a giraffe and being a good giraffe are the same thing – it is constituting yourself as a giraffe (by digesting nutrients, repairing your body) that makes you into a giraffe at all. But this is not paradoxical (so Korsgaard claims). It only looks paradoxical if you look at one particular time-slice of the giraffe's life, and ask 'is it constituting itself now?'. From the point of view of a particular time this question makes little sense – if the giraffe exists, it is a giraffe, and doesn't need to do anything more to be a giraffe. But this point of view is misleading. All it shows is that being a giraffe, or any living thing, is not an event or state, instead it is an activity. To be a living thing is to engage in the activity of being that living thing. As it is with living things, so it is with self-constitution. To be a person is just to be engaging in the activity of making yourself one. This only sounds paradoxical because we automatically envisage the situation from the time-slice perspective. This time-slice view forces us into a dilemma – either we are already made, in which case we do not need to do anything else (and have a full and determinate set of practical identities) or we have not yet been made, in which case making ourselves into ourselves would be of no help (we have no practical identities from which to start the process). But, again, this ignores the point that self-constitution is an activity, just as being a

living thing is. Choosing between the picture of ourselves as either full or empty selves leads to all sorts of problems. It seems as if we must be full, otherwise we will have nothing upon which to base our choices. However, if we are full, then we will not really be free – there will be something forcing our hand: our pre-existing determinate self. Instead we should see a person as a thing that is engaged in the activity of constituting themselves in action, rather than taking a particular stage of that person's life and asking how *it* got made a person – looking at things from this perspective *does* lead to the appearance of paradox, but it is the perspective that is at fault, not Korsgaard's view, she argues.

There are other questions we might have with Korsgaard's view: why are the hypothetical and categorical imperatives the constitutive principles of the activity of self-constitution?[154] Is her appeal to teleology metaphysically kosher? However, the main problem with Korsgaard's constitutivism, I think, is that it does not deliver the result she wants from it. We can see this if we consider one claim that Korsgaard needs to get the whole framework going. In order for constitutivism to provide a response to sceptical challenges we must have a reason for engaging in the activity in question. What Korsgaard needs to claim, then, is that for beings like us rational agency is inescapable. The reason why the categorical

[154] For Korsgaard's answer to this, see (2009, ch. 3). It's fairly easy to see how we might ground the hypothetical imperative in this way – it does look like there is something defective with an action which does not connect the means employed to secure an end and the end in the right kind of way: such a bodily movement might not count as an action at all. What Korsgaard needs, however, is the claim that actions which fail to embody the categorical imperative are defective *qua* actions. Korsgaard thinks her argument against particularistic willing above (which we will return to when we come to Geuss's criticism of her view) is enough at least to secure the FU. However, the move from that to the full-blown categorical imperative is, again, obscure. However, this does not matter much for our purposes. The criticism I shall make against Korsgaard will still hold even if she could give a principled argument for the claim that being guided by the categorical imperative is constitutive of action. The other way of attacking constitutivism (claiming that the constraints the constitutive principles of action place on agents are too weak to derive the requirements of morality) is taken up by Kieran Setiya (2003), (2007).

imperative has force over us is because it is one of the constitutive principles governing rational agency, and you can't stop yourself being a rational agent.

## 5.4 The Inescapability of Agency

David Enoch objects to constitutivism by challenging the inescapability of agency:

> Classify my bodily movements and indeed me as you like. Perhaps I cannot be classified as an agent without aiming to constitute myself. But why should I be an agent? Perhaps I can't act without aiming at self-constitution, but why should I act? If your reasoning works, this just shows that I don't care about agency and action. I am perfectly happy being a shmagent – a nonagent who is very similar to agents but who lacks the aim (constitutive of agency but not shmagency) of self-constitution. I am perfectly happy performing shmactions – nonaction events that are very similar to action but that lack the aim (constitutive of actions but not shmactions) of self-constitution. (Enoch, 2006, 179)

Turning to Korsgaard's example of house-building, Enoch asks us to imagine a builder who, when we point out to him that he's falling short of the constitutive standards governing house building by doing such a shoddy job, replies by saying something like 'Fine, you've proved to me that I don't care about house-building. But this just means I'm really a shmouse builder' and carries on regardless. Enoch is making the point that Korsgaard's argument only establishes the conditional: If you care about being an agent, then you should follow the categorical imperative. But this opens up space for the

normative sceptic to raise their scepticism in a different form. They can agree that the conditional is true, and then go on to ask why they should care about being agents. The position we are left in is this – if shmagency is a genuine possibility for someone to take up, then the constitutivist attempt to ground the categorical imperative fails.

Enoch then outlines Connie Rosati's response to this problem. In brief, she argues that the standards constitutive of agency are in a way self-vindicating. This is because the challenge raised by the sceptic above *depends upon* the exercise of these capacities. In the same way that we can dismiss scepticism about logic due to the fact that the sceptic will have to use *some* logic to get their argument going, we can also dismiss the sceptic's concerns about the status of rational agency. Enoch, however, contends (following Wright, 1991) that this sort of response gets the dialectical status of sceptical challenges all wrong. When the sceptic uses the very tools they are undermining, we should treat their challenge as a sort of *ad hominem* argument. They are arguing that, even with the tools you have at your disposal (the laws of logic, or facts about rational agency), *your* position collapses into a form of scepticism. They are allowed to use our claims against us to launch their challenge. Thus Rosati's response won't do (Enoch, 2006, 182).

Enoch raises another problem that looms even if the constitutivist *can* show that agency is inescapable and shmagency impossible. Enoch asks why we should think of the necessity of agency as being a kind of *normative* necessity. Ok, the sceptic can say, I see that I have to be an agent, but why is this not simply a causal necessity? Enoch drives the point home against David Velleman's and Rosati's constitutivism by pointing out that they argue against views that try to ground normativity in terms of fulfilment of our desires. Velleman

and Rosati reject this sort of view because basing normativity on the desires that we happen to find ourselves forced to have seems unacceptably arbitrary. But how is grounding normativity in the fact that we find ourselves forced to be agents in any better shape? (Enoch 2006, 179). I think that this concern is particularly pressing for Korsgaard. Recall (from ch. 1) her argument against an evolutionary-based theory of morality. Such a theory of morality would be inadequate because even if we showed that as a matter of our evolutionary heritage we felt ourselves forced to act in accord with the dictates of morality by inner drives that had been selected for, this does not yet answer the normative question, because we can still ask whether we *should* allow ourselves to be carried away with those drives, even if, as a matter of psychological fact, we cannot avoid it. If Enoch is right, we can now run the same objection against Korsgaard's own theory – suppose, we can say, as a matter of fact I have to be an agent, and thus have to perform actions which are governed by the categorical imperative. That still does not show that the categorical imperative really is normative, even if I can't avoid being guided by it.

