How to be a child, and bid lions and dragons farewell:

The consequences of moral error theory

David James Hunt

A thesis submitted to the University of Birmingham for the degree of Doctor of Philosophy

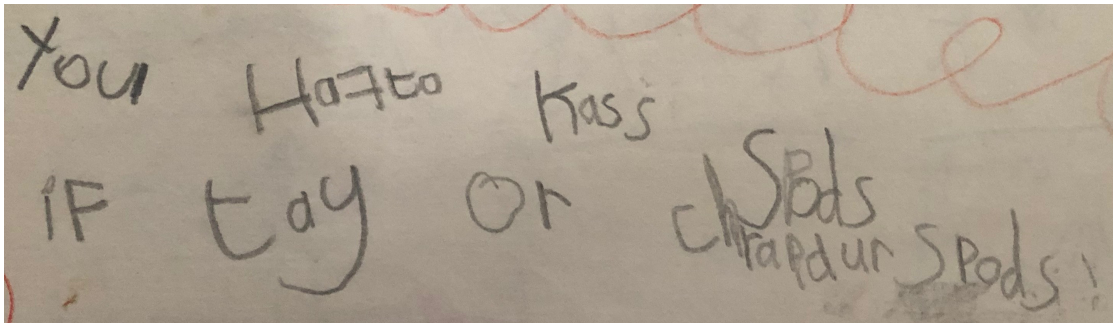
Department of Philosophy
College of Arts & Law
University of Birmingham
December 2019

Abstract

Moral error theorists argue that moral thought and discourse are systematically in error, and that nothing is, or can ever be, morally permissible, required or forbidden. I begin by discussing how error theorists arrive at this conclusion. I then argue that if we accept a moral error theory, we cannot escape a pressing problem – what should we do next, metaethically speaking? I call this problem the 'what now?' problem, or WNP for short. I discuss the attempts others have made to respond to the WNP, and in each case I show that the responses fail to be satisfying. I then propose a new response to the WNP, which I call revolutionary relativism. I define revolutionary relativism, explain why it is preferable to the existing responses to the WNP, and defend it against the most problematic objections I anticipate that opponents might raise. I conclude that revolutionary relativism succeeds where previous WNP responses fail, and that if we accept a moral error theory, we should become revolutionary relativists.

Dedication

This thesis is dedicated to my mom.



We've come a long way.

Acknowledgements

I am deeply indebted and grateful to my supervisor, Jussi Suikkanen, for his guidance, support and constructive criticism over the years it has taken to complete this thesis. I am also grateful to the audiences at the University of Birmingham and at the Joint Session of the Aristotelian Society & the Mind Association who offered helpful criticism when I presented a draft version of part of this thesis as a paper.

On a more personal level, words cannot express my gratitude to my parents for their belief, support and enthusiasm, or to Laura for her certainty that I would succeed and her willingness to listen to me bang on about whatever I was (enthusiastically but no doubt largely incomprehensibly) preoccupied with on any given day.

Table of Contents

Chapter 1. Introduction

Moral error theorists invite us to imagine a world not without religion, countries or possessions, but without something arguably even more fundamental – morality. At least, that is one way of reading the conclusion error theorists reach, which is that all of morality is systematically, unavoidably in error, and that as a result there is nothing – indeed there *can never be anything* – which is morally required, permitted or forbidden.[1] As conclusions in philosophy go, this is about as dramatic as it gets. Yet if the error theorists are right, we cannot simply accept their conclusion and leave it at that. A profound and daunting question immediately looms: what on earth should we do about it? This thesis is an attempt to provide a new answer to that question.

One of the functions of the introduction to a thesis such as this is often to explain why the subject under examination is important, and why the contribution which the thesis makes to the topic matters. In this case, this step almost seems redundant. If issues in philosophy can ever be important at all, and surely they can, then there must be few issues more important than whether there is any kind of moral reason why, for example, we shouldn't all just murder each other right now. And there can be few philosophical questions more pressing than what we should do if we come to accept the potentially shocking conclusion that no such reason exists. If it also turns out that none of the answers to that question which people have offered so far is satisfactory, as I will argue, then finding a new answer which *is* satisfactory is a near-vital task.

[1] Although error theories about other domains of discourse exist (see e.g. Miller 2013 pp. 105-108 for a discussion of an error theory of colour), throughout this thesis I will refer almost exclusively to *moral* error theory. That being the case, I will sometimes omit the word moral and refer to just *the* error theory, *an* error theory, or simply error theory. Unless specified otherwise, wherever terms such as error theory or error theorist are used, it may be assumed that it is a moral variety I am referring to.

Broadly speaking, the thesis has three parts. The first part, comprising chapters 2 and 3, sets the stage by outlining why and how moral error theorists argue that we must accept an error theory of traditional morality. Not only is this important in allowing the reader to orient themselves within the philosophical terrain at hand, it also introduces and defines the terms of debate and thus the conceptual underpinnings and commitments of many of the arguments in the chapters which follow. It is not my aim here to defend error theory, but it is crucial to what follows to explain how and why others do so. And one cannot have a comprehensible error theory of something without being clear from the outset what that something is.

Accordingly, chapter 2 introduces moral error theory and explains in detail how error theorists typically view traditional morality – the nature of moral thought and discourse, whether and how moral judgements can be true, what is really going on when we utter a sentence such as 'torture is wrong', and so on. And at the end of the chapter I will offer a formulation of the error theory which cuts across the variations between different error theorists, and forms the basis for the discussion throughout the rest of the thesis.

Having established what exactly it is that error theorists typically believe morality consists in, chapter 3 will lay out why they think it is infected with systematic error. I will focus on two sophisticated and influential arguments for a moral error theory, and show why I believe that the shortcomings of one argument highlight why the other argument is effective. I will also take note of the commitments error theorists typically take on in the course of their arguments, and which we must therefore be careful not to violate as we go on to figure out what to do if error theorists are right.

In the next broad section of the thesis in chapters 4 and 5, I will prepare the ground for my own proposed response to the truth of a moral error theory, and then offer the response

itself. In chapter 4, I explain why coming to accept a moral error theory means that we are unavoidably confronted by something I call the 'what now?' problem. While others have discussed the existence of such a problem and sought to respond to it, I feel that it can sometimes be under appreciated quite how pressing a problem it is. I attempt to correct this by explaining in some detail what the problem is, why we cannot avoid confronting it, and how important it is that we do so successfully. I lay out some 'ground rules' for how we might respond to the 'what now?' problem, including highlighting several of the commitments of the arguments for error theory. I then discuss the main responses to the 'what now? problem proposed by others to date. In each case I argue that the response in question fails to respond adequately to the problem at hand. The failure of all of the main ways of responding to the 'what now?' problem which others have argued for to date means that we must find a new way of responding.

In chapter 5, I describe in detail my own, new response, which I call revolutionary relativism. Briefly, I argue that if we accept an error theory of morality, we should respond to this by adopting a form of relativism which replaces the beliefs about moral reasons for action which we previously embraced with beliefs about practical norms which are accepted by our communities. The precise form of relativism I argue we should adopt has numerous features which may not be typical among conventional relativist views. I therefore explain in detail what the atypical or novel features of my proposed form of relativism are, and why they are advantageous in the post-error-theory context.

The final main part of the thesis in chapters 6 and 7 is devoted to defending the proposal I make in chapter 5, bearing in mind the commitments 'inherited' from the arguments for error theory and the reasons for the failure of previous responses to the 'what now?' problem. In chapter 6 I argue that my proposal is preferable to the existing responses to the 'what now?'

problem. For each existing response, I recap the objections I raised in chapter 4, and then show why revolutionary relativism can avoid or better cope with each objection.

In chapter 7 I defend my proposal against several key direct objections. First I discuss the strongest challenges to traditional forms of relativism, and show that my proposal can cope with them and thus stand on its own two feet, philosophically speaking. Then I discuss the strongest objections to my proposal which could be made specifically in the post-error-theory context. No one can predict and anticipate every potential objection to any view, but the objections I discuss in this chapter are the most problematic I can foresee.

In chapter 8 I draw all of the above together and offer my conclusion. To state my conclusion as succinctly as possible, I believe that revolutionary relativism can succeed where others fail and be a satisfying response to the 'what now?' problem. As such, revolutionary relativism is an important new contribution to this area of metaethics.

Chapter 2. Introduction to Moral Error Theory

While its roots stretch back much further in time, the point of departure for moral error theory in most contemporary debates is J. L. Mackie's 1977 book, *Ethics: Inventing Right and Wrong*.[2] In *Ethics*, Mackie argues that we cannot avoid the conclusion that morality itself is fundamentally and systematically flawed, and as a result, that all affirmative first order moral judgements (for example that torture is wrong) are in error (1977 p. 49). In modern parlance, Mackie and subsequent error theorists ultimately argue that that there is nothing – nor can there ever be anything – which is morally permitted, required or forbidden. As a result, error theorists conclude that whenever anyone judges that *torture is wrong* or that *we are morally obliged to help others in need*, they are always mistaken.

The impact and implications of this, both in terms of debates within philosophy and in terms of our everyday moral lives, cannot be underestimated. Indeed, Simon Robertson is hardly exaggerating when he writes 'The opening chapter of John Mackie's *Ethics: Inventing Right and Wrong* reset the metaethical agenda of the late twentieth century' (2008 p. 107).[3] Moral thought and discourse are so fundamentally intertwined with human society and psychology that many people would find it difficult to accept that there is nothing which we morally ought or ought not to do. After all, the committed moral error theorist must agree that not only is there no moral reason to pay taxes or to refrain from breaking promises, but also that there is nothing morally wrong with rape, that we have no moral grounds on which to criticise the

[2] For an insightful commentary on some of the historical precursors to Mackie's moral error theory, see Olson 2014, part 1.

[3] Also deserving of a mention in this context is John Burgess, who was working along similar lines to Mackie at around the same time, but whose paper *Against Ethics* remained unpublished until much later (see Burgess 2007).

state-ordered crucifixion of teenagers, and conversely that there is nothing morally good about acting to prevent children – even one's own children - dying of starvation.[4]

Virtually anyone capable of understanding moral error theory will find these latter claims jarring, even alarming. We have grown up in societies where many kinds of action, especially actions which are brutally violent or predatory, are seen as simply evil – never, ever to be permitted and always to be emphatically opposed. Likewise, the idea that some actions are morally good arguably plays an important role in our wellbeing by contributing towards a feeling that we are doing the right thing; that we are living good lives. Whether absorbed from our societies or arrived at by independent thought, moral considerations are deeply embedded in our identities as human beings.

Probably the most natural reaction to learning about moral error theory, then, is to wonder how exactly the error theorists can arrive at such stark conclusions. As compelling as all this may be, however, my aim here is not to defend a moral error theory. Mackie's original arguments have been the subject of no small amount of discussion already, and numerous philosophers have subsequently taken up - and significantly advanced - the cause of error theory. Most influential among them is Richard Joyce, who, along with others such as Jonas Olson and Bart Streumer, has re-presented moral error theory in a way which avoids many of the pitfalls which commentators find in the arguments presented by Mackie in *Ethics*. In this thesis I will have little to add to the excellent work of these philosophers and others like them in discussing arguments for and defences of error theory. Rather, my aim here is to consider what we can or should do if the error theorists are right. If there is indeed nothing which is morally required, permitted or forbidden, and we come to accept and understand this, then

[4] The crucifixion of teenagers has been the subject of outcry in recent years, see e.g. Mezzofiore 2015.

what, if anything, should we do about it? I call this the 'what now?' problem, and the ultimate aim of this thesis is to propose a new response to this problem.

In order to prepare the way for my proposal, however, we must first get clear about what today's error theorists actually say. Only then will we be equipped to consider possible responses to the 'what now?' problem from an informed position and in a way which will allow us to check for inconsistencies with arguments deployed by error theorists *en route* to an error theory, and so be sure that the responses we are considering are not self-undermining. I will begin in sections 2.1 to 2.1.4 by outlining the nature of traditional morality, i.e. the morality which most people currently believe in, according to error theorists.[5] Then in section 2.2, I will draw together the chapter as a whole and give a clear and concise statement of the error theory as I will understand it throughout the rest of this thesis.

I will then move on in chapter 3 to explain why error theorists argue that a moral error theory is inescapable. I will start with a brief overview of Mackie's seminal discussion, and then move on to present what most current commentators take to be the best kind of argument for a moral error theory. Having established all of this, we will then be in a position to consider the motivating question behind this thesis in an appropriately informed and rigorous way – if the error theorists are right, what should we do about it?

[5] A note on terminology which it will be important to bear in mind throughout this thesis. Whenever I use the term morality without qualification, I am referring to this 'traditional' or 'folk' morality which error theorists claim is the best analysis of moral thought and discourse as understood and used by most people who have not yet accepted the truth of a moral error theory (or otherwise come to understand moral thought and discourse in a way which is relevantly consistent with accepting a moral error theory).

2.1 The nature of morality according to error theorists

A persuasive argument for an error theory of any domain of thought or discourse must begin with an analysis of the target domain. Only if that analysis a) is more plausible than any competing analyses of the target field and b) leads inevitably to an error theory should we accept the error theory in question is true.[6] Accordingly, moral error theorists must begin by arguing for a particular analysis of traditional morality. It is not my task to argue for or against error theorists' typical analysis of traditional morality. But for my thesis to make sense to the reader as we move on, it is vital to establish exactly what error theorists claim about morality. In other words, if we are going to consider what we should do if the error theorists are right, we will need to know exactly what the error theorists are right *about*.