What the constitutivist needs to do to respond to this is tell us more about the type of necessity involved in their claim that we are forced to be agents. If it is mere causal necessity then we can run Korsgaard's own arguments against her. But it's hard to see what form of necessity would do the job. Enoch argues that if the necessity in question is normative Korsgaard is left with one unexplained normative claim, which can only be given a realist treatment, in which case constitutivist constructivism fails as both constitutivism and constructivism and collapses into a complicated form of normative realism. This neglects one possibility open to the constitutivist – to give this normative claim, again, a cognitivist anti-realist treatment. However, the constitutivist has given us no clue how we

might do this, and even if we can do it concerns about respecting the independence condition are likely to loom even larger.

Enoch's first problem, however, can be solved if we follow the line offered by Luca Ferrero (2008). Ferrero basically argues that even if shmagency is possible, *choosing* shmagency is impossible. The idea is this: Enoch underestimates the strength of the constitutivist's position by putting his points in a linguistic way. It's not that we can't properly be classified as agents if don't follow the constitutive standards governing action, instead we won't *be* agents if we don't follow those standards. What ignoring the constitutive standards of agency involves is trying to choose to be nonagents, but such a choice is still a choice and hence governed by the relevant principles. We cannot choose to be unbound by them.

This position, for one thing, has the counterintuitive consequence that no-one ever chooses to commit suicide. That aside, the argument invites an obvious response from the anti-constitutivist: 'ok, you've shown me I can't choose my way out of agency. But that just means I will have to shmoose my way out of it, where shmoosing is a lot like choosing (in that it reliably brings about changes in future behaviour on the basis of the consideration of reasons) except that shmoosing is not bound by the principles constitutitive of choosing.'. The constitutivist's response is equally obvious – if you are weighing up considerations as reasons you are choosing, not shmoosing. And then the anti-constitutivist says 'ok, I will weigh up shmondsiderations as sheasons then'.

I think that we can break into this deadlock by considering what the constitutivist has to say about people who violate the categorical imperative on at least some occasions – people

who just ignore its demands from time to time. Such people seem to be possible. In fact, all of us are such people, to some extent. But few of us are, presumably, nonagents because of it. What Korsgaard seems to think is that our practical identities are robust enough to allow some violations of principles which stem from them. This claim seems to be required just by the bald facts of how humanity acts. However, if we continually fall short of the constraints placed on us by our practical identities then we will no longer embody the right level of 'psychic unity' and our identities will dissolve. This might be the right way of thinking about things, but it leads to problems for Korsgaard.

Rather than thinking about a general normative sceptic, think about a sceptic who asks why they should perform a particular action. Korsgaard's reply seems to be 'well, if you do too many things like that then your rational agency will dissolve away and you'll no longer be an agent'. This invites the reply 'ok, well in that case I won't do it too many times, but you've provided me no reason to not do it in this case'. The problem is not just that Korsgaard, in the face of the obvious facts, has to admit that sometimes we do things in violation of the categorical imperative without losing our agency. The problem is deeper – that in this particular case violating the categorical imperative is not even wrong: this is because violating the categorical imperative is forbidden because doing so threatens your rational agency. But the constitutivist has to admit that one violation of the categorical imperative doesn't stop you being an agent. So any particular violation of the categorical imperative doesn't dissolve your rational agency. And if conforming to the categorical imperative is only required because doing so preserves your rational agency, then you have no reason to follow it in this case.

The problem is even more pressing when we consider Korsgaard's ambitions. Korsgaard acknowledges that the burdens morality places on us can be quite severe – there are cases where morality requires that we give up our lives, for example. To explain this Korsgaard says (1996) we must explain why the consequences of violating the demands of morality can be as bad or worse than what morality demands from us, and on her picture we can do this – sometimes morality can demand we give up our lives, and this can be sensible because the alternative is dissolution of our practical identity, which amounts to a kind of death. However, I think the constitutivist has to acknowledge that the 'psychic' impacts of violating the demands of morality can often be less bad than what morality demands of us (is it really always the case that ignoring a requirement to give up your life leads to a complete dissolution of identity, for example? Looking at humanity it's hard to see how it could be[155]).

There are two possible responses here – Korsgaard could argue that her considerations against particularist willing above rule this out – the maxims we will have to be universal. Even if this is true, though, it doesn't yet give us the right result. We can imagine a maxim that, though universal (in that it prescribes the same action in the same circumstances) prescribes an *immoral* action. As long as the relevant circumstances for that action come up

[155] Evidence for this claim might be found in looking at fictional portrayals of, and real-life stories about, criminals (who violate the demands of morality). Consider Marlo Stanfield from *The Wire* – a rum character who commits all manner of morally wrong acts. But Marlo seems to exhibit the strongest sense of personal identity of any character in the show, one that intimately involves morally wrong actions – when faced with the possibility of living the life of a reputable business man he finds himself compelled, by the very nature of who he is, to engage in criminal behaviour which has the potential to threaten this new lifestyle. Now it could be that Marlo is halfway towards complete dissolution as a person. But I am sceptical enough about human nature to not be convinced of that. This portrayal seems to me convincing, and if it is then this is evidence against the claim that violating the categorical imperative, as a matter of psychological fact, leads to dissolution of identity. We could give numerous other, similar, cases.

rarely enough, then we will not have to be worried about acting on that maxim leading to the end of our personality.

The second response is to say that when people fail to be guided by the categorical imperative they are not really acting. So, again, it's not possible to choose to act in a way not guided by the categorical imperative. However, the people in question are still *doing* something wrong (even if that behaviour doesn't amount to a full blown action). If we follow this response then it makes it hard to see why we would blame people for the bad things they do. After all, on this line they aren't really freely choosing their action, because the things that they are doing are not really actions. What I'm getting at is this: if we call these bodily movements that are not performed in accordance with the categorical imperative nonactions, that does not stop us wanting to say that they are wrong. We want to say that each particular bodily movement should have been an action, and hence performed in line with the categorical imperative. But it looks as if this avenue is no longer open to the constitutivist.

What I think this shows is that we do not need to be concerned so much with the inescapability of agency. Instead we should ask the constitutivist what they say about people who fail to act in accord with the categorical imperative. If their constitutivism is strong enough that these bodily movements no longer count as actions then it's hard to make sense of why we think they are wrong – they aren't actions at all, so aren't really bound by the categorical imperative. If, however, we permit that people can act in violation of the categorical imperative without losing their identity, as long as they don't do it too much, then the constitutivist will not be able to get a moral theory with the strength they

want – they won't be able to answer the normative sceptic who asks of any particular obligation why they should fulfil it.

There is another problem for the constitutivist constructivist to consider.

## 5.5 The Standard Objection to Kant

It is in response to what Korsgaard calls the 'standard objection to Kant' (**SOK**) that she flags up the details of her way of thinking about a maxim; we touched on this above but is worth returning to now because I think her response to it generates a problem for neo-Kantianism. For Kant, the morally virtuous agent is one who has a good will: they do their duty for duty's sake. When we consider a particular case though, this starts to look problematic. Imagine a daughter who goes to visit her mother who is ill in hospital. If we asked why she does this, imagine if she responded 'Simply because it was my duty'. This sounds shockingly cold. There are a number of reasons you could visit your mother – to make her feel loved, to brighten her day up, because you love her and want to see her, etc. In some cases, these reasons might not weight heavily for you, and your sense of duty intercedes and you visit the old bat anyway. But this shouldn't be the typical case. The truly virtuous person should visit her mother for the other reasons, not just because it is her duty. What has happened, in effect, is that Kant's theory makes the virtuous person fetishistic, in

the sense discussed in chapter 2, this is the **SOK**.[156] Korsgaard is worried about this objection and thinks it can be surmounted by defending a particular thesis about action.

She wants to contrast her own conception of action with a 'Millian' one, where "All action is for the sake of some end, and rules of action, it seems natural to suppose, must take their whole character and colour from the end to which they are subservient." (Mill 1998, 2) On this model whether an action is good or not depends upon what effects it produces. This *production* conception of action is one that Korsgaard thinks is deeply ingrained in philosophers, even Kantians. Kantians sometimes put forward their moral principles as 'side-constraints' – as restrictions on the right ways to realise certain ends. If we think of action as something that is judged on the basis of the ends it realises, then these side-constraints will appear mysterious. If action is judged by what it brings about, how can the way in which it brings an end about matter to its moral value?

Korsgaard wishes to overturn this state of play with a different conception of action (one she claims is also found in Aristotle and Kant). She argues that we need to distinguish between *acts* and *actions*. The *act* is what the Millian has in mind when they talk about actions. But, for Korsgaard an *action* is not merely performing an act, there is something else as well. In Kant (so Korsgaard claims) the description of an action is a maxim. The maxim has this structure:

Act + Purpose

└ Action ┘

[156] And it looks as if things are worse than for the externalist who the fetishism argument was targeted against above. There it emerged that the externalist could dodge the worst of the fetishism charge by pointing out that having a *de dicto* desire to the right thing is compatible with having *de re* desires for the right-making features we'd expect a non-fetishist to be moved by. In this case it looks like Kant is claiming that *all* moral actions should be motivated by duty.

On this model, to perform an action is to perform an act for some particular purpose – to bring about some end. The categorical imperative test therefore applies to an act done for some particular purpose: you ask not whether it is permissible to lie, but whether it is permissible to lie in order to save someone's life. So, despite the traditional view of Kant, he has no time at all for general moral principles as far as they attach to *act* types.

With this new conception of action we can hope to dissolve the **SOK**. Whether some action is good, and thus should be done out of duty, is determined not only by the act, but also by the purpose the act is intended to bring about. So, it is not the act of visiting your mother that you should do out of duty. As it stands, 'visiting your mother' is just a description of an act, and so is not yet a candidate for moral judgement (this is why utilitarianism is not even a moral theory, according to Korsgaard, as it tries to judge acts, which Korsgaard claims just aren't the kind of things that can have moral statuses attached to them). Instead, what is up for evaluation is the action: 'visiting your mother in order to show her you love her'. Now this, the action described by a maxim with the structure of act + purpose, is what you should do out of duty. So acting out of duty is not some cold, fetishistic process. Acting out of duty involves doing an act for the sake of some purpose – and this purpose can be something warm and touchy-feely, like making your mother feel loved.

There are two things we might wonder about this – whether this is an accurate representation of Kant; and whether this response to the **SOK** works. I am only concerned here with the second question. We can start to make problems for Korsgaard by asking what is supposed to be playing a role in motivation here, the duty or the purpose?[157]

[157] It should be possible to stay agnostic about what theory of motivation we are committed too – for the Humean, it won't be the duty or purpose playing a motivational role, rather it will be the desire to do one's duty, or the desire to fulfil that purpose which will. However, it should be possible to state the

So we can ask whether the duty (or desire to do your duty) plays any motivational role. Korsgaard could answer no. This means that the fact that something is your duty is in a way entirely epiphenomenal – it being the case that some action is your duty will not make a difference to anything you actually do. This makes it hard to see how Kant's claim that the morally virtuous person is one who acts out of duty is even a candidate for being true – if we take this option it makes little sense to say that anyone ever 'acts out of duty', at least on an obvious reading of what this means (that the duty motivates you). Call this option (**A**) (I shall return to this later).

So let's imagine instead that Korsgaard answers 'yes, duty plays a motivational role'. Now we can ask what would happen if somebody saw that something was their duty, but failed to be motivated by the purpose picked out by the description of the action. (So suppose I could see that it was my duty to visit my mother in order to show her I loved her, but that I didn't want to show her I loved her). What would happen in this case? Would I still perform the act (visiting my mother)? If we answer 'no' then Korsgaard is claiming that you need both the purpose, and your duty to perform the action. Call this option (**B**). If we answer 'yes' then we are faced with another question – what would happen if we wanted to fulfil the purpose, but we lacked the recognition that the action was part of our duty? Would we still perform the act then? It might seem obvious that the answer has to be 'no' – remember we are considering the options for someone who thinks that duty has some motivational role, and we have already bracketed off the position that says you need both the purpose and the duty (**B**). The only thing left with motivational impact in this case would then be the duty, and if we take that away then it must be the case that we won't perform the action. However, this ignores a possible position that claims that duty has motivational force, the

---

objection in either Humean or anti-Humean terms. I shall just talk of 'the duty' and 'the purpose', a Humean could fill this in with 'the desire to do your duty' and 'the desire to fulfil that purpose'

purpose also has motivational force, but you don't need both together in order to perform the action. The way to make this claim is to say that you performing the action is in a way overdetermined. Call this option (**C**). The other option is to answer 'no' to the last question – that is to say that it is that you only need the recognition of a duty to secure a morally virtuous agent doing the right thing (option **D**). I wish to claim that none of these possible accounts of the relationship between duty and motivation is attractive.

**D** is the least helpful option. Here we are claiming that the morally virtuous agent is motivated by acting out of duty alone. This means that the purpose is entirely epiphenomenal, in the way the duty was on option **A**. This is straightforwardly fetishistic, and would mean that for all her new-fangled machinery Korsgaard has no response to the **SOK**. She can't mean to occupy this position.

**C** claims that you would do the right act if you had the purpose, or if you had the duty, or if you had both. However, when you have both it is not the case that one or other's motivational force is switched off – they both still have full motivational force in the case where both are present. This view does at least seem to get us around the original problem – the morally virtuous person is one who acts out of duty. But this does not preclude them from also acting with a certain purpose in mind. So they can meet the Kantian standard of virtue without being fetishistic. However, the picture it gives us seems very strange – our actions (when we are acting virtuously) would be overdetermined, in the sense familiar from debates in the philosophy of mind. It seems difficult to understand how you could make sense of this view. In any case, if this is what Korsgaard intends, she needs to do more work to motivate it.