Probably the most direct way to approach this is to examine what error theorists typically claim is going on when we sincerely make a judgement which can be expressed by a simple indicative subject-predicate sentence, such as 'torture is wrong'. I will run through and briefly explain each of the important features of error theorists' typical claims about this kind of judgement and about sincere utterances of sentences which express this kind of judgement. In what follows, we should bear in mind that error theory is a second order view. Error theorists are not primarily concerned with the moral status of one kind of action versus

[6] I draw here form Daly & Liggins' discussion (2010 p. 219) of an argument by Crispin Wright against error theory. That discussion is couched in terms of discourse, whereas here I also include moral thought, i.e. moral judgements as well as moral sentences. Note that Daly & Liggins refer to the best analysis, as opposed to one which is more plausible than any competing analyses. I prefer my own formulation because I believe that it can be doubted whether any analysis of a domain of discourse which leads to an error theory can be properly called the best analysis unless is it also plausibly the correct analysis. For if two analyses are very nearly as plausible as one another, but one of them leads to an error theory, I believe a significant proportion of people would consider this a reason to find the analysis which does not result in an error theory the more plausible of the two. This is similar but slightly different to Wright's worry which Daly & Liggins discuss. However for economy's sake I set this issue aside here.

another. Rather, error theorists are concerned initially with analysing what right*ness* and wrong*ness* are, and whether anything can instantiate them.

2.1.1. Assertion

In non-moral contexts, basic subject-predicate sentences are typically used to make assertions. For example, when we speak the words 'the earth is flat', the most straightforward interpretation of this utterance is as an assertion of the proposition *the earth is flat*, i.e. as the speaker presenting this proposition as *true*.[7] This interpretation is borne out by the way we use such sentences in conversational contexts – conversationally appropriate responses to 'the earth is flat' might include 'no it isn't' or 'I couldn't agree more'. These responses are not how we respond to non-assertoric uses of language, such as commands or expressions of desire, which are not truth-apt. Rather, they are typical of how we respond to assertions which present propositions as true, and which can be doubted or believed.[8]

Moral error theorists typically claim that the same is true in the moral case. Joyce in particular is very explicit about the fundamental assertoricity of moral discourse, saying that at their most basic level, 'moral utterances turn out to be *assertions*' (2001,p. 15, emphasis original). According to error theorists, when we say 'torture is morally wrong', we are (typically primarily) asserting – i.e. presenting it as true - that *torture is morally wrong*, or alternatively that *people morally ought not to torture others*. Again, this interpretation is plausibly borne out by the way we use basic indicative moral sentences in conversational contexts. Burgess

[7] See e.g. Wright's claim that there is a 'basic, platitudinous connection of assertion and truth: asserting a proposition-a Fregean thought-is claiming that it is true' (1992 p. 23).

[8] There are of course numerous other uses for this kind of sentence which would be permissible in English – a speaker uttering 'the earth is flat' could be being sarcastic, or could be lying, and so on. But these are not the most basic primary uses of such sentences, and their effect arguably depends on a background linguistic convention of interpreting subject-predicate sentences as assertions. These non-assertoric uses of language are something I will return to in later chapters. But for now, we need only consider the most basic, sincere use of this kind of sentence.

offers examples: 'One who says "Abortion is as wicked as murder," may meet with the response, "I doubt that very much!" or "Do you sincerely believe that?"' (2007 p. 429). These responses are conversationally appropriate, and support the claim that moral discourse is primarily assertoric.[9]

2.1.2. Belief

At the same time as making an assertion, when we sincerely utter 'the earth is flat', we are also typically expressing a mental state - a judgement about the flatness or otherwise of the earth. There are numerous kinds of mental state which we can express through language – hopes, fears, beliefs, expectations, demands and so on. But in the non-moral case of sincerely uttering 'the earth is flat', it seems most reasonable to take the relevant mental state to be a belief. As with assertion, this is borne out by the way we use such sentences in conversational contexts; if someone says 'the earth is flat', it would be conversationally appropriate to agree or to contradict them, whereas it would be nonsensical to reply as if to an expectation or command or by saying 'don't worry, one day you'll look back on it and laugh'.

This is not the place for a full account of the exact nature of beliefs. But a reasonably standard account of beliefs would include that belief is a mental state which consists in having an attitude towards some proposition such that the believer takes the proposition to accurately represent the state of affairs in the world. One common way of putting this is to say that beliefs have a mind-to-world direction of fit, i.e. that what is in the mind – the propositional content of the mental state of belief - aims to fit with the way things are in the world.[10] Typically, this means that beliefs are sensitive to evidence; should we believe that *p*, and then

[9] The term assertoric is often connected with speech act theory. Indeed it would be possible to rephrase some of the above in terms of speech acts, e.g. in terms of perlocutionary aims (see e.g. Austin 1962 p. 94*ff.*). But I take the term assertoric to be sufficiently intuitive that further digression into such territory is not required here. For an overview of speech act theory, see Green 2017.

[10] See e.g. Railton 1994 for a discussion of the influential idea that beliefs by definition 'aim at truth'.

encounter evidence that *not-p*, we will typically cease to believe that *p*, and instead come to believe that *not-p*. Thus if we believe that the world is flat, but then encounter conclusive evidence that the world is round, we will typically cease to believe that the world is flat, or else risk of being considered irrational or delusional. This contrasts with non-cognitive attitudes such as desires, which are commonly thought to have a world-to-mind direction of fit, and so to not be sensitive to evidence in the same way - if we desire that *p*, and have evidence that *not-p*, then we will typically continue to desire that *p*, and be motivated to change the way things are in the world such that *p* obtains.[11]

This brings out a further distinction between beliefs and other mental states such as desires, namely that beliefs are commonly thought to be motivationally inert, i.e. there seems to be no necessary link between having a belief and being motivated to act. By contrast, desires are commonly thought to be motivational, i.e. my having a desire seems to play an important causal role in my acting so as to bring that desire about. For example, if I have a desire relating to the printer on my desk, say a desire to print a document, then this seems to play a key role in causing me to use the printer on my desk to print a document. Yet my belief that there is a printer on my desk seems unlikely to result in any action in particular unless it is accompanied by a desire which would be satisfied by using the printer.

Again, moral error theorists typically think the same is true in the moral case, and argue that moral judgements are beliefs. This view is known as moral cognitivism.[12] According to cognitivists, when we sincerely say 'torture is morally wrong', we are expressing a belief that

[11] For an overview of belief, see Schwitzgebel 2015. For more on direction of fit and the sensitivity of beliefs to evidence, see Humberstone 1992 (which includes discussion of various authorities in the field, hence my claim that a standard view of belief would include the features discussed above) and Smith 1994 p. 7 & §4.6, pp. 111-116.

[12] Since I will hereafter use the unqualified term cognitivism exclusively to refer to cognitivism about morality, i.e. the view that moral judgements are beliefs, and will not use it to refer to other domains of cognitivism (e.g. the school of thought in psychology which sought to respond to behaviourism), I will frequently omit the 'moral' from here onwards.

acts of torture have a property of moral wrongness, or alternatively that it is morally obligatory for all agents to refrain from torture. And again, this is borne out by our use of moral terms, and Burgess' examples quoted above serve to support a cognitivist interpretation of morality as well as the view that moral discourse is assertoric.

A further reason why error theorists are typically cognitivists about morality is that, if torture is wrong, it is commonly thought that this is somehow an objective matter. We speak and apparently think as if whether or not torture is wrong is something we know, not as if the wrongness of torture is a matter of our non-cognitive attitudes such as our desires, expectations, hopes and so on. There is an apparent objectivity or externality about putative moral facts which is consistent with moral facts being the kind of thing we believe in, and thus with moral judgements being beliefs.[13]

This is not to say that moral error theorists, or cognitivists more generally, deny that moral utterances *can* express mental states other than beliefs. For it is abundantly clear from observing our moral discourse that we do indeed use it to express emotions, issue commands and so on. But these other ways in which we may use moral discourse are pragmatic rather than semantic – they are matters of what we can do with moral utterances rather than what the linguistic meaning of moral utterances is. In terms of their linguistic meaning, cognitivists hold that moral utterances conventionally express beliefs. Analogies can readily be drawn to many other areas of discourse, for example an exasperated parent shouting 'two plus two is four!' at a child who has got their sums wrong can easily be interpreted as expressing emotions, or demanding something of the child. But this does not affect the fact that the primary, conventional role of 'two plus two is four' in our discourse is to assert a proposition

[13] As Mackie puts it, 'It is a very natural reaction to any non-cognitive analysis of ethical terms to protest that there is more to ethics than this, something more external to the maker of moral judgements, more authoritative over both him and those of or to whom he speaks' (1977 p. 32). This phenomenology, i.e. the apparent 'externality' of moral facts is also discussed by Smith (1993).

and (arguments in the philosophy of mathematics notwithstanding) express a belief. And moral error theorists take this to be the primary role of indicative sentences in our moral discourse.

2.1.3. Moral normativity

Now we turn to moral normativity, the sense in which if torture is morally wrong, that means in some way or other that one ought not or is obliged not to torture. There is a quality to moral obligation which, although intuitively quite easy to understand, is tricky to pin down and isolate when discussing metaethics.[14] Suppose that we believe that torture is wrong, and we encounter a person who habitually tortures others. The moral wrongness of torture, according to our belief, means that we will likely condemn the torturer's actions, and perhaps we will attempt to convince or even compel her to refrain from torture in the future. We may well also feel outrage or a desire to punish her for her moral transgressions. Underlying all of these reactions is a sense in which we believe that the wrongness of torture somehow makes it the case that people should not engage in torture; that there is an authoritative moral obligation or reason for all agents not to torture other agents. And it is a distinctive feature of moral reasons that they apply to all agents, regardless of agents' desires or ends.

Various terms are used to attempt to capture this sense of normativity, such as 'objective prescriptivity', 'authority', the somewhat colloquial 'non-evaporatibility' (Joyce 2001 p. 35) and 'practical clout' (Joyce 2006 p. 57), or the more formal 'count[ing] in favour of or requir[ing] certain courses of behaviour, where the favouring relation is irreducibly normative' (Olson 2014 p. 118). Olson even gives a list (2014 p. 117) of further attempts by others to find an appropriate term. All of these locutions are intended to convey the sense in which

[14] *Cf.* Hattiangadi 2006, p. 228: 'The distinction between hypothetical and categorical 'oughts' is difficult to draw'.

according to error theorists, if one is morally obligated to pursue a course of action, one has that obligation regardless of one's desires, ends, interests or, often, beliefs. Moral facts entail (or simply are) reasons for action which apply to all moral agents simply by virtue of the fact that they are moral agents.[15]

The term I will use is non-institutional categorical reason, or categorical reason for short. To make it as clear as possible what categorical reasons are, consider the following three kinds of practical reason. First, a hypothetical reason, i.e. one which depends on an agent's desires or ends. Say you want to get to town by 11 o'clock, and I know that the bus which stops at the nearby bus stop at 10.30 will get you to town on time. I might say 'you ought to get the 10.30 bus'. This 'ought' implies that you have a reason to do something, but that reason is dependent on your desire to get to town by 11 o'clock. Anyone who lacks such a desire, including you if you change your mind about your plans, also lacks a reason to catch the 10.30 bus.

Second, consider what is often called an institutional categorical reason, i.e. one which applies to agents regardless of their desires or ends, but only in virtue of some institution they are participating in or a role they are playing. For example, if I am playing chess, it is my turn, and I am playing as black, then I have a reason not to move any white pieces, regardless of whether I might wish to do so. The rules of chess and the fact that I am currently playing the game dictate that this is the case. So the reason I have is independent of my desires, but it only applies because I am playing chess, and it does not apply to anyone who is not playing chess.

[15] 'Moral agent' will be taken here to mean any agent capable of moral deliberation and practical choice, i.e. anyone who can make a judgement about what the morally right and/or wrong course of action might be, and who can choose to act accordingly. Further refinements of the terminology of agency etc. are unnecessary for present purposes.

Finally, consider torture. If we believe that torture is morally wrong, according to error theorists this means we believe that all agents are obliged – that is, have a moral reason – not to torture others. And this reason is categorical in a non-institutional sense in that it is an authoritative reason for all agents, no matter what their desires or ends, and no matter what role they might be playing. So to return to our torturer, if she replies to our moral outrage by saying 'But I *really* want to torture people!', we will not reply 'Oh, in that case it's fine, go ahead'. And neither will our moral condemnation lessen if we discover that the torturer has significantly different moral beliefs, according to which torture is morally permissible. Rather, it seems intuitively plausible that we would say that the torturer was wrong, and we were right.