**B** claims that you need both the purpose and the duty in order to act. Either one on their own won't do, but together they are jointly sufficient. This also seems to solve the problem: you can have the touchy-feely purpose AND the duty – so you can be a good person on Kant's account and avoid fetishism. However, it seems false to claim that you need the duty *and* the right purpose in order to act well. In fact, the anti-Kantian could just reformulate their challenge: the morally virtuous person is someone who is motivated by the heart-warming purpose of the act (e.g. to make their mother feel loved) on its own. If the duty is also necessary, then it means that the considerations about love would not be enough on their own, and this still seems too fetishistic. It is still true that you would not visit your mother unless it was your duty. What this reveals is that having duty play an essential role in motivation allows the anti-Kantian space to restate their argument. Having duty play this role appears to taint the whole action.

There is something more that Korsgaard could say on option **A**. **A** says that the duty is motivationally inert. How then can we make sense of the claim that a good action is one done out of duty? One way might be to distinguish between explanatory (or motivational) and justifying reasons. So the fact that something would show your mother that you loved her explains your action (in that it explains your motivation) whereas the fact that it is also your duty merely justifies your action. So, to be a fully good agent is to be motivated by the warm cuddly features of the action you bring about, but to be justified by the fact that doing the action is your duty. You only become fetishistic when the duty functions as a motivational reason.

This, though, can't be what Korsgaard intends. Korsgaard wants to defend an internalism about reasons, where having a reason is intrinsically motivational (1986). The only way for you to not be moved by a reason is for you to suffer some form of practical irrationality

(depression, fear, etc). So, if in order to respect the Kantian conception of a virtuous agent you must have a reason that doesn't motivate you (but merely justifies) then the only way to be virtuous would be to be practically irrational. This obviously can't be Korsgaard's intention. So, **A** is (on one reading) unavailable to Korsgaard and on another a failure (it leaves us without a substantial account of what 'acting out of duty' is). **B** seems to be only some improvement and leaves open the possibility for the anti-Kantian to refine the **SOK**. **C** solves the problem, but involves a strange metaphysical view that Korsgaard needs to do more work to motivate. And **D** straight-forwardly fails to tackle the **SOK**. My conclusion is that if the **SOK** is a problem, then Korsgaard's use of the distinction between actions and acts to try to dissolve it is inadequate as it stands.

To sum up, then. We began this chapter wanting to be able to give a clear formulation of Korsgaard's positive metaethical suggestions. I've argued that if we formulate neo-Kantian constructivism as a form of cognitivist anti-realism along the lines of Wright's account of judgement-dependent qualities we secure a number of benefits – we get a formulation of constructivism that gives the constructivist most of the features they want from a metaethical theory, but which is also clearer to evaluate. When formulated this way, the credibility of constructivism depends upon the constructivist's ability to give a kosher derivation of the categorical imperative. I then attempted to lay out Korsgaard's argument for the bindingness of the categorical imperative in terms which, again, make it easier to evaluate. This argument has a number of possible gaps, some of which I have not been able to investigate. Korsgaard tries to plug some of these gaps by invoking a type of constitutivism. This part of her view runs into serious problems accounting for what goes on the case of people who violate the categorical imperative. If, then (as I have argued) the credibility of neo-Kantian constructivism depends upon being able to give an argument for the categorical imperative, and that argument depends on an implausible form of

constitutivism, then I have to conclude that neo-Kantian constructivism fails. Finally I looked at whether Korsgaard could avoid the **SOK** and argued that her view, as it stands, does not.

# CONCLUSION

This thesis has been based on the presumption that Korsgaard's forays into metaethics are worth engaging with, even if ultimately her arguments against her metaethical competitors can be demonstrated to fail, and her own position faces serious difficulty. Reflection on neo-Kantian constructivism has enabled me to reach a number of, I hope, independently interesting conclusions.

In the first chapter we saw that, in the process of making space for Korsgaard's argument against realism, that there are good reasons to think that our conception of metaethics should be more expansive than the one sometimes put forward, and that there are surprising connections between normative ethics and metaethics.

To get clear on precisely what Korsgaard's complaint against realism is I suggested that we read her as concerned with the motivational import of moral judgements (this is not a characterisation she would rush to endorse herself, but the idea is that we could capture a lot of what might animate someone's concerns with realism by thinking about things in this way – even if Korsgaard wouldn't be happy with this reading, perhaps it would be useful to someone who, like Korsgaard, takes her normative question seriously). Putting things this way allowed me to reach conclusions about Smith's treatment of amoralism, his argument from fetishism, and van Roojen's battery of considerations in favour of internalism. There I argued that there are no compelling reasons to commit to internalism, and that if this

reading of Korsgaard is what was bothering us then the realist can surmount the difficulty by demonstrating there is a viable form of externalist moral realism available.

Chapter three considered another way of reading Korsgaard – as offering a generalised anti-voluntarism argument. This argument afflicts both realism and Korsgaard's own position, and thus reveals something interesting not about realism but about the nature of normative explanations. To avoid this argument the realist needs a form of reductionism. I then considered whether Cornell realism and Finlay's reductivism are tenable positions for the moral realist. Finlay's position has a number of strengths, but reflection on the account of analyticity he appeals to opens up problems for his methodology. Cornell realism can defend itself against the attack launched on its semantic programme by Horgan and Timmons by getting clear on what that programme is doing. In addition, it's not obvious that the Cornell realist's strategy for earning ontological rights for moral properties fails, despite the objections it faces. The upshot of all this is that there are two viable versions of realism that survive one of both ways of construing Korsgaard's argument.

Korsgaard's complaints against expressivism prompted an investigation of the main issue with that metaethical position – the Frege-Geach problem. We saw there how careful attention to what the Frege-Geach problem is getting at (explaining the compositionality of language using only elements available to an expressivist) militates against at least Ridge's version of hybrid expressivism. We also saw how other versions of hybrid expressivism are more properly construed as sophisticated types of moral realism. The outcome of this discussion is that hybrid expressivism offers little hope to the expressivist, and the fortunes of expressivism then seem to depend on how much complexity we are willing to permit in

our semantic theory to secure expressivism's great metaphysical and epistemological benefits.

Finally (chapter five) I've tried to offer a new way of understanding Korsgaard's own positive position. Neo-Kantian constructivism is best thought of as a version of cognitivist anti-realism for two reasons: this way of doing things secures most of the features a neo-Kantian constructivist wants from a metaethical theory; and it gives us a clear way of presenting and evaluating the constructivist's claims. When we set things up this way attention shifts to whether the Kantian can give a viable argument for the categorical imperative. When we look at how Korsgaard gives that argument we learnt a number of things: that its viability depends upon a type of constitutivism; that constitutivism depends upon various dubious claims about the inescapability of agency; that the argument fails to respect the conditions on giving a judgement dependent treatment of moral qualities; that we need to be concerned with the truth of the psychological claims Korsgaard needs; that the argument depends upon a limited conception of what human freedom involves. In addition I argued that Korsgaard's argument against a standard objection to Kant is inadequate as it stands. For these reasons I conclude that her own positive metaethical position, though worth consideration, fails.

# BIBLIOGRAPHY


Altham, J. 1984: "The Legacy of Emotivism", in G. Macdonald and C. Wright (eds.) *Fact, Science and Morality* (Oxford: Basil Blackwell).

Armstrong, D.M. 2004: *Truth and Truthmakers* Cambridge: Cambridge University Press.

Ayer, A.J. 1936: *Language, Truth & Logic* New York: Dover.

Bar-On, D. 2004: *Speaking My Mind: Expression and Self-Knowledge* Oxford: Oxford University Press.

Bar-On, D. and Chrisman, M. 2009: "Ethical Neo-Expressivism" *Oxford Studies in Metaethics* **4** Oxford: Oxford University Press.

Barker, S. 2000: "Is Value Content a Component of Conventional Implicature?" *Analysis* **60**, 268-279.

Barker, S. Ms: *Global Expressivism* Available at: http://eprints.nottingham.ac.uk/696/1/BOOKGE.pdf

Beebee, H, and Sabbarton-Leary, N. 2010: "Are Psychiatric Kinds 'Real'?" *European Journal of Analytic Philosophy* **6**, 11-28.

Blackburn, S. 1973: "Moral Realism" reprinted in his *Essays in Quasi-Realism* Oxford: Oxford University Press, 1993. 111–29.

Blackburn, S. 1984: *Spreading the Word* (Oxford: Oxford University Press).

Blackburn, S. 1993a: *Essays in Quasi-Realism*. (Oxford: Oxford University Press).

Blackburn, S. 1993b: "Realism, Quasi or Queasy?", in J. Haldane and C. Wright (eds.) Reality, Representation, and Projection (Oxford: Oxford University Press).

Blackburn, S. 1993c: "Circles, Finks, Smells and Biconditionals" *Philosophical Perspectives* **7**, 59–80.

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

Blackburn, S. M.s: "All Souls Night" URL: www.phil.cam.ac.uk/~swb24/PAPERS/Allsoulsnight.htm

Boghossian, P. 1996: "Analyticity Reconsidered" *Nous* **30**, 360-91.

Boghossian, P. 1997: "Analyticity", in Wright, C. And Hale, B. (eds.) *A Companion to the Philosophy of Language* Oxford: Blackwell.

Boghossian, P. 2011: "Review of Truth in Virtue of Meaning" *Australasian Journal of Philosophy* **89**, 370-74.

Boisvert, D. 2004a: “What is Expressivism” at *PEA Soup* URL: http://peasoup.typepad.com/peasoup/2004/06/the_embedding_o.html

Boisvert, D. 2004b: “Four Kinds of Expressivism” at *PEA Soup* URL: http://peasoup.typepad.com/peasoup/2004/06/the_embedding_o_1.html

Boisvert, D. 2004c: “The Objection From Truth Ascriptions” at *PEA Soup* URL: http://peasoup.typepad.com/peasoup/2004/07/the_embedding_o.html

Boyd, R. 1988: “How to be a Moral Realist” in *Essays on Moral Realism* G. Sayre-McCord (Ed.) Ithaca: Cornell University Press.

Boyd, R. 1991: “Realism, Anti-Foundationalism and the Enthusiasm for Natural Kinds” *Philosophical Studies* **61**, 127-48.

Boyd, R. 2010: “Realism, Natural Kinds and Philosophical Methods” in Beebee, H. And Sabbarton-Leary, N. (eds.) *The Semantics and Metaphysics of Natural Kinds* New York: Routledge.

Bradley, D.J. 2011: “Functionalist Response-Dependence Avoids Missing Explanations” *Analysis* **71**, 297-300.

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

Brink, D.O. 1997: “Moral Motivation” *Ethics* **108**, 4-32.

Burge, T. 1979: “Individualism and the Mental” in Rosenthal, D. (ed.) *The Nature of the Mind* London: Oxford University Press.

Chrisman, M. 2008: “Expressivism, Inferentialism, and Saving the Debate” *Philosophy and Phenomenological Research* **77**, 334-58.

Chrisman, M. 2010: Expressivism, Inferentialism, and the Theory of Meaning. In Michael Brady (ed.), New Waves in Metaethics. London: Palgrave-Macmillan.

Cohen, G.A. 1978: *Karl Marx’s Theory of History: A Defence* New Jersey: Princeton University Press.

Colyvan, M. 2011: "Indispensability Arguments in the Philosophy of Mathematics", *The Stanford Encyclopedia of Philosophy* Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2011/entries/mathphil-indis/>.

Copp, D. 2000: “Milk, Honey, and the Good Life on Moral Twin Earth” *Synthese* **124**, 113-137.

Copp, D. 2001: “Realist-Expressivism: A Neglected Option for Moral Realism”, *Social Philosophy and Policy* **18**, 1-43.

Copp, D. 2009: “Realist Expressivism and Conventional Implicature” *Oxford Studies in Metaethics* **4** Oxford: Oxford University Press.

Chalmers, D. 2002: “Consciousness and its Place in Nature” In Stich, P. and Warfield, T. (eds.) *Blackwell Guide to the Philosophy of Mind*. Oxford: Blackwell.

Chalmers, D. 2010: “Merely Verbal Disputes” *Arche research seminar* audio available: http://jonathanichikawa.net/podcast/

Chalmers, D. 2010: “Constructing the World” *John Locke Lectures* audio available: http://www.philosophy.ox.ac.uk/lectures/john_locke_lectures/past_lectures

Dancy, J. 2000: “The Particularist's Progress” in Hooker, B. and Little, M. 2000 (Eds.) *Moral Particularism* Oxford: Clarendon Press.

Dancy, J. 2004: *Ethics Without Principles* Oxford: Clarendon Press.

Dancy, J. 2005: “Moral Particularism” found on the *Stanford Encyclopedia of Philosophy* URL: http://plato.stanford.edu/entries/moral-particularism/

Darwall, S., Gibbard, A. and Railton, P. 1992: “Toward Fin de Siècle Ethics: Some Trends” *Philosophical Review* **101**, 115-89.

Davidson, D. 1967: “Truth and Meaning”, reprinted in his *Inquiries Into Truth and Interpretation* Oxford: Oxford University Press 1984.

Divers, J. and Miller, A. 1994: “Why Expressivists About Value Should Not Love Minimalism About Truth”, *Analysis* **54**, 12-19.

Divers, J. and Miller, A. 1995: “Platitudes and Attitudes: A Minimalist Conception of Belief”, *Analysis* **55**, 37-44.

Dreier, J. 1990: “Internalism and Speaker Relativism” *Ethics* **101**, 6-26.

Dreier, J. 2004: “Meta-Ethics and the Problem of Creeping Minimalism” *Philosophical Perspectives* **18**, 23–44.

Dretske, F. 2004: “The Case Against Closure” in Steup, M. (ed.) *Contemporary Debates in Epistemology* Oxford: Blackwell.

Dupre, J. 1993: *The Disorder of Things* Cambridge, MA: Harvard University Press.

Enoch, D. 2006: “Agency, Shmagency: Why Normativity Won’t Come from What is Constitutive of Agency” *Philosophical Review* **115**, 169–198.

Evans, G. 1973: “The Causal Theory of Names” *Proceedings of the Aristotelian Society* **47**, 187-212.