Moral error theorists typically hold that this inescapability, this normative independence from an agent's desires, ends or interests which I am calling categorical normativity is an essential part of morality.[16] Joyce goes so far as to claim that 'we might well think that it is the whole point of having a moral language' (2000 p. 463). Joyce's point is that if we attempt to analyse moral discourse and behaviour in any way which leaves this definitively moral characteristic out of the picture, moral error theorists typically respond that it is not morality which we are analysing, but rather something weaker which merely *appears* to be morality. In Joyce's parlance (e.g. 2001 p. 3), it is a 'non-negotiable' commitment of moral thought and discourse that moral reasons are categorical.

2.1.4. Truth conditions

Error theorists' commitments to the assertoricity of moral discourse and to cognitivism entail that moral beliefs are truth-apt, i.e. can be true or false. But what are the truth conditions of

[16] See e.g. Olson 2010 p. 3, 'What makes moral facts queer is that they make demands from which we cannot escape'.

moral beliefs - in virtue of what are they true or false? As before, consider a non-moral belief that the earth is flat. This can be understood as a belief that the earth has a certain property of flatness. This belief will be true or false dependent on the (natural) facts of the matter - if we think of all of the possible worlds in which there is an earth, then the belief will be true in all those possible worlds in which the earth has the property of flatness, and false otherwise.

For error theorists, the same is true of moral beliefs. That is to say, a belief which ascribes a moral property, e.g. a belief that torture is wrong, is true or false depending on the (moral) facts of the matter - if the actual world is one in which torture has the property of moral wrongness, then the belief is true, otherwise it is false.

For Mackie, moral properties and facts (facts about whether the relevant properties are instantiated) cannot be natural properties and facts, i.e. the kinds of properties and facts which are the domain of the natural sciences. This is because the kinds of natural properties which feature in scientific descriptions of the world do not include any normative properties, i.e. properties which make it the case that we ought or ought not to do anything in particular. Even the natural fact that torture causes suffering does not, according to Mackie (1977 p. 33), make it the case that I ought to refrain from torture unless I also have some desire or end which would be served by so refraining. This, Mackie argues, falls short of the way we think and speak about morality – if torture is wrong, we typically regard that as somehow normative for all agents, regardless of their desires or ends. That someone really wants to torture others is not generally thought to affect whether torture is morally wrong and therefore ought not to be engaged in. Thus moral properties must be, according to Mackie, non-natural properties.[17]

[17] Here Mackie is drawing on a famous passage in Hume's *A Treatise of Human Nature* (Hume 1739 SBN469/T 3.1.1.27), wherein Hume observes that although many philosophers overlook it, there can be no causal link between an 'is' and an 'ought', that is, between a natural and a moral fact. When

More modern error theorists have tended to focus less on the natural or non-natural properties which certain acts or situations may instantiate, and have concentrated instead on the categoricity of moral normativity. Comparatively slight differences in focus or terminology aside, this is the basis of Joyce's sustained argument for an error theory (2001), the most substantial part of Olson's argument (2014 chapter 6), and is the strategy recommended to all would-be error theorists by Robertson (2008). Thus for today's error theorists, for a moral belief such as *torture is wrong* to be true, i.e. for torture to have the property of moral wrongness, it must be the case that there exists a reason for all agents not to torture others which is authoritative for all agents, regardless of those agents' desires or ends, and regardless of any institution they might be participating in. [18]

2.2. Summary, and statement of the error theory

To briefly recap, then, error theorists typically make various claims about the nature of traditional morality – the morality which we, the 'folk' typically speak and think in terms of today. Error theorists claim that the way we think and talk about morality demonstrate that when we judge that *torture is wrong*, that judgement is a belief, and so is truth apt, i.e. capable of being right or wrong. Accordingly, error theorists typically hold that moral discourse is

considering the example of a cruel act, Mackie says that we think that 'it is wrong because it is a piece of deliberate cruelty. But just what *in the world* is signified by this "because"?' (1977 p. 41, emphasis original). Following Hume, Mackie sees no way in which the natural facts of a situation could give rise to normative facts.

[18] It is a matter of debate whether problematic features such as categorical normativity are *entailed* by the content/truth conditions of moral claims, or are *presupposed* by moral claims. I have framed the matter here broadly along the lines of the entailment view (i.e. that 'x is wrong' entails that there are categorical reasons not to x, because there being such reasons is part of the truth conditions of 'x is wrong'). But it should be noted that it is also possible to formulate moral error theory in terms of moral claims presupposing, rather than entailing, the existence of relevant categorical reasons. According to this variant of error theory, since there are no categorical reasons, this presupposition fails. Yet the existence of categorical reasons is not part of the truth conditions of moral claims, and so moral claims are not false, but rather fail to refer, and are thus meaningless. See e.g. Kalf 2013, Perl & Schroeder 2019, Olson 2014 p. 14-15, Sobel 2019, p. 3. Nothing here is intended to turn on this distinction. But readers who take issue with this may substitute 'fail to be true' or 'meaningless' where appropriate without, I believe, fatally undermining my discussion.

assertoric – when we express a moral belief that torture is wrong, we are making a positive claim about the world, and not simply expressing our feelings on the matter or speaking *as if* there were some kind of moral fact about torture. To be more precise, error theorists claim that when we assert that torture is wrong, we are (possibly *inter alia*, but nonetheless vitally) claiming that it is a fact that torture has some kind of moral property such that all agents, regardless of their desires or ends, have inescapable, authoritative reason to refrain from torture. One upshot of this group of claims about traditional morality is that moral beliefs and moral claims can be true only if the relevant kinds of normative properties exist. But – and this is a very weighty philosophical 'but' indeed – error theorists then claim that there are not - and in fact cannot be - any properties of the kind required for beliefs and claims which ascribe moral properties to be true.

Drawing all this together more formally, I can now offer a formulation of the error theory which I will treat as typical for the rest of this thesis. Naturally individual error theorists will have their own more detailed and specific formulations, but mine is a broad formulation which captures the key features of modern moral error theories while avoiding taking sides in the more intricate debates between defenders of specific error theories.

1. Moral judgements are beliefs.
2. Moral utterances are typically assertoric, and express moral beliefs.
3. Moral beliefs ascribe moral properties.
4. Moral beliefs are true only if the actions or situations they are about have the moral properties which those beliefs ascribe to them.
5. No action or situation has or could ever have moral properties.
6. Therefore all moral beliefs and utterances which ascribe moral properties are false.

The foregoing sections have laid out premises 1-4. While these premises are not universally accepted, they are broadly plausible, and are accepted by many philosophers. However premise 5 is quite a philosophical bombshell, and is the focus of intense debate. It is also the crux of the error theory as a whole. Therefore, having laid out the key features of error theorists' view of traditional morality in this chapter, I will turn in the next chapter to the arguments error theorists deploy in support of premise 5, and thus in support of the conclusion that a moral error theory is unavoidable. This will conclude the 'stage setting' part of my thesis, and I will then move on to discuss what we can and should do if we accept that the error theorists are right.

Chapter 3. Arguments for a Moral Error Theory

In the previous chapter I established what typical error theorists claim are the key features of morality, and offered a formulation of a 'master' argument for error theory which broadly captured the main features of the various more nuanced arguments advanced by error theorists to date. At the heart of all moral error theories is a claim which I rendered as

5. No action or situation has or could ever have moral properties.

In this chapter, I will explain how error theorists defend this claim, and thus how they move from their analysis of morality to the conclusion that we must embrace an error theory to the effect that there are no affirmative moral facts, no authoritative moral obligations, and nothing which we are, or can ever be, morally required, permitted or forbidden to do.

The foundational serious and sustained attempt to argue for a moral error theory comes from Mackie in *Ethics* (1977). Most famously, Mackie deployed an 'argument from queerness', i.e. an argument that if we examine various aspects of moral thought and discourse, we find that in order for our moral thought and discourse to be successful, moral facts must be unlike anything else with which we are acquainted. This gives us strong grounds, Mackie argues, to be sceptical that moral facts exist.

As seminal as Mackie's discussion was, however, subsequent error theorists have largely abandoned most of the avenues of argument Mackie identified, and have concentrated on one central aspect of Mackie's discussion.[19] Therefore, while what Mackie has to say is

[19] For a detailed discussion of the arguments found in *Ethics*, including reasons why error theorists should abandon most of Mackie's lines of thought and focus mainly on the normativity issue, see Olson 2014, chapters 5 and 6. Robertson makes a similar argument (2008). And Joyce actually presents a version of one of Mackie's other arguments even though he admits that he finds it implausible, simply

interesting, and important as the basis of the subsequent debates which led to more refined and detailed arguments for error theory, it is not economical in the context of this thesis to dwell on Mackie himself other than to give him due credit as one of the main figures in the history of moral error theory. Having done so, I will now move directly on to discuss moral normativity, and what most metaethicists today take to be the best argument for a moral error theory.

Section 3.1 will introduce the normativity arguments. I will discuss two main varieties of normativity argument, Olson's argument in §3.2, and Joyce's argument in §3.3. It will be important for later chapters to be very clear about these arguments, and so the discussion I will present will be quite detailed - especially Joyce's argument, which will occupy §3.3.1-3.3.8. Finally in §3.4 I will draw together chapters 2 and 3 by summing up the key features of moral error theory and the principal argument for it, and so conclude the 'stage setting' part of my thesis. I will then move on in chapter 4 to argue that the error theory confronts us with a significant problem: if we accept that the error theorists are right, what now?

3.1. Introduction to the normativity arguments

As a general strategy, error theorists' arguments for a moral error theory typically have two parts. First, error theorists seek to identify one or more commitments of moral thought and discourse which are definitive of *moral* thought and discourse *per se*. Joyce calls these 'non-negotiable' commitments (2001 p. 17.).[20] Should any theory fail to take account of even one non-negotiable commitment of moral thought and discourse, it will fail to be a theory of

to demonstrate the structure of the argument about normativity which is his actual argument for error theory (2001 chapter 1).

[20] Note that I will be assuming the analysis of morality described in the previous chapter from here on, even if I do not specify this. Essentially any characterisations of morality may hereafter be read with an implicit preface of 'according to typical error theorists, as outlined in chapter 2' unless stated otherwise.

morality. Second, error theorists then attempt to show that at least one of the identified non-negotiable commitments is fatally problematic, for example that it is sufficiently queer in some way that we should be sceptical that it exists (as e.g. Olson argues), or that we can make no sense of it (as Joyce argues). In combination, these two steps lead us to an error theory of morality.

According to typical error theorists, a key non-negotiable commitment of moral thought and discourse which can be shown to be fatally problematic is the kind of non-institutional categorical normativity I described in §2.1.3.[21] To show how, I will discuss the two most developed and effective normativity arguments deployed by error theorists to date, those from Olson and Joyce. I believe that Joyce's argument is the more successful of the two, and covers key areas which Olson's argument cannot address. However, the reader may disagree, and in any event it is instructive to contrast the two, and so both arguments have a place here.

3.2. Olson's argument

3.2.1. Reducible versus irreducible normativity

We saw earlier in §2.1.3 that one way to understand the desire/end independence of moral facts is to draw a distinction between hypothetical, institutional (also called weak categorical) and categorical reasons for action. Prior to 2014, Olson observed this distinction, and 'maintained that moral facts are queer because they are or entail *categorical reasons*' (2014 p. 117, referring to Olson 2010, emphasis original). However, more recently he came to believe that the distinction is better articulated in terms of reducibly and irreducibly

[21] I leave open the possibility of identifying multiple non-negotiable commitments of moral discourse. Though error theorists typically agree that (at least some variant of) the normativity argument presently under discussion is sufficiently conclusive to underwrite a moral error theory, we should not rule out other arguments which could be made.

normative favouring relations. Thus his view is that 'moral facts are queer [to the extent that we should be skeptics about their existence] in that they are or entail facts that count in favour of or require certain courses of behaviour, where the favouring relation is irreducibly normative' (2014 p. 118). We can simplify this terminology to a distinction between reducibly and irreducibly normative reasons.[22]

Reducibly normative reasons are reasons which reduce to non-normative facts about what promotes the satisfaction of an agent's desires, non-normative facts about correctness norms, or non-normative facts about an agent's roles or rule-governed activities.[23] Slightly expanding on the explanation I gave in the previous chapter, consider the following examples of how reasons can reduce in relevant ways. In the case of desire, we might say that if I desire to get to town before 12.30, then I have a reason to catch the train at 11.30. This reason is reducible to the non-normative fact that the 11.30 train gets to town at 12.15, the non-normative fact that I wish to get to town before 12.30, and what Olson calls the favouring relation between these facts and the action of catching the train. This favouring relation is itself reducible to a further non-normative fact, namely that performing the relevant action (i.e. catching the 11.30 train) while the other facts obtain will (very likely) satisfy my relevant desire. Hence Olson calls this kind of favouring relation a reducibly normative favouring relation. Given the way these facts combine to constitute my reason to catch the train, I will cease to have a reason to do so if I cease to have the relevant desire.

Similarly, in the case of correctness or rule-governed activities, we might say that a tennis player has a reason to serve the ball such that it bounces in a particular marked area of the

[22] Olson himself simplifies his terminology in this way (*ibid*.).