Field, H.H. 1980: *Science Without Numbers: A Defence of Nominalism* Oxford: Blackwell.

Ferrero, L. 2009: “Constitutivism and the Inescapability of Agency” *Oxford Studies in Metaethics* **4** Oxford: Oxford University Press.

Finlay, S. 2004: “The Conversational Practicality of Value Judgement,” *Journal of Ethics* **8**, 205-223.

Finlay, S. 2005: “Value and Implicature,” *Philosophers’ Imprint* **5**, 1-20.

Finlay, S. Forthcoming: *A Confusion of Tongues*.

FitzPatrick, W.J. 2005: “The Practical Turn in Ethical Theory” *Ethics* **115**, 651-91.

Frankenna, W. K. 1939: “The Naturalistic Fallacy” *Mind* **48**, 464-77.

Geach, P.T. 1958: “Imperative and Deontic Logic”, *Analysis* **18**, 49-56.

Geach, P.T. 1960: “Ascriptivism” *Philosophical Review* **69**, 221–5.

Geach, P.T. 1965: “Assertion” *Philosophical Review* **74**, 449–65.

Gert, J. 2006: “Problems for Moral Twin Earth Arguments” *Synthese* **150**, 171 - 183.

Geuss, R. 1996: “Morality and Identity” in *The Sources of Normativity* Cambridge: Cambridge University Press.

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

Gibbard, A. 2003: *Thinking How to Live* Cambridge, MA: Harvard University Press.

Goff, P. 2007: "A Non-Eliminative Understanding of Austere Nominalism" European Journal of Philosophy **16**, 43–54.

Hale, B. 1986: “The Compleat Projectivist”, *Philosophical Quarterly* **36**, 65-84.

Hale, B. 1993a: “Can There be a Logic of Attitudes?”, in J. Haldane and C. Wright (eds.) *Reality, Representation, and Projection* Oxford: Oxford University Press.

Hale, B. 1993b: “Postscript”, in J. Haldane and C. Wright (eds.) *Reality, Representation, and Projection* Oxford: Oxford University Press.

Harman, G. 1975: “Moral Relativism Defended” *Philosophical Review* **84**, 3-22.

Harman, G. 1986: “Moral Explanations of Natural Facts” *Southern Journal of Philosophy* **24**, 57-68.

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

Hare, R.M. 1970: “Meaning and Speech Acts” *The Philosophical Review* **79**, 3–24.

Haukioja, J. 2006. “Why the New Missing Explanation Argument Fails, Too” *Erkenntnis* **64**, 169–75.

Hawthorne, J. 2004: "The Case for Closure" in Steup, M. (ed.) *Contemporary Debates in Epistemology* Oxford: Blackwell.

Henning, T. 2011: "Moral Realism and Two-Dimensional Semantics" *Ethics* **121**, 717-48.

Hobbes, T. (1651) 1991: *Leviathan* Edited by Richard Tuck, Cambridge: Cambridge University Press.

Hood, C. 2010: *Ethics and Intention* unpublished PhD Thesis, University of Birmingham.

Hooker, B. and Little, M. 2000 (Eds.): *Moral Particularism* Oxford: Clarendon Press.

Horgan, T. and Timmons, M. 1991: "New Wave Moral Realism Meets Moral Twin Earth" *Journal of Philosophical Research* **16**, 447-65.

Horgan, T. and Timmons, M. 1992: "Troubles for New Wave Moral Semantics: the 'Open Question Argument' Revived", *Philosophical Papers* **21**, 153-75.

Horgan, T. and Timmons, M. 1996: "From Moral Realism to Moral Relativism in One Easy Step" *Crítica* **28**, 3-39.

Horgan, T. and Timmons, M. 2000: "Copping Out on Moral Twin Earth" *Synthese* **124**, 139-152.

Horgan, T. and Timmons, M. 2009: "Analytical Moral Functionalism Meets Moral Twin Earth" In Ian Ravenscroft (ed.), *Minds, Ethics, and Conditionals: Themes from the Philosophy of Frank Jackson* Oxford: Oxford University Press.

Hussain, N. 2010: "Error Theory and Fictionalism" in Skorupski, J. (ed.) *The Routledge Companion to Ethics* London: Routledge.

Hussain, N. and Shah, N. 2005: "Is Constructivism an Alternative to Realism?" Unpublished Ms.

Hussain, N. and Shah, N. 2006a: "Misunderstanding Metaethics: Korsgaard's Rejection of Realism" in *Oxford Studies in Metaethics* **1**: (2006), 265-94.

Hussain, N. and Shah, N. 2006b: "Metaethics and its Discontents – A Case Study of Korsgaard" Unpublished Ms URL: http://www.amherst.edu/~npshah/Shah/papers/md.pdf

Ichikawa, J., Maitra, I. and Weatherson, B. 2011 (online preview): "In Defence of a Kripkean Dogman" *Philosophy and Phenomenological Research* URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1933-1592.2010.00478.x/full

Jackson, F. 1992: "Critical Notice" *Australasian Journal of Philosophy* **70,** 475 –88.

Jackson, F. 1998: *From Metaphysics to Ethics* Oxford: Oxford University Press.

Jackson, F. and Pettit, P. 1990: "Program Explanation" *Analysis* **50**, 107-17.

Jackson, F. and Pettit, P. 1993: "Structural Explanation in Social Theory" in David Charles and Kathleen Lennon (eds.) *Reduction, Explanation, and Realism* Oxford: Oxford University Press.

Jackson. F. and Pettit, P. 1995: "Moral Functionalism and Moral Motivation" *Philosophical Quarterly* **45**, 20-40.

Jackson, F. and Pettit, P. 1998: "A Problem For Expressivists" *Analysis* **58**, 239-51.

Jackson, F., Pettit, P. and Smith, M. 2000: "Ethical Particularism and Patterns" in Hooker, B. and Little, M. 2000 (Eds.) *Moral Particularism* Oxford: Clarendon Press.

Johnston, M. 1989: "Dispositional Theories of Value" *Proceedings of the Aristotelian Society* **63** (Suppl.), 139–74.

Johnston, M. 1993a: "Objectivity Refigured: Pragmatism Without Verificationism" In *Reality, Representation and Projection,* eds. J. Haldane and C. Wright Oxford: Oxford University Press.

Johnston, M. 1993b: "Remarks on Response-Dependence" Unpublished Ms.

Johnston, M. 1998: "Are Manifest Qualities Response-Dependent?" *The Monist* **81**, 3–44.

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

Kaplan, D. 1989: "Demonstratives" in Almog, J., Perry, J., and Wettstein, H. (eds.) *Themes From Kaplan* New York: Oxford University Press.

Kauppinen, A. 2007: "The Rise and Fall of Experimental Philosophy" *Philosophical Explorations* **10** 95-118.

Kölbel, M. 1997: "Expressivism and the Syntactic Uniformity of Declarative Sentences" *Critica* **29,** 3–51.

Korsgaard, C. 1986: "Skepticism about Practical Reason" *Journal of Philosophy* **83**, 5–25.

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

Korsgaard, C. 1997: "The Normativity of Instrumental Reason." In Cullity and Gaut, eds., *Ethics and Practical Reason* Oxford: Oxford University Press.

Korsgaard, C. 2003: "Realism and Constructivism in Twentieth-Century Moral Philosophy" *Journal of Philosophical Research* APA Centennial Supplement, 99–122.

Korsgaard, C. 2009: *Self-Constitution* Oxford: Oxford University Press.

Kripke, S. 1980: *Naming and Necessity* Oxford: Blackwell.

Leiter, B. 2001: "Moral Facts and Best Explanations" *Social Philosophy and Policy* **18**, 79-101.

Leiter, B. 2005: “The Hermeneutics of Suspicion: Recovering Marx, Nietzsche, and Freud” *University of Texas Law, Public Law Research Paper* **72**.

Lenman, J. 2003: “Disciplined Syntacticism” *Philosophy and Phenomenological Research* **66**, 32-57.

Lenman, J. 2006: "Moral Naturalism", *The Stanford Encyclopedia of Philosophy* Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2008/entries/naturalism-moral/>.

Lenman, J. 2008: “Against Moral Fictionalism” *Philosophical Books* **49**, 23-32.

Levy, N. (2011): “Resisting Weakness of Will” *Philosophy and Phenomenological Research* **82**, 134-55.

Lewis, D. 1979: “Counterfactual Dependence and Time's Arrow”, *Nous* **13,** 455–76.

Lewis, D. 1984: “Putnam’s Paradox” *Australasian Journal of Philosophy* **62**, 221-36.

Lillehammer, H. 1997: “Smith on Moral Fetishism” *Analysis* **57**, 187-95.

Lillehammer. H. 2007: *Companions in Guilt* London: Palgrave.

Lowe, E.J. 1995: “The Truth About Counterfactuals” *Philosophical Quarterly* **45**, 41-59.

McDowell, J. 1998: “Values and Secondary Qualities” in *Morality and Objectivity: A Tribute to J.L. Mackie*. T. Honderich (ed.) 110-25. London: Routledge (1985). Reprinted in *Mind, Value and Reality* 1998 Cambridge MA: Harvard University Press.

McDowell, J. 1998: *Mind, Value and Reality* Cambridge, MA, Harvard University Press.

McFarland, D. 1999: “Mark Johnston’s Substitution Principle: a New Counterexample? *Philosophy and Phenomenological Research* **59** 683–89.

McFarland, D. and Miller, A. 1998: “Response dependence without reduction?” *Australasian Journal of Philosophy* **76**, 407-25.

McKeever, S. and Ridge, M. 2005: “What Does Holism have to do with Particularism?” in *Ratio* **18**, 93-103.

Machery, E. Mallon, Nichols, S. and Stich, S. 2004: “Semantics, Cross-Cultural Style” *Cognition* **92**, B1-B12.

Machery, E. Mallon, Nichols, S. and Stich, S. 2009: “Against Arguments from Reference” *Philosophy and Phenomenological Research* **79**, 332–356.

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

Melia, J. 1995: “On What There’s Not” *Analysis* **5**, 223-29.

Mill, J.S.M. 1998: *Utilitarianism* Crisp, R. (ed.) Oxford: Oxford University Press.

Miller, A. 1995: "Objectivity Disfigured: Mark Johnston's Missing-Explanation Argument" *Philosophy and Phenomenological Research* **55**, 857–68.

Miller, A. 1996: "An Objection to Smith's Argument for Internalism" *Analysis* **56,** 169–74.

Miller, A. 1997: "More Responses to the Missing-Explanation Argument" *Philosophia* **25**, 331–49.

Miller, A. 2001: "The Missing-Explanation Argument Revisited" *Analysis* **61**, 76–86.

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

Miller, A. 2007: *Philosophy of Language* Oxford: Routledge.

Miller, A. 2009: "Reply to Nelson" *Australasian Journal of Philosophy* **87**, 337-41.

Miller, A. 2010: "Non-Cognitivism", in J. Skorupski (ed.) *The Routledge Companion to Ethics* Routledge: London.

Moore, G.E. 1903: *Principia Ethica* Cambridge: Cambridge University Press.

Nelson, M. 2006: "Moral Realism and Program Explanation" *Australasian Journal of Philosophy* **84**, 417-28.

Papineau, D. Ms.: "The Rise of Physicalism" URL: http://sas-space.sas.ac.uk/881/1/D_Papineau_Rise..pdf

Peacocke, C. 2000: *Being Known* Oxford: Oxford University Press.

Pettit, P. 1991: "Realism and Response Dependence" *Mind* **100**, 587–626.

Pettit, P.1996: "Realism and Truth" *Philosophy and Phenomenological Research* **56**, 881-8.

Pettit, P. and Menzies, P.. 1993: "Found: The Missing Explanation" *Analysis* **53**, 100–09.

Price, R. 1948: *A Review of the Principal Questions in Morals* Raphael, D.D. (ed.) Oxford: Oxford University Press.

Price, H. 1994: "Semantic Deflationism and the Frege Point", n S.L. Tsohatzidis *Foundations of Speech Act Theory: Philosophical and Linguistic Perspectives* London: Routledge.

Price, H. Forthcoming: *Naturalism Without Mirrors* Oxford: Oxford University Press.

Prichard, H.A. 1912: "Does Moral Philosophy Rest on a Mistake?" *Mind* **21**, 21-37.

Pufendorf, S. (1672) 1934: *On The Law of Nature and of Nations* (C.H. Oldfather and W.A Oldfather trans.) Oxford: Oxford University Press.

Putnam, H. 1975: "The Meaning of 'Meaning'" in *Mind, Language and Reality* Cambridge: Cambridge University Press.

Quine, W.V.O. 1935: "Truth by Convention" in *The Ways of Paradox and Other Essays* New York: Random House.

Quine, W.V.O. 1948: "On What There Is" *The Review of Metaphysics* **2**, 23-40.

Quine, W.V.O 1951: "Two Dogmas of Empiricism" *Philosophical Review* **60**, 20-43.

Railton, P. 1986: "Moral Realism" *The Philosophical Review* 163-207. Reprinted in Darwall, S., Gibbard, A. and Railton, P. *Moral Discourse and Practice* 1997 Oxford: Oxford University Press.

Railton, P. 1989: "Naturalism and Prescriptivity" *Social Philosophy and Policy* **7**, 151-74.

Ridge, M. 2006: "Ecumenical Expressivism: Finessing Frege", *Ethics* **116**, 302-36.

Ridge, M. 2007a: "Epistemology For Ecumenical Expressivists", *Proceedings of the Aristotelian Society* Supplementary Volume **81**, 83-108.

Ridge, M. 2007b: "Ecumenical Expressivism: The Best of Both Worlds?", in R. Shafer-Landau (ed.) *Oxford Studies in Metaethics* **2** Oxford: Oxford University Press.