[23] It could be argued that rather than 'the satisfaction of an agent's desires', this might be better phrased as 'the fulfilment of agents' non-cognitive attitudes' or something similar. For example, a desire-like attitude such as hope could ground agents' reasons in a similar manner to desires. I limit the discussion here to desires in order to avoid complicating matters too much at this early stage.

court when they are serving. This reason is reducible to the non-normative fact that hitting the ball in this fashion is correct according to the rules of tennis, the non-normative fact that the agent is currently playing tennis, and a favouring relation between these facts and hitting the ball in the prescribed fashion – a relation which in turn reduces to the non-normative fact that performing the prescribed action while the other facts obtain will conduce to successfully playing tennis. So long as the agent is playing tennis, these facts ground a practical reason for the agent to hit the ball as prescribed. But facts about the rules of tennis cannot ground similar practical reasons for any agent not currently serving in a game of tennis.

And in the case of an agent's roles, we might say that a ship's helmsman has a reason to steer the ship according to the captain's orders. This reason is reducible to the relation between the non-normative fact that performing such tasks is part of the role of being a ship's helmsman, the non-normative fact that the agent is indeed a helmsman, and the favouring relation between these facts and performing the action of steering the ship as the captain orders – a relation which in turn is reducible to facts about properly carrying out the duties of helmsmen. This does not hold for any agent who is not a helmsman (or for the agent in question if she ceases to be a helmsman).[24]

None of these reducibly normative kinds of reasons are problematic for error theorists like Olson. This is because these reasons reduce to facts and relations about which there are no queerness worries – natural, non-normative facts and about agents' desires, roles and so on, and the relevant relations between them, are fine. Yet contrast these types of reason with moral normativity. If we take a putative moral fact, say that it is morally wrong to torture people for fun, then according to error theorists' view of how morality works, this entails that

[24] I am assuming in these examples that the agent is engaging sincerely in these activities/roles. It is of course possible to complicate the picture by stipulating that the agent intends to throw the tennis game for a bet, or intends to be a disobedient helmsman. But hopefully there are enough sporting tennis players and diligent helmsmen that the examples make sense as presented and therefore do their job!

we have a reason to refrain from torture. This reason is irreducibly normative in that it is not reducible to the above kinds of non-normative facts and relations.

If it is a fact that torture is morally wrong, then according to error theorists' analysis of morality, we have reason to refrain from torture. And unlike the above examples, this reason does not cease to apply to us if we have or lack any particular desires, if we are playing a game whose rules require torture, if we are professional torturers and so on. In all of these cases, the reason to refrain from torture which is entailed by the fact that torture is morally wrong continues to be a reason for any agent, regardless of any contingent facts about that agent.[25] In Olson's parlance, the fact that torture is wrong entails facts which count in favour of refraining from torture. Yet moral facts and the favouring relations they may bear to agents' actions cannot be reduced to non-normative facts in the same way as the examples above. Rather, if we attempt to break down moral normativity in the same manner, we will always encounter a putative favouring relation which essentially tells us 'you just have to ϕ, whether you like it or not' (where ϕ denotes an action of some kind), and which cannot be further explained in terms of non-normative facts. Thus we can see that the target of Olson's normativity argument, the location of the queerness he attributes to moral facts, is the favouring relation between moral facts and the courses of action they count in favour of.

[25] Strictly speaking, according to Olson the moral wrongness of torture entails a fact (which does not itself need to be normative), say that torture causes human suffering, and that fact grounds a reason for all agents to refrain from torture, irrespective of whether they want to avoid inflicting suffering. That any agent has such a reason is entailed by the fact that torture is morally wrong. Thus there is an extra link or two in Olson's chain of reasons. I omit this detail because I worry that this potentially equates to a circularity along the lines of 'the reason torture is morally wrong is that torture is morally wrong'. Whether this is an issue is not important for my discussion. I take the above discussion to get the presently required point across even with this omission.

Olson formulates his normativity argument accordingly (2014 p. 123-4) (I will add the prefix O, denoting Olson, to the steps in order to avoid confusion with other arguments which follow): [26]

O1. Moral facts entail that there are facts that favour certain courses of behaviour, where the favouring relation is irreducibly normative.

O2. Irreducibly normative favouring relations are queer.

O3. Hence, moral facts entail queer relations.

O4. If moral facts entail queer relations, moral facts are queer.

O5. Hence, moral facts are queer

If this argument is sound, then Olson takes himself to have successfully made the case for the queerness of moral facts and properties, and can then move on to pressing the argument that we should be skeptics about the existence of anything which is queer in the specified way.

3.2.2. Problems with Olson's normativity argument

Let us appraise Olson's argument. Olson points out (2014 p. 124-135) that error theorists may not have to agree with O1, and that at least one commentator, Stephen Finlay, has disagreed with it at length (Finlay 2008, 2009 & 2011). However my task is not to defend error theory, and in this chapter we are explicitly examining those arguments for an error theory which are predicated on something like the view of moral normativity captured by O1. That being the case, and since in the context of the foregoing sections here Olson's premise is plausible, I will

[26] I have amended Olson's numbering system of P12, C5 etc. in order to avoid confusion in the present context.

grant O1. If O2 is true, then O4 seems unlikely to be particularly contentious. The premise I wish to focus on is therefore O2. It seems to me that this is the key premise in Olson's argument, since if it cannot be defended, his argument fails to locate any queerness in moral normativity, and hence fails to provide any support for the overall argument that moral facts are queer in a fatally problematic way. Given this, Olson's whole argument would seem to ride on whether O2 can be satisfactorily defended.

The issue with O2 is that Olson actually says very little to defend it. Essentially, his support for this premise boils down to a few sentences culminating in 'When the irreducibly normative favouring relation obtains between some fact and some course of behaviour, that fact is an irreducibly normative reason to take this course of behaviour. Such irreducibly normative favouring relations appear metaphysically mysterious. How can there be such relations?' (2014 p. 136). At face value, this is not an argument at all, but simply a rhetorical question. And as is universally acknowledged in philosophy, rhetorical questions are not arguments. Olson seems to be saying that irreducibly normative reasons are *queer* because they appear *mysterious*. This kind of near-synonymic table thumping does not tell us anything.

Olson observes that opponents (specifically non-naturalists, who claim that morally normative facts or properties exist, though are unlike the kind of natural facts or properties which are the domain of the natural sciences) can refuse to acknowledge that there is anything queer about such favouring relations, and suggests that their strategy for establishing this could be a 'companions in guilt' argument. This would involve locating other, less contentious examples of irreducible normativity, and claiming that their existence shows that instances of irreducible normativity are not therefore puzzling.[27] However, why should non-naturalists respond in this way? They could reply much more directly that Olson's assertion about the

[27] For an example of this kind of strategy, see Cuneo 2010.

queerness of irreducible normativity is seriously impoverished, and needs considerable supplementation before it presents a challenge which they would have to counter. I will highlight five areas which non-naturalists might point to.

The first area is the unqualified use of '*appear* metaphysically mysterious'. When used in this way, the word 'appear' implies that we apprehend the mysteriousness in question via some kind of intuition. Yet while Olson may have this intuition, many philosophers do not. Olson therefore owes us an explanation of why we should have the same intuition he does. While it might be extravagant for opponents to respond to Olson by demanding an entire theory of epistemic intuition, it seems they would be warranted in demanding an explanation of how intuition shows us that irreducibly normative reasons are queer. For Olson to simply state that he has such an intuition (if this is even how we should read him) is unlikely to convince anyone else. And if Olson does not intend his use of 'appear' in this intuitive way, it is not clear how he does intend it.

Second, whatever it is that Olson means by queer, it is not clear that it amounts to the same thing as mysterious. Many things are mysterious in the sense that they are difficult or even impossible to understand for most people. Numerous phenomena described by quantum mechanics, for example, are certainly mysterious in this way. Yet we do not take this mysteriousness to be reason to doubt their existence. On a much more mundane scientific level, it is a mystery to me why moving certain metals in a magnetic field generates an electrical current. However I do not doubt for a moment that it does. Thus even if we did share Olson's intuition that irreducibly normative reasons are mysterious, this does not force us to accept that they are queer in any way which makes them ontologically suspect.

Third, we might read Olson as suggesting that the queerness of irreducibly normative reasons lies in their difference to other forms of (reducible) normativity. Yet there is no reason to

think that simply because something is different to other things that it does not exist. Examples are ubiquitous – stones are different to colours, mass is different to money, and so on. Even if something is so different from all other things that it is of a unique kind, this is not necessarily a bar to believing it exists. Theists would claim that God was entirely unlike other entities, but this difference can hardly be claimed to support atheism in and of itself. Olson owes us an account of why irreducibly normative reasons are different *in a way which is problematic*.

Fourth, we might think that Olson is appealing to parsimony. In short, if we can explain moral thought and discourse without reference to irreducibly normative reasons, then those kinds of reasons may seem queer because they are surplus to our explanatory requirements. But we must be careful to note where this appeal to parsimony occurs in the argument. At this stage, we are still arguing about whether irreducibly normative relations are queer. Non-naturalists are yet to be convinced that they are queer, and such relations are therefore still on the table as part of a non-naturalist explanation of moral discourse and behaviour. To say that they are surplus to explanatory requirements at this stage is to beg the question against the non-naturalist.[28] Given that irreducibly normative reasons appear in non-naturalists' explanation of moral discourse and practice, Olson needs to show that irreducible reasons are queer without reference to them being explanatorily surplus.

Finally, and most basically, Olson has not demonstrated that the burden of proof lies on non-naturalists. We can read his question 'How can there be such relations?' as a demand for a general theory of reason relations which explains irreducibly normative favouring relations. Yet non-naturalists can respond with their own question – why should there not be such

[28] David Enoch, for example, argues that 'irreducibly normative truths [are] deliberatively indispensable – they are, in other words, indispensable for the project of deliberating and deciding what to do' (Enoch 2011a chapter 3).

relations?[29] If Olson's analysis is correct, we already have widespread acceptance of irreducibly normative reasons, evinced by the putative belief that moral facts exist. The burden of proof would appear at best to be in no mans' land. If the charge of queerness is to be made to stick, Olson needs to demonstrate that the burden of proof lies with non-naturalists, in order that they have a charge to answer. Until he does so, it is not clear that non-naturalists should have any reason to question their existing theories of practical reason.

The general worry highlighted by these five points is that in order for Olson's normativity argument to be convincing, he needs to supply a lot more argumentation to explain why exactly irreducibly normative favouring relations are problematically queer, and why non-naturalists should bear the burden of proof of explaining how such relations might work. This is where I will turn to discuss Joyce, as his normativity argument takes this as its principal theme.

3.3. Joyce's normativity argument

To refresh our memory by making the terms of the debate explicit again, as is typical of error theorists, Richard Joyce would largely agree with the last chapter's characterisation of moral discourse as a field of discourse which aims at asserting facts about the world and of moral judgements as beliefs about those facts. He also agrees that one of the most important, definitive features of moral discourse is that moral obligations are or entail practical reasons – that is, reasons for action – which are irreducibly normative (2006 p. 191). Admittedly Joyce favours terms such as 'strong categorical imperative' (2001 p. 37), but by this he means something very similar to an irreducibly normative reason.[30] He summarises this by saying 'In

[29] Indeed, several people have tried to give an account of how there can be such relations – see Wedgewood (2007), Enoch (2011a) and McDowell (1985) for examples.
[30] It might be desirable to unify Olson's and Joyce's terminology here. However, given the complex and nuanced nature of the arguments, I fear this 'translation' would be in danger of leading to confusion.

short, when we say that a person *morally* ought to act in a certain manner, we imply something about what she would have reason to do regardless of her desires and interests, regardless of whether she cares about her victim, and regardless of whether she can be sure of avoiding any penalties' (2001 p. 134, emphasis original). Observing the discussion of institutional roles above, and the later polemic between Joyce and Finlay (Finlay 2008 2011 & Joyce 2011, 2012), we should perhaps add 'and regardless of any institutions she may be participating in'.

Accordingly, Joyce formulates his normativity argument as follows (2001 p. 77), where x is any moral agent, and ϕ is a placeholder for an action which a moral agent may be required to perform or refrain from performing as a result of their moral obligations (I have added the label J to the steps to denote Joyce):

J1. If x morally ought to ϕ, then x ought to ϕ regardless of what his desires and interests are.

J2. If x morally ought to ϕ, then x has reason for ϕing

J3. Therefore, if x morally ought to ϕ, then x can have a reason for ϕing regardless of what his desires and interests are.

J4. But there is no sense to be made of such reasons.

J5. Therefore, x is never under a moral obligation. [31]

Suffice it to say that for the purposes of this section, we can read Olson-circa-2014's 'facts that count in favour of or require certain courses of behaviour, where the favouring relation is irreducibly normative' as functionally equivalent to Joyce's (and Olson's previous) talk of 'non-institutional categorical reasons' – noting as we do so (and as I noted above) that Olson's updated terminology helps us to be clearer about the targets of moral error theory in some ways.

[31] Joyce later (2012 p. 2) puts his argument in a more stripped down form, as follows:

Comparing this with Olson's argument in §3.2.1, we can clearly see that J4 is very similar to O2, the key step in Olson's argument, which I argued he failed to adequately support. Where Joyce differs crucially from Olson is that in order to support J4, he does indeed supply a theory of normative reasons which goes beyond asserting that such reasons are queer, and actually seeks to show that we cannot make sense of them at all. Therefore Joyce's normativity argument is not intended to support an argument from queerness.[32] Rather it is intended to be a standalone argument for a moral error theory. Unless non-naturalists (or anyone else) can respond with an account of practical reasons which shows that sense can be made of moral obligation, Joyce argues that a moral error theory is inescapable. The burden of proof, according to Joyce (*ibid*.) lies firmly on his opponents.

3.3.1. Joyce's theory of normative reasons

In order to see how Joyce's theory of normative reasons supports his conclusion that we must accept a moral error theory, I will run through the theory quite briskly. This will then frame the discussion in later chapters, forming the basis of what I will take to be an analysis typical among error theorists unless they specify otherwise. Before I begin, we should be careful to observe that Joyce does not take his normativity argument to be definitive of moral error theory. As the later polemic between Joyce and Stephen Finlay (Finlay 2008, 2011 & Joyce 2011, 2012) illustrates, Joyce sees his normativity argument simply as a strong argument in favour of moral error theory. But if his (or indeed any) normativity argument fails, this does not mean that error theory is false (*cf.* Joyce 2012 pp. 2-3). To be clear, I am focusing on

---

A. Conceptually, morality requires non-institutional categorical imperatives.
B. But such things are indefensible.
C. Therefore, moral discourse is bankrupt.
I have chosen to use the slightly fuller version above since it makes it clearer that normativity is the focus.

[32] This is not to say that a proponent of a queerness argument could not make use of Joyce's arguments in this area; there is no obvious incompatibility.

Joyce's normativity argument here not because it is the only possible way to support an error theory, but because it is Joyce's main argument, and is one of the most sophisticated, successful and influential arguments deployed in support of a moral error theory thus far.

Joyce's strategy is to investigate the concept of *having a reason*, and establish a plausible theory of how we can make sense of an agent having an authoritative reason to act, in virtue of which they ought to act in a specific way (i.e. a theory of normative reasons). He then compares this theory with the kind of reason for action entailed by moral obligation. As we have seen, the crucial point of comparison with moral reasons will be whether, given a plausible theory of normative reasons, we can make sense of non-institutional categorical moral reasons and obligations having any authority. Put more simply, the question is whether moral reasons can have the 'practical oomph' our moral discourse commits us to them having.

As noted above, Joyce's arguments in this area are mainly intended to support J4. I will present a distillation of Joyce's arguments in order to pick out the salient points, with the discussion proceeding via the following stages:[33]

H1. Moral reasons purport to be non-institutional categorical reasons, and thus authoritative for all agents (i.e. they are inescapable).

H2. If x is a reason for an agent, S, x must be authoritative for S.

H3. Institutional and hypothetical reasons are authoritative, but not for all agents (i.e. they are escapable).

[33] I have used the label H to denote Hunt, i.e. my own presentation of Joyce's discussion. A more economical formulation may be possible (*cf.* footnote 31 above), but my presentation here should help to make things clearer.

H4. Reasons based on practical rationality are the best candidates for reasons which are authoritative for all agents.

H5. Practical rationality cannot give rise to the same reasons for all agents.

H6. Because of H4 and H5, there are no authoritative non-institutional categorical (i.e. inescapable) reasons.

H7. H1, H2 and H6 taken together entail that there are no categorical reasons of the kind required by moral normativity.

Premises H1 and H3 were explained in the previous chapter (see §2.1.3). To unpack and explain the remaining premises, I will proceed as follows: section 3.3.2 will discuss H2 and how it leads into H4 and H5. I will then explain the reasoning behind H4 in §3.3.3. The arguments supporting H5 are more complex, and so I will split them into several sections to help clarify things. First, §3.3.4 will introduce the various kinds of reason an agent can have, according to Joyce, and will examine how Joyce views the question of whether an agent is practically rational. Then §3.3.5 will specify how the various kinds of reason which an agent may have are to be defined. Following this, I will pause in section 3.3.6 to take stock of the argument thus far.

There will then be two sections looking at the various facets of the question whether practical rationality can yield authoritative reasons for all agents, which is the bone of contention in H5. Section 3.3.7 will consider whether the account up to this point might suggest that moral reasons can be grounded in desires which all agents have, and why Joyce dismisses this notion. Section 3.3.8 will consist in a discussion of whether practical rationality can nonetheless yield reasons which are authoritative for all agents in virtue of what it means to be practically

rational, a suggestion Joyce rejects. Finally, in section 3.4 I will discuss the implications of Joyce's theory of normative reasons for moral error theory in general.

3.3.2. Reasons *for*

We begin the discussion of H2 with a question: what is it for an agent to have a reason for action which is authoritative; which is a genuine reason *for them*? Another way of understanding this is to consider whether, when deliberating upon a suggested reason for action, an agent can sensibly reply 'But what does that mean to me?'.[34] If an agent can sensibly respond with a question like this, Joyce believes that the reason in question cannot be authoritative for them.

Consider a reason which we uncontroversially accept is a reason *for us* to act. If I were to say 'I understand that the fact that the 11.30 train will take me where I want to go at the time I want to go there means I have a reason to catch it, but why should I get on that train?', then I would appear to either be failing to understand the terms I was using, or more simply failing to say anything sensible. The fact that I have a goal I wish to accomplish, and a true belief that actions are available to me which are a way of accomplishing it, is one way of understanding what having a normative reason *is*.

Contrast this with the fact that we can acknowledge and fully understand that, say, tennis players have a reason to serve the ball in a particular way. But if we are not tennis players, it makes no sense to consider that to be a reason *for us* to do anything in particular. Thus when told that one should serve in the manner prescribed, we can sensibly ask, 'But why should I

[34] Joyce 2001 employs various formulations of this question - 'Why?', p. 39, 'But what's that to me?', p. 41, 'So what?', p. 81, 'In virtue of what?', p. 82, 'Why should I ϕ?', p. 83 and so on. I take it that, given the context, the equivalence of variations on this theme is clear to the reader.

do that?'. Asking this question demonstrates that we do not consider ourselves to have a reason which is a genuine reason for us to act.

In order for a normative reason to be a genuine, authoritative reason *for us*, the reason must resemble the train example; it cannot be the case that we can both acknowledge it and yet still ask 'But what is that to me?'. As Joyce puts it, 'Any adequate theory of normative reasons must make out reasons to be precisely those things that forestall a "So what?" response' (2001 p. 81).

If we are to make sense of the concept of *having a reason* which fits with the non-institutional categorical nature of moral reasons, we should expect moral reasons to resist this questioning in the same way. Moral reasons purport to be reasons *for all moral agents*; according to error theorists' analysis of traditional morality, an agent's moral reasons do not track standards of correctness-according-to-morality which can be dismissed if we see ourselves as acting outside of the institution of moral conduct. Rather, moral obligations entail (or simply are) reasons which are authoritative for everyone, regardless of their desires, ends, or any institution in which they may be participating. The tyrant who orders genocide does not escape censure from others by claiming that other people's moral obligations have no authority over tyrants. On the contrary, a moral transgressor who says 'yes, I am fully aware that my actions are morally abhorrent, but what has that got to do with me?' seems to typical participants in moral discourse to have made an error, or to be trying to 'worm their way out' of censure which is nonetheless still deserved. This is precisely because we, the outraged observers, take the reasons entailed by moral facts to be authoritative *for everyone*, including tyrants.

Joyce's strategy is therefore to find the theory of normative reasons which has the best chance of grounding moral reasons in such a way that they can forestall 'So what?' responses. Then,

if he can show that *even this* theory of normative reasons cannot underwrite the putative categorical authority of moral reasons, then he will have demonstrated that the kind of categorical normativity to which morality is non-negotiably committed cannot be made sense of, and so an error theory must follow. Section 3.3.3 will explain why Joyce focuses on practical rationality, and will provide support for what I have labelled H4. Subsequent sections will provide support for H5, and then give the conclusions to be drawn from Joyce's theory of normative reasons.

3.3.3. Practical rationality: authoritative for all

Joyce suggests (2001 p. 49-51) that the best chance of forestalling 'so what?' responses might be offered by grounding moral reasons in the requirements of practical rationality, which he calls 'the framework that tells us what our reasons for acting are' (2001 p. 49). The thought here is that requirements of practical rationality can forestall 'so what?' responses, and so constitute reasons which are authoritative for all agents because the very act of asking the question 'but what does that mean to me?' in response to a suggested reason for action commits one to practical rationality.[35] Thus to ask a question like 'I know that's a requirement of practical rationality, but what has it got to do with me?' makes no sense. This is because asking that question in itself demonstrates that one is 'in the business' of considering and potentially accepting reasons, of taking them to be authoritative, and so on. Therefore no agent can sensibly accept that they have a reason to ϕ which is a requirement of practical rationality but then reply 'but what has that got to do with me?'. Practical rationality, Joyce argues, thus offers hope of grounding the kind of universally authoritative reasons which morality requires.

[35] While Joyce goes on to disagree with the specifics of what she says on the matter (2001 §5.4), this idea of tying the authority of practical reasons to rationality recalls influential work by Christine Korsgaard (1996, especially chapter 3). I will return to this shortly, in §3.3.8.

In fact, this basis in practical rationality seems to be the only promising candidate for successfully forestalling 'so what?' responses, since the authority of normative reasons is always likely to be questionable unless the very act of questioning commits one to the principles on which the theory of normative reasons under consideration is founded. If moral reasons turn out not to be requirements of practical rationality, it is hard to see how they could ever escape the 'So what?' problem and have legitimate authority for all moral agents.

Joyce's task therefore becomes one of analysing practical rationality to see if we can use it to make sense of the kinds of universally authoritative categorical reasons it promises to ground. If we can, then it may be possible to forge an appropriate link between moral reasons and requirements of practical rationality and so to rescue morality from error. But if we cannot, a moral error theory beckons.

3.3.4. What it is to be practically rational: Molly and the cake

The question which therefore arises is, 'what does being practically rational consist in?'. Joyce's answer is that one is practically rational to the extent that one is guided by one's subjective reasons (2001 p. 54). This, of course, requires a taxonomy of reasons which defines and explains what an agent's subjective reasons are, and how they might differ from any other kinds of reasons an agent could have. In order to show how Joyce lays his taxonomy out, let us consider one of his examples, Molly.[36]

Molly spies a cream cake. She likes eating cream cakes, and has a desire to eat it. She is tempted by it, we might say. Absent any other considerations, it seems to fit our everyday

[36] I have condensed and/or paraphrased Joyce's discussions of 'Molly' somewhat here for economy. I have also changed Molly's motivation from losing weight to avoiding hyperlipidaemia.

way of thinking about reasons to think that Molly therefore has a reason to eat the cake. However Molly is also watching her diet. She has thought about this at length, and has decided it is in her best interests to limit her intake of saturated fats (i.e. doing so conduces to her long term ends), and desires to do so. She believes that eating the cake is likely to frustrate this desire. So, we can say that on a different level to her initial reason to eat the cake, she also has a reason not to eat it. On top of this considerable quandary, unbeknownst to Molly, she will be marooned on a barren desert island tomorrow, and would in fact benefit from the extra calories gained by eating the cake now (assuming, plausibly, that not dying of starvation is also something she wants). This is commonsensically understood as constituting another reason for Molly to eat the cake.

There seem to be various kinds of reason which might guide Molly's conduct here. In order to see how Joyce accounts for each of them, and how to understand what Molly's subjective reasons are, we must start with a brief discussion of Hume.

3.3.5. Non-Humean instrumentalism

According to Joyce's reading of Hume, the reasons which it is rational for an agent to be guided by must be grounded in the agent's desires and beliefs. Furthermore, a genuine reason for an agent consists in the conjunction of a desire and a belief about how to satisfy that desire.[37] To desire a state of affairs and to believe that ϕing will bring about that state of affairs *is* to have a reason to ϕ. There is, on Joyce's reading of Hume, very little more that can be said, since Hume considers what an agent desires to be beyond rational appraisal, hence the famous quote that ''Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger.' (Hume 1739 T 2.3.3.6, SBN 415-416). Joyce (2001 chapter 3)

[37] It must be pointed out that this is not an uncontroversial reading of Hume, and it is not clear that Hume would have accepted it. See e.g. Sayre-McCord 2008 for an alternative view. Nonetheless, Joyce proceeds in this vein.

develops this Humean account into what he calls non-Humean instrumentalism via a critical engagement with Michael Smith's views, primarily those put forward in *The Moral Problem* (1994) and 'Internal Reasons' (1995a).

For Joyce, Hume's account is slightly impoverished in that it cannot explain certain common sense features of what it is to have a reason. For example, it suggests that when we have competing desires, the thing we have the 'real' reason to do is simply whatever will satisfy our strongest current desire. But this does not seem to be able to make sense of the competing reasons Molly has in the example given above. Yes, we could interpret Molly's case in this way, and say that in the end her choice boils down to a straight forward competition between her desire to eat the cake and her desire to watch her diet. But it seems a richer account, according to which we can in fact rationally appraise an agent's reasons, will describe the situation better.

To provide this richer account, Joyce draws on Smith, and makes a distinction between an agent's objective, subjective and irrational reasons. An agent's objective reasons are those which conduce to the satisfaction of the desires the agent would have if they were fully rational (see Smith 2010 p. 2). Smith considers being fully rational to mean deliberating flawlessly and in full knowledge of all the relevant facts.[38] Since Molly, were she in this epistemic position, would know about her impending sojourn on a desert island, she would acknowledge that she has a reason to eat the cake now. Yet it is not reasonable to expect Molly to have this level of knowledge, thus we would not call her irrational if she did not act upon this objective reason.