Ridge, M.. 2009: "The Truth in Ecumenical Expressivism", in David Sobel and Stephen Wall (eds.) *Reasons For Action* Cambridge University Press.

van Roojen, M. 1996: "Expressivism and Irrationality", *Philosophical Review* **105**, 311-35.

van Roojen, M. 2006: "Knowing Enough to Disagree: A New Response to the Moral Twin Earth Argument" In Russ Shafer-Landau (ed.) *Oxford Studies In Metaethics* **1** Oxford: Oxford University Press.

van Roojen, M. 2010a: "Moral Rationalism and Rational Amoralism" *Ethics* **120**, 493-525.

van Roojen, M. 2010b: "Reply to Shafer-Landau" at *PEA Soup* URL: http://peasoup.typepad.com/peasoup/2010/07/ethics-discussions-at-pea-soup-mark-van-roojens-moral-rationalism-and-rational-amoralism-with-commen.html

Rorty, R. 1979: *Philosophy and the Mirror of Nature* New Jersey: Princeton.

Sainsbury, R.M. and Tye, M. 2011: "An Originalist Theory of Concepts" *Aristotelian Society Supplementary Volume* **85**, 101-24.

Salmon, N. 1986: *Frege's Puzzle* Cambridge, MA: MIT Press.

Sayre-McCord, G. 1997: "'Good' on Twin Earth" *Philosophical Issues* **8**, 267-292.

Scanlon, T. 2009: "Being Realistic about Reasons" *John Locke Lectures* audio available: http://www.philosophy.ox.ac.uk/lectures/john_locke_lectures/past_lectures

Schroeder, M. 2005: "Cudworth and Normative Explanations", Journal of Ethics and Social Philosophy **1**, 1-27.

Schroeder, M. 2008a: "What is the Frege-Geach Problem?", *Philosophy Compass* **3/4**, 703-720.

Schroeder, M. 2008b: "How Expressivists Can and Should Solve Their Problem with Negation" *Noûs* **42**, 573-99.

Schroeder, M. 2008c: *Being For: Evaluating the Semantic Program of Expressivism* Oxford: Oxford University Press.

Schroeder, M. 2008d: "Expression for Expressivists" *Philosophy and Phenomenological Research* **76**, 86-116.

Schroeder, M. 2009: "Hybrid Expressivism: Virtues and Vices", *Ethics* **119**, 257-309.

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

Schroeder, M. 2010: "Getting Noncognitivism Out of the Woods" *Analysis* **70**, 129-139.

Schroeder, M. 2011: "What Matters About Metaethics" In Peter Singer (ed.), *Does Anything Really Matter? Responses to Parfit*.

Schroeder, M. forthcoming: "Finagling Frege" (unpublished MS).

Searle, J. 1962: "Meaning and Speech Acts" *Philosophical Review* **71** 423–32.

Setiya, K. 2003: "Explaining Action" *Philosophical Review* **112**, 339-93.

Setiya, K. 2007: "Cognitivism About Instrumental Reason" *Ethics* **117**, 649-73.

Shope, R.K. 1978: "The Conditional Fallacy in Contemporary Philosophy" *Journal of Philosophy* **75**,397-413.

Sinclair, N. 2009: "Recent Work in Expressivism" *Analysis* **69**, 136-147.

Sinclair, N. 2011: "Moral Expressivism and Sentential Negation", *Philosophical Studies* **152**, 385-411.

Smith, M. 1994: *The Moral Problem* Oxford: Blackwell.

Smith, M. 1994b: "Why Expressivists About Value Should Love Minimalism About Truth" *Analysis* **54**, 1 - 11.

Smith, M. 1994c: "Minimalism, Truth-Aptitude and Belief" *Analysis* **54**, 21 - 26.

Smith, M. 1997: "In Defense of 'The Moral Problem': A Reply to Brink, Copp, and Sayre-McCord" *Ethics* **108**, 84-119.

Soames, S. 2002: *Beyond Rigidity* Oxford: Oxford University Press.

Soames, S. 2003: *Philosophical Analysis in the Twentieth Century* vols **1** and **2** New Jersey: Princeton University Press.

Soames, S. 2010: *Philosophy of Language* New Jersey: Princeton University Press.

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

Stratton-Lake, 1999: "Why Externalism is not a Problem for Ethical Intuitionists" *Proceedings of the Aristotelian Society* **99**, 77-90

Street, S. 2008: "Constructivism about Reasons" in R. Shafer-Landau (ed.) *Oxford Studies in Metaethics* **3** Oxford: Oxford University Press.

Street, S. 2010: "What is Constructivism in Ethics and Metaethics?" *Philosophy Compass* **5,** 363-84.

Sturgeon, N. 1985: '"Moral Explanations" in D. Copp and D. Zimmerman (eds.), *Morality, Reason and Truth* New Jersey: Rowman and Allanheld.

Sturgeon, N. 1986: "Harman on Moral Explanations of Natural Facts" *Southern Journal of Philosophy* **24**, suppl., 115-42.

Sturgeon, N. 1992: "Nonmoral Explanations" *Philosophical Perspectives* **6**, 97-117.

Suikkanen, J. 2007: "The Proto-Frege-Geach Problem" at *PEA Soup* URL: http://peasoup.typepad.com/peasoup/2007/05/the_protofregeg.html

Suikkanen, J. 2010: "Non-Naturalism: The Jackson Challenge" in R. Shafer-Landau (ed.) *Oxford Studies in Metaethics* **5** Oxford: Oxford University Press.

Surgener, K. and Miller, A. Under review: "Against Ecumenicism in Metaethics".

Tappolet, C. 1997: "Mixed Inferences: A Problem for Pluralism About Truth Predicates", *Analysis* **57**, 209-10.

Unwin, N. 1999: '"*Quasi*-Realism, Negation and the Frege-Geach Problem" *The Philosophical Quarterly* **49** 337–52.

Unwin, N. 2001: "Norms and Negation: A Problem for Gibbard's Logic" *The Philosophical Quarterly* **51** 60–75.

Urmson, J. 1968: *The Emotive Theory of Ethics* New York: Oxford University Press.

Way. J. 2010: "The Normativity of Rationality" *Philosophy Compass* **5,** 1057-68.

Wedgwood, R. 2007: *The Nature of Normativity* Oxford: Oxford University Press.

Wedgwood, R. 2010: "Schroeder on Expressivism: For – or Against?" *Analysis* **70**, 117-29.

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

Williamson, T. 2007: *The Philosophy of Philosophy* Oxford: Blackwell.

Wright, C. 1988: "Realism, Anti-Realism, Irrealism, Quasi-Realism", reprinted in his *Saving The Differences* Cambridge MA: Harvard University Press 2003.

Wright, C. 1991: “Scepticism and Dreaming: Imploding the Demon” *Noûs* **25**, 205.

Wright, C. 1992: *Truth and Objectivity* Cambridge, MA: Harvard University Press.

Zangwill, N. 1992: “Moral Modus Ponens”, *Ratio* **5**, 177-93.