[38] Note, this is Smith's view of full rationality. Joyce's view differs crucially, but at this stage Joyce borrows Smith's terminology for the sake of clarity. These arguments are complicated, and it helps the reader to get a grip on one thing at a time before going into further subdivisions of the theory. The differences between Smith's and Joyce's views and how they affect the argument will be explained shortly.

Subjective reasons are those reasons which are based on those desires an agent has which would survive (or have survived) a process of deliberation. To put this another way, they are the reasons an agent takes on reflection to be their objective reasons, given their epistemically restricted viewpoint. In our example, Molly has thought about her various desires and appetites and decided upon a goal of watching her diet, and therefore refraining from eating cream cakes (presumably among other things). We might say that she has adopted this as an end or a project. This desire, coupled with a belief, which Molly thinks is justified, that eating the cake will frustrate her desire, constitutes a subjective reason to refrain from eating the cake. Having no knowledge of what tomorrow will bring, she takes this to be an objective reason.

Finally, Molly's desire (which I have called a temptation) to eat the cake is what Joyce calls an irrational desire, in that although she may satisfy a current desire by eating it, she would also be frustrating an end which she has adopted after significant deliberation. Therefore while her desire to eat cake, coupled with a belief that eating the cake will satisfy this desire, constitutes a reason for Molly, it is not a reason which would survive deliberation. In fact, acting upon it would frustrate the ends and desires which Molly has thought long and hard about. Thus Joyce calls it an irrational reason (albeit one which it would be fortunate for Molly if she acted upon it, given that she will soon become a castaway).

3.3.6. Summary thus far

So, how does this account of the various reasons an agent might have fit into Joyce's theory of normative reasons, and then into the wider argument for a moral error theory? Remember that Joyce's overall strategy is to build the most plausible analysis of moral normativity he can, and show that even this analysis leads to an error theory because we cannot make sense of it.

At the current point in this process, the aim is to build a picture of a kind of practical reasons which can fulfil the role of moral reasons in our thought and discourse. In order to underwrite the form of normativity to which traditional moral thought and discourse are non-negotiably committed, this must include a way to avoid the possibility of agents evading the authority of normative reasons by forestalling any 'but what does that mean to me?' response. The only apparent viable route to an account of reasons which can deliver this inescapable quality required of moral normativity was taken to be via understanding the reasons in question as requirements of practical rationality. Therefore the theory of normative reasons developed here must provide a way of discerning what it is practically rational for an agent to consider a reason for them to act.

In my reconstruction of Joyce's argument, the tripartite subjective/objective/irrational classification of reasons vitally serves this strategy because it establishes two main ideas. First, it establishes what Joyce considers an agent's subjective reasons to be. As we saw above, Joyce's view is that an agent is practically rational insofar as they are guided by their subjective reasons, and we now have a fuller definition of what 'practically rational' means for him.

In the context of the wider argument, this definition will therefore allow Joyce to specify how we are to understand requirements of practical rationality, i.e. the kinds of reasons which can forestall 'So what?' responses, and thereby be inescapable. If moral obligations are to entail inescapable reasons, the existence of which our moral discourse commits us to, we now know how Joyce thinks they must be formed, i.e. that they must function as subjective reasons and be appropriately based on agents' desires. If, on the other hand, moral reasons cannot be shown to function as subjective reasons do here, they cannot be requirements of practical rationality. This means they will be unable to forestall 'so what?' responses, and will therefore

fail to be inescapable. Given the commitments of our moral discourse, this will point towards a moral error theory.

Second, this three-way classification of reasons introduces the idea that an agent's authoritative practical reasons can encompass reasons which are grounded in the desires of a counterfactual, epistemically improved version of themselves. Namely, the definition of objective reasons given includes the desires an agent would have *if they were fully rational*. In simple terms, this means that we can have reasons to pursue what we would want if we thought about the matter properly, even if we have not actually thought about it properly ourselves. However this is deceptively simple. Much turns on how we interpret 'properly', and how demanding a view of full rationality we can take.

We will turn to this shortly, and investigate how different views of full rationality can affect what reasons we have, and how Joyce's view on this issue leads to H5. First, however, section 3.3.7 will briefly discuss another implication of the instrumentalist view described above, an implication Joyce rejects.

3.3.7. Universal reasons, part 1: Universal desires

Joyce's theory so far suggests that there may be an opportunity to rescue moral thought and discourse from an error theory by weakening the claim that morality entails or presupposes the existence of practical reasons which are authoritative for all agents, regardless of those agents' desires or ends. We saw in §2.1.3 that error theorists typically have no issue with the authority of hypothetical reasons. Consider, then, what could happen if there were 'a *desire* which all humans have (reliably, on all occasions), a desire from which no human could escape' (Joyce 2001 p. 61, emphasis original). If this desire – whatever it might be - were served by acting in apparently morally good ways such as refraining from murder and torture, keeping

promises and so on, then an error theory could be avoided. For universally authoritative moral reasons could be grounded by hypothetical imperatives based on this universally felt desire, with no reliance on the problematic kind of categorical reasons with which the likes of Joyce and Olson take issue.

The problem here is in specifying what the relevant desire might be. Joyce considers and rejects both self-interest and any sense of 'natural sympathy' with others. We can too easily imagine an agent (Joyce uses the example of Plato's character Gyges) who is sincere and secure in their community, part of a loving family, but who in their secret dealings with others outside their community is driven by selfish motivations to break promises, murder and so on, and who is unfazed by the suffering this causes.[39] So long as they can keep their skulduggery secret, it seems that hypothetical reasons to act morally can have no authority for such an agent. They will enjoy the material (and possibly, depending on their psychology, emotional) benefits of their antisocial behaviour without cost to the sense of community and fulfilment which their life at home affords. Thus agents' self-interest can be served without their having to avoid putatively immoral behaviour. And any sense of natural sympathy seems at best a very limited brake on agents' conduct. If this seems implausible, we need only to think back to historical examples of widespread slavery to find examples where even secrecy was not required to retain the love and respect of family and friends, and where natural sympathy clearly failed to ground any desires strong enough to reliably prevent the enslavement of other human beings. Indeed, I think there is an even more problematic case than those Joyce discusses – we cannot even depend on all agents reliably, on all occasions, having a desire to *keep living*. If even so basic a desire is not universal, then avoiding an error theory by invoking an inescapable desire shared by all agents at all times is surely a doomed endeavour.[40]

[39] For the original depiction of the character Gyges, see Plato 2008, book 2.

[40] Mark Schroeder attempts to deal with something like this problem on behalf of the Humean in his book *Slaves of the Passions* (2007). Since my task in this section is to explain error theory rather than

3.3.8. Universal reasons, part 2: The limits of 'full rationality'

Returning to my reconstruction of Joyce's argument, I said above that one way opponents might try to reject H5 and therefore avoid an error theory is to show that it could be possible to ground authoritative, categorical moral reasons in the requirements of practical rationality. This is sometimes called constitutivism, and has been influentially defended by Smith (e.g. 1994) and Korsgaard (e.g. 1999). While the details differ significantly, constitutivists share the idea that rational agency, when defined in a particular way which they describe, can give rise to authoritative practical reasons for all fully rational agents. Thus if all fully rational agents as such have authoritative reasons to behave morally, a moral error theory can be avoided.

To see how this works, we begin with practical rationality itself. It is common to begin with a relatively undemanding view of practical rationality. Not least, this is because defending a comparatively simple, undemanding view is easier and often more successful than defending a more demanding view when the simpler view would have sufficed to explain the phenomenon at hand. Thus standard views of practical rationality include that a practically rational agent must be instrumentally rational (i.e. can understand a link between their actions and their outcomes) and must be responsive to reasons which they take to be authoritative for them. A practically rational agent must also be capable, through deliberation, of arriving at and intending to act upon a minimally coherent set of authoritative reasons – for example they must be able to understand that they cannot routinely act so as to bring about both *p* and *not-p* by the same action.[41] A hugely influential view of practical

defend it, I will avoid digression into discussing Schroeder's arguments. But for a detailed discussion, see e.g. Enoch (2011b).

[41] Failure to grasp this principle is sometimes colloquially referred to as 'cakeism', referring to the saying that one cannot have one's cake and eat it.

rationality along these lines was famously set out by Bernard Williams in 'Internal and External Reasons' (1981).[42]

The problem with this undemanding view of rationality in the current context is also highlighted in Williams' paper – it cannot be used to show that all agents share specific practical reasons, and so cannot underwrite the kinds of reasons to which morality is non-negotiably committed. As we saw above, Joyce rejects the suggestion that there might be desires or motivations which are shared by all agents. And Williams sums up the implications of this quite bluntly, 'Can we define a notion of rationality where the action rational for A is in no way relative to A's existing motivations? No.' (1981 p. 112).

The solution for constitutivists is to defend a richer view of practical rationality which *can* yield authoritative practical reasons for all agents and thus underwrite moral normativity.[43] Korsgaard bases her view on Kant's Categorical Imperative, and argues that for an agent to *be an agent* and thus genuinely to act, that agent must be able to will that there were a universalisable law that all agents must act as the agent intends to act. Korsgaard then hopes that all agents who constitute themselves as agents by having a will of this form will pursue the same kinds of goals as a result, and thus demonstrate that practical reasons which are grounded in rational agency itself can be universally authoritative.

In a similar fashion, Smith introduces the notion of 'fully rational' (i.e. fully informed, flawlessly deliberating and having no relevant false beliefs) counterparts of current agents (1994 §5.9). Consider an agent, Anne, and her fully rational counterpart, Anne^{+}. Smith argues it is part of the concept of *having a reason* that in order for Anne to have a reason to ϕ, it must be the

[42] I will return to Williams in significantly more detail in §5.2.3.

[43] More specifically, constitutivists may defend views of what actions are, or what agency is, or other similar related notions. But in the present context, these are all sufficiently closely related to practical rationality that I am content to gloss over the differences in order to focus on the general idea which underlies the various specific formulations.

case that Anne^{+} would want Anne to ϕ. But more than this, it must also be the case that all fully rational agents would want Anne to ϕ – the desires of all fully rational agents about what current agents are to do must converge (*ibid.* p. 166-167). So long as the desires of all fully rational agents would converge in this way, this then allows Smith to argue for the existence of authoritative universal practical reasons which escape an error theory.[44]

These are of course very sophisticated, detailed views which it would be inappropriate to delve into in great detail here.[45] But they face a common objection highlighted by Enoch (2006) which cuts across the detail. We already saw in §3.3.5 that Joyce rejects Smith's argument that agents must be fully informed to be fully rational. But Enoch argues more generally (2006 §3) that all constitutivist views fail to provide the sense of authority for practical reasons which could ground the kind of categorical reasons morality requires. This is because whatever constitutivists add to an undemanding view of practical rationality in order to arrive at their richer view, we can essentially respond in a now familiar vein: 'so what?'.

In Korsgaard's case, Enoch argues that we can reply that agency and action may well require all sorts of things including universalizability and so on, but that if so, we are quite content with failing to act or to be agents. We can go on our merry way, failing to will that our actions should be universalised and so on, with no apparent meaningful penalty. Likewise, as I discussed in §3.3.5, we can apparently be perfectly practically rational, yet still reply to Smith 'fine, an idealised version of me, with information I do not have, may want me to do x or y.

[44] Initially Smith was not necessarily committed to the claim that the desires of all fully rational agents *actually would* converge (see e.g. Smith 1994 p. 173-174), and has admitted that he was tempted to embrace an error theory (Smith 2010). But in more recent work (starting approximately with the aforementioned 2010 article), he has moved to arguing that the required convergence would emerge. Smith's later view remains controversial (see e.g. Smith 2015 and Bukoski's 2016 critique thereof), and since my task is to explain rather than defend error theory, I omit further discussion of the topic here.
[45] For a more detailed discussion of constitutivism in general, see Katsafanas 2018.

But I don't have that information, and it makes most sense to me to do z'. Enoch's point is that whatever 'enrichment' is proposed by constitutivists in their search for the requisite kind of normativity, that enrichment itself will always undermine the authority of the reasons which the view in question can yield.

Drawing together these last two subsections, then, we can see that premise H5 in my reconstruction of Joyce's argument is credible - practical rationality cannot give rise to the same reasons for all agents. And even if it can, then we should be suspect about the authority of those reasons. Since the strategy here was to seek a way of grounding practical reasons which are authoritative for all agents, the strategy fails.

3.4. Conclusion

I said earlier that that Joyce's overall strategy is to build the most plausible analysis of moral normativity he can, and show that an error theory is unavoidable because we cannot make sense of moral normativity *even on this analysis*. Drawing the various subsections of §3.3 together, it seems that Joyce is arguably successful in this. In the wider context, this means that Joyce succeeds where I argued that Olson fails, in providing a theory of practical reasons which results in a moral error theory. In doing so, Joyce provides what is often regarded as the strongest argument for a moral error theory defended to date.

I do not believe that this description of Joyce's view would be controversial among most commentators. But for the avoidance of doubt, some evidence to back up the claim: as well as the explanation I have provided here of why Joyce's arguments succeed where others fail, at the time of writing, a Google Scholar search yields over 900 citations for *The Myth of Morality*. Simon Robertson argues that error theorists in general are best advised to take an approach along Joyce's lines, and he explicitly sees the aim of his article 'How to be an Error

Theorist' as providing additional justification for Joyce's approach (2008 footnote 22). Joyce is taken alongside Mackie as a paradigmatic error theorist by Stephen Finlay (2008), and Russ Shafer-Landau describes Joyce's view as 'surely the most elegantly written, comprehensive and well-argued defense of a moral error theory yet to appear' (2005 p. 108).

In what follows, then, when I discuss what we should do if we accept a moral error theory, I will have roughly Joyce's view in mind unless indicated.[46] Specifically, this will typically mean an error theory based on the view that non-institutional categorical practical reasons (such as the reasons implied by moral obligations) cannot be authoritative for agents. I will now move on in the next chapter to argue that if we accept the truth of a moral error theory, this confronts us with an urgent problem which demands a response.

[46] It is vital to present Joyce's argument here, as it frames important parts of the debates that will follow. But it should be noted that even if we are not convinced by Joyce's argument, the question of what we should do if some other version of a moral error theory turned out to be true would still be independently interesting.

Chapter 4 – What Now?

4.1. Why the 'what now?' problem arises, and how we might respond to it

The impact of accepting a moral error theory should not be underestimated. Moral thought, discourse and behaviour are so fundamentally intertwined with human society that many people would find it difficult to accept that there are no positive moral facts, much less know what to do in their absence. Our moral practices are so important to us that one may even doubt that we could get by without them without descending into some kind of murderous, debauched chaos. After all, as I pointed out at the start of chapter 2, the committed moral error theorist must agree that not only is there no moral reason to refrain from breaking promises or to pay taxes, but also that there is nothing morally wrong with rape, that we have no moral grounds on which to criticise the state-ordered crucifixion of teenagers in Saudi Arabia, and conversely that there is nothing morally good about working to prevent even one's own children dying of starvation.

Moreover, morality plausibly plays an important role in our wellbeing by contributing towards a feeling that we are doing the right thing and living good lives. And we frequently manifest powerful reactive attitudes such as violent resentment or profound gratitude which revolve around the moral relationships we have with others.[47] Moral considerations inarguably

[47] See Strawson's classic paper on relationships and reactive attitudes 'Freedom and Resentment' (1962). Strawson argues that It may indeed be psychologically (or even phenomenologically) impossible for us to abandon what he calls 'reactive attitudes' or 'participant attitudes', at least for more than just short periods of time. Such attitudes would, according to Strawson, include much of what we might typically think comes under the heading 'moral considerations'. See Strawson 1962, particularly §4. Strawson's paper was itself recently the subject of a volume of critical discussion to mark its 50th anniversary, Shoemaker & Tognazzini 2015.

colour and inform the way virtually everyone acts and understands their relationship with the wider world.

This creates two major problems for error theorists. The first is that their arguments will often be met with fierce opposition or even dismissed purely because of the conclusion they support. This is highly emotive territory. Therefore, error theorists must be exceedingly careful to build their arguments from firm, believable foundations and to state their case with clarity and thoroughness. The foregoing chapters should have given the reader an appreciation that contemporary error theorists are indeed painstaking about these things, and this is partly why I devoted as much space to explaining error theory as I did. Recall that I began by discussing some very basic features of moral discourse and only after that did I build towards presenting some quite detailed, more difficult material. When arguing for a conclusion which will almost inevitably be difficult for others to accept, or when presenting such arguments in a credible light as I have tried to do, one must attempt to make an especially convincing case.[48]

The second major problem is the one I will turn to now, and with which the remainder of this project will be concerned: if we accept a moral error theory, what should we do next?[49] I will begin by arguing that a problem confronts error theorists which defenders of most other metaethical theories do not obviously face. I call this problem the 'what now?' problem (for economy I will frequently abbreviate this to WNP). The philosophers involved in the WNP debate clearly agree that if a moral error theory is true, we need to think about what to do next. Otherwise, there would be no WNP debate about what we should do post-error-theory.

[48] It has even been argued that the error theory is impossible to believe, but that that does not stop it being true – see Streumer 2013a & 2017 chapter 9.

[49] Various formulations of this question will be summed up in what follows by the phrase 'what now?'. These two words should be read as encompassing everything from their comparatively polite, literal meaning to the sort of crushing, awe-stricken despair which afflicts Nietzsche's madman in *The Gay Science* (2001 §125).

Yet in the literature to date there has been little in the way of explicit articulation of the problem itself, and quite how pressing a problem it is for error theorists. Sections 4.1.1-4.1.3 will move towards plugging this hole. In my view, there are three key questions about the WNP, each of which I will seek to answer. First, how does the problem arise? Second, what is the 'what now?' question really asking? And third, how are we to evaluate responses to the problem – what are the rules of the game? Armed with answers to these questions, we will be much better placed to assess responses to the WNP - both others' responses and my own.

I will then move on to examine the main responses to the problem offered by others to date. Broadly speaking there are four main kinds of response which have been defended: abolitionism, conservationism, revolutionary fictionalism and revolutionary expressivism. I believe that all of these responses to the WNP are inadequate for various reasons. In sections 4.2-4.5 I will discuss each in turn, and in each case I will argue that the position described is fatally problematic. Section 4.6 will then conclude this chapter by highlighting the need for a new response to the WNP which does not face the same fatal objections which beset the other responses. I will then go on in the next chapter to introduce my own, new response to the 'what now?' problem, revolutionary relativism. It is important to remember that throughout the rest of this thesis, I will assume the truth of a moral error theory, i.e. that there are no moral facts or properties, nothing we morally ought or ought not to do, and so on.[50]

[50] Mackie may seem to be curiously absent in what follows. This is not because he did not offer views on what might follow after accepting error theory. Quite the contrary – a large part of *Ethics* discusses just that. Rather, it is because there is some confusion as to what his position was. He seemed to consider himself broadly a conservationist in *Ethics*, whereas later he spoke in terms of a 'moral overlay', which could be thought to suggest fictionalism. It has also been suggested that he was a substitutionist (Jaquet & Naar draw attention to remarks from 1976 which might support this, 2016 footnote 7), and that he became an abolitionist before his death (see Hinckfuss 1987 chapter 1, footnote 24 & Garner 2007 p. 501*ff.*). To add to the confusion, Nolan *et al.* read Blackburn as claiming that Mackie was an abolitionist, despite using a quote from Blackburn which rather suggests to me that Blackburn thought Mackie wanted to preserve morality in some modified form (Nolan *et al.* 2005 footnote 5). This confusion over Mackie's true position means it is more straightforward to largely omit him here. There is plenty of material on the topic from other philosophers for present purposes.

4.1.1. Why the WNP arises

The WNP arises from error theory in two ways. The first, which appears commonly in the relevant literature, grows out of an intuitive sense that moral beliefs are *useful* to us, that they help regulate our conduct in beneficial ways by acting as constraints on our more malevolent or short-sightedly self-interested impulses. If, for example, there is a widespread, deeply rooted belief which tells us not to steal, then most people would find it intuitively plausible that less stealing will occur. Alternatively, mutually beneficial cooperation may seem more likely (or is even made possible) if we have moral beliefs which tell us that we ought to keep our promises and uphold agreements.[51]

Yet if we accept a moral error theory, we thereby take ourselves to have established that the moral beliefs which may have previously guided our behaviour cannot be true. And it is often thought that we tend to give up beliefs which we know to be false.[52] Yet without moral beliefs, one may well worry that there will be insufficient non-moral reasons to regulate our behaviour in these beneficial ways. And if our behaviour is not so regulated, the worry goes, human society could descend into chaos. This powerful worry places a perceived burden on error theorists to show why, even if their theory is true, it does not come at a catastrophic cost. In contrast with other metaethical views, if we accept a moral error theory, it seems we

[51] This echoes David Gauthier's contractarianism, according to which, even if we are motivated principally or exclusively by self-interest, we will be better off if we treat each other morally. This is because individuals who cooperate can achieve more beneficial outcomes than individuals acting alone, and therefore we can realise much greater aggregate benefits if we agree to adhere to moral norms such as keeping our word, refraining from predating upon one another and so on, despite the fact that this might curb what we would otherwise want to do out of short-term selfish motives. See e.g. Gauthier 1987. I will return to Gauthier in more detail in chapter 7.

[52] There will be more to say about whether or not we must jettison false beliefs in what follows, both as we progress through this chapter and later in chapter 6. But for now, I think it plausible that we do have this intuition.

are in danger of losing something hugely important. This is one reason why error theorists appear to face a WNP when other metaethicists do not.[53]

Bolstering this point is the fact that it will be significantly to error theorists' advantage if they can show us a way forward whereby at least some significant subset of the apparent benefits of traditional morality can be retained. Matt Lutz even suggests (2014) that it is part of the role of WNP responses to make error theory more palatable to those who might doubt its veracity because of their distaste for its conclusions (e.g. the positions on rape and starving children I just mentioned) by seeking to convince opponents that some of the potential negative consequences of error theory can be avoided if we respond to the WNP in the right way. I take a slightly less accommodating line in that I believe that, if the arguments for error theory are sound, whether their conclusions are unpalatable is irrelevant. The truth is sometimes ugly. Having said that though, even if undermining opposition to error theory is not what WNP responses are *for*, it might nonetheless be an effect they have which is favourable for error theorists. Thus it behoves error theorists to have something substantial to say about the WNP even if only as a way of reducing the potential psychological barriers to accepting their theory. And for those who accept error theory, responding adequately to the WNP may show that we can retain important benefits which might have been thought lost.

The second way the WNP arises has not, as far as I am aware, been noted in the literature, but to my mind makes finding a solution even more pressing. It is that we cannot escape the WNP. If we accept the error theory, we *will* do *something* next. It does not matter whether we abandon moral thought and discourse, or try to replace them with something else, or even

[53] An argument might be made that other metaethical theories do face a less obvious WNP, but that they respond to it so immediately that it goes unnoticed. For example, non-cognitivists argue that moral judgements are not beliefs. If they stopped there, the obvious question would be 'well what are they, then?' – a close variation of the 'what now?' question discussed here. It is only because non-cognitivists automatically to go on to tell us what they think moral judgements *are* that a WNP does not obviously arise for non-cognitivism.

try to forget about the error theory altogether and carry on as we were before – each of these alternatives constitutes a response to the WNP even if we do not realise the problem exists. It is therefore incumbent upon metaethicists in general, and particularly upon error theorists, to seriously consider what the available responses to the WNP are, and which might be the best among them.

It is part of the core activity of philosophy to make sense of the world. So if we are to be abolitionists and abandon so important (and so elaborately contested) a field of thought and discourse as morality, I would argue that we ought to know why we are doing so. A case must be made, to show *why* it is the best thing we could do out of the available options. And if we are not to abandon morality, or if we are to seek to retain some key aspects of it while jettisoning the rest, then we need to truly understand what it is we are doing in those cases too. Whatever we do in response to the WNP, we need to know what our options are and whether what we are doing is the best thing we could do under the circumstances – whether it brings about the greatest benefits and generally promotes human flourishing more than the alternatives. Otherwise we fail as philosophers.

4.1.2. What the 'what now?' in the WNP is really asking

My second question was about what is really being asked in the question 'what should we do if we accept a moral error theory?'. There are two key things to mention here. One is that the 'should' here cannot be read as a moral 'should'. To ask a moral error theorist what we should morally do if error theory is true is to negate the very grounds for asking the question. Beyond this constraint, however, there is no onus on those who may pose the WNP to specify what sense of 'should' is intended – at this stage it can be left vague. Rather, the onus is on those who wish to respond to the WNP to specify the sense of 'should' on which their response relies (which I will discuss in §4.1.3.).

The second key thing to consider is to whom responses to the WNP should be addressed. Is it a question about what metaethicists should do, or perhaps society more generally? Again, error theory is different from most other metaethical theories in this respect. Traditional, descriptive metaethical theories and the resultant arguments between metaethicists need not necessarily concern society in general very much. Metaethicists focus on analysing a commonplace phenomenon (i.e. morality), but for the most part, the rest of society simply gets on with its moral affairs. 'The folk' do not need to know about sophisticated theories of action or normativity to be competent users of moral terms and to legitimately consider moral judgements an important part of their lives. Consider an analogy: two biologists happen upon an egg while walking in a park. One says 'This egg is merely a vessel for the bird which will one day hatch from it, it is the beginning of an existence, a promise which has yet to be fulfilled'. The other shakes her head and replies 'No, not at all. This egg is the culmination of a fascinating process of fertilisation, a complex construct of albumen, calcite and so on. It is the *raison d'être* of the bird which laid it'. While they argue, the egg continues to be simply an egg. And to a passer-by, none of this debate necessarily affects whether they think of eggs as things to be incubated and hatched, or as breakfast. In fact, to make the analogy more realistic we might add that although they can see the egg, the passer-by hasn't noticed the biologists and can't hear the debate.

Error theory is different. When a claim is made about what we should do in response to the WNP, the 'we' in question cannot be limited to metaethicists. The upshot of error theory is not that moral judgements should be analysed as *x* or *y* kind of attitude, but that *there can be no true moral judgements*. The phenomenon being examined has been destroyed in the course of the examination, one might say. Thus, if error theory is true, then that fact (and what we should do about it) concerns anyone and everyone who ever makes a moral judgement. Of course, the community of metaethicists involved in these debates is not all

that large, but the material discussed in those debates is of direct relevance to the everyday lives of virtually all human beings. The 'we' in 'what should we do?' potentially includes everyone, whether they are metaethicists, whether they are error theorists, or whether they are unaware of anything discussed in the course of this project.[54]

4.1.3. The rules of the game for WNP responses

My final question was about how we should evaluate responses to the WNP. Several responses have been proposed, and of course my overall aim here is to provide a new response. So, what are the rules of the game in this endeavour, and how are we to tell a good response from a bad one? I will highlight several issues which must be borne in mind as we go on.

i) Normative Circularity Constraint

Perhaps the most obvious issue to highlight is the counterpart of the 'no moral 'shoulds'' constraint I mentioned above. Just as the question 'what should we do?' may not invoke a moral sense of should, neither may the responses to the WNP rely on any sense of 'should' about which we accept an error theory. Most glaringly, this includes any moral sense of 'should'. For it would be inconsistent for moral error theorists to respond to the WNP in terms of what we should *morally* do. So the 'what should we do now?' question can be neither asked nor answered in a moral sense. But we must also bear in mind the commitments of error theorists *en route* to their error theories. For example, Joyce relies on the incoherence

[54] This assumes that having true beliefs is something which matters to or is in the interests of most or all people. While I consider this intuition respectable at this point in the discussion, it may run contrary to arguments for conservationism and for a niche view known as propagandism (see e.g. Cuneo & Christy 2011 & Joyce 2005 §5). I will have much more to say about conservationism in §4.2 and in subsequent chapters. And I will touch upon propagandism in §4.5, though I will not engage too extensively with it, for reasons I will give in that section.

of not just moral norms, but of the broader category of non-institutional categorical norms in making his case for a moral error theory. Therefore he cannot consistently give an answer to the WNP which relies on *any* kind of non-institutional categorical sense of 'should' (and, since his way of arguing for error theory is quite influential, neither can a lot of other error theorists). Thus we can establish a general constraint on all WNP responses:

> **Normative Circularity Constraint**: No response to the WNP may rely upon any form of normativity about which the philosopher in question claims an error theory holds, or which contradicts any commitment of the arguments for said error theory.[55]

This may not be too much of a problem for typical moral error theorists, since as we saw in §2.1.3, they have no issue with instrumental norms which predicate authoritative reasons for action based on securing things which we desire. Thus arguing that we should adopt a given WNP response because it will deliver benefits which we plausibly desire - for example, that adopting it would promote mutually beneficial cooperation among agents, when the alternative is murderous chaos - is not problematic for typical moral error theorists. Whereas arguing that we should adopt a WNP response because that would be the moral thing to do would be self-undermining for moral error theorists.

This may be more subtle than it initially appears. For example a WNP response based on desire satisfaction may end up recommending that we adopt a given strategy on the basis that it will bring about *the greatest degree of desire satisfaction for the greatest number of people*. This would be very close to a form of utilitarianism – an ethical theory – and will therefore require careful argumentation in order to avoid violating the normative circularity constraint. Similarly, there are those who defend error theories which extend beyond morality to all

[55] Later on in chapter 5, I will further refine and subdivide this constraint (see §5.2). But for the time being, the NCC captures the relevant idea without complicating matters prematurely.

normative judgements and discourse, for example Bart Streumer (2017). I am interested in specifically moral error theorists and so will not dwell on this here, but for obvious reasons it will be particularly tricky for error theorists about all normative judgements to argue that we *should* do anything at all in response to the WNP. [56]

ii) Previously dismissed views

Next, although it may seem to go without saying by this point, it bears repeating that responses to the WNP occur only in a post-error theory context. This might be thought to give some extra leeway to WNP responses, as opposed to views about the metaethics of conventional morality. The theories here no longer need to describe conventional moral thought and discourse. In the current context, that job has already been done, and the matter has been decided – in the WNP context, error theory is true. The task now is to consider what attitudes we can and should have towards matters about which we previously held traditional moral beliefs. Therefore many of the theoretical options ruled out by error theory – for example expressivism, rejected as a descriptive theory by Mackie for lacking a way to account for the apparent normativity of moral facts and properties – are back on the table. For, while error theorists might consider them inadequate as descriptions of conventional morality, it is open to philosophers operating in the WNP context to show that such previously rejected theories might nonetheless preserve certain benefits of morality, and therefore be worth adopting as we move into a post-error theory world.

[56] See §4.5, footnote 108 of this thesis for an example of a WNP response which seeks to do this and, I argue, falls foul of the NCC despite efforts to avoid doing so.

iii) Plausible implementation

The above does not mean, however, that those who respond to the WNP can say anything they like, so long as they can show that their view might provide certain benefits. Post-error theory recommendations must be strategies which we could actually adopt in our everyday activities. Bluntly put, recommendations which cannot consistently or plausibly be acted upon cannot help us solve the WNP. Therefore responses to the WNP must be consistent with the commitments of error theory, and must be coherent theories which plausibly and directly bring about (or at least substantively contribute to) the benefits claimed. This is why I again take a slightly firmer view than might seem to be implied by Lutz, for example. He can be read (2014) as suggesting that a cost/benefit analysis is the primary way to assess WNP responses.[57] I take a more stringent line, examining each response for inconsistency and questioning whether it could realistically be implemented before reaching any kind of cost/benefit analysis. Should a response fail either of these tests, it will be ruled out entirely.

For example, if a given WNP response suggests that we should adopt a view based on the fact that it would be the morally right thing to do, it will thereby violate the Normative Circularity Constraint (NCC) discussed above. It will therefore be rejected as inconsistent with error theory and/or the commitments of the arguments for error theory. If the response satisfies the NCC, but suggests that we should do something which we are not capable of doing – as a flippant example, let's say the response is that we should grow wings and fly around because doing so will bring about a more cohesive society, and we desire societal cohesion - it will be rejected as being incapable of plausibly being implemented.

[57] I do not mean to claim that Lutz *really is* willing to ignore inconsistency or implausibility when assessing WNP responses. Rather I wish to underline that my approach is somewhat more demanding than simply asking 'which theory gets us the most benefits?'.

iv) Delivery of claimed benefits

When it comes to considering the claimed benefits of adopting a given WNP response, there are a number of things to bear in mind. At this preliminary stage I do not wish to lay out a comprehensive list, as it serves both my overall argument and the reader better to let those who seek to respond to the WNP have their say, and then to judge the arguments on their own merits. But even from the outset, we can anticipate certain issues around the claimed benefits of each response. Most obviously, there is the question of whether the proposed response actually would deliver (or at least substantially contribute to delivering) the claimed benefits. To return to the exaggerated example I just gave, it is not obvious that flying around would help deliver a more cohesive society, even if we could do it. Responses to the WNP must show that there is a plausible connection between adopting their recommendations and realising the benefits claimed. Otherwise there may be no point adopting them.

v) Costs and benefits

Additionally, we will need to know *for whom* the claimed benefits will accrue. If everyone in society will benefit, then that is all well and good. But things are seldom so clear cut, and where there may be disparities, we will need to be aware of them if we are to make properly informed choices. Relatedly, we might consider the subject of the 'should' in any suggestion that 'we should ϕ'. I mentioned above that WNP responses must be addressed (or at least addressable) to all of us in a broad sense, but should they be interpreted as advice that we each follow individually, or as collective advice for groups of people – i.e. is it a matter of each individual ϕing, or does the proposal require that groups ϕ *en masse*? In each of these cases, there may be varying consequences, including potential effects on whether the suggestion in question can plausibly be implemented.

We will also need to consider not only the implementation of the strategy in question, but also whether we need to worry about the inculcation costs involved – if a strategy would bring about certain benefits if implemented, but the costs of doing so in terms of education and so on would be enormous, perhaps we would be left with a negative aggregate benefit.[58]

vi) Is morality beneficial?

Finally, it must be noted that we should not assume that traditional morality delivers unqualified, universal benefits. As we will see, there is at least one way of responding to the WNP which strenuously argues against the claim that conventional morality is beneficial at all. And as reflected by the NCC above, in the WNP context, morality itself cannot be thought to be intrinsically good, and therefore to provide us with categorical reasons to include it (or something closely resembling it) in our lives. It is therefore up to each philosopher to argue for the specific benefits (or costs) they think traditional morality provides, and then to argue why their chosen WNP response delivers those benefits (or avoids those costs). Any responses to the WNP which suggest retaining moral thought and discourse (or anything closely resembling them) must show us *why* we should retain them. Moral thought and discourse (or anything closely resembling them) must earn their place in our lives.

To draw all of this together, then, the 'what now?' problem is real, and confronts error theorists in a way in which it does not obviously confront defenders of other metaethical theories. Responses to the WNP, while of primary interest to metaethicists, must be addressed (or at least addressable) to the whole of society. And good responses must have certain features – *inter alia*, they must be consistent with the commitments of error theorists,

[58] Brad Hooker discusses similar issues using the term 'internalization costs', see Hooker 2000, especially chapter 3.

they must make a case for the benefits they claim to secure rather than just assume them, and they must be strategies which can be implemented in real life situations.

To date, there have been four principal varieties of WNP response, which I will call conservationism, abolitionism, revolutionary fictionalism and revolutionary expressivism.[59] In §4.2-4.5, I will explain each of them, and show why they are inadequate. I will also touch in §4.5 on a couple of other more niche WNP responses which have appeared in the literature, and explain why I will not be engaging further with them. I will then conclude this chapter by arguing that a new WNP response is required, and by drawing out the lessons which can be learned from the failure of the responses discussed. I will then move on in the next chapter to begin to present and defend my own response.

4.2 Conservationism

Having accepted a moral error theory, and thus taking ourselves to have established that there are no moral values, we might find ourselves worried that there will be insufficient non-moral reasons to regulate our behaviour. As I wrote above, we might fear that in the absence of moral rules, human society will descend into chaos. Without a moral injunction against murder, won't people be much more likely to become murderers? At the heart of this worry is a sense that moral rules are *useful* to us, that they help regulate our conduct in beneficial ways by acting as constraints on our more malevolent or short-sightedly self-interested impulses.[60] This is intuitively quite plausible for many people. If, for example, there is a

[59] It is common to use terms such as 'revolutionary' to describe the post-error-theory variants of fictionalism and expressivism, in order to distinguish them from their hermeneutic cousins. While I will sometimes include the 'revolutionary' in what follows, the firmly post-error-theory WNP context here means that wherever I use the terms fictionalism and expressivism without qualification, the revolutionary variant is the one I have in mind unless specified otherwise.

[60] This view of morality's usefulness is certainly well known to error theorists from Mackie onwards (see e.g. Mackie 1977 p. 43: 'We need morality to regulate interpersonal relations, to control some of the ways in which people behave towards one another, often in opposition to contrary inclinations') but is also found outside error theory. For example Gauthier's influential view in *Morals by Agreement* (1987)

widespread, deeply rooted social institution (e.g. a moral rule) which tells people not to steal, then most people would be willing to believe that less stealing will occur. One might also think that keeping promises and sticking to agreements at least most of the time are practices which make mutually beneficial cooperation possible in the first place, and so having widespread moral beliefs that everyone ought to do so is likely to be very useful indeed. This usefulness is the motivation for conservationism – conservationists' motivation is fundamentally prudential. Pigden claims (2007 p. 445) that he is a conservationist, and that Mackie was one as well (though see footnote 50, above). But the most developed account of conservationism to date comes from Olson (most recently, 2014 ch.9), and therefore Olson's account will be my focus here.

Olson argues that moral beliefs are of (at least net) benefit to society, and that the most straight-forward way to access the benefits of having moral beliefs after we accept a moral error theory is by continuing to have moral beliefs, even if they are false. As he puts it, 'I argue that moral error theorists are [best] advised to recommend what I call *conservationism*, i.e. preservation of ordinary (faulty) moral thought and discourse' (2014 p. 178, emphasis original). As I understand it, Olson's argument (2014 §9.3) can be summarised as follows (OC stands for Olsonian conservationism):

> OC1. Having moral beliefs (i.e. beliefs which entail or presuppose that there are moral properties and facts) at least most of the time is of net benefit to society.
>
> OC2. We can form and/or maintain beliefs not only for evidential reasons, but also for prudential reasons which are independent of any evidence we might have.

bases morality on the constraint of self-interested individuals' behaviour for mutual benefit. I will return to Gauthier's view in §7.4.1.

OC3. For a proposition, *p*, we can have an occurrent belief that *p* in one context, while simultaneously being disposed to believe that *not-p* in other contexts.

OC4. Therefore, if we accept error theory, we should on prudential grounds choose to have moral beliefs in most contexts (i.e. in everyday life), while simultaneously remaining disposed to believe that all positive moral beliefs are false in other contexts (i.e. in the philosophy seminar room).

Many people will find OC1 plausible, as I discussed above. Crucially, abolitionists will not, but given that it seems quite intuitive, and that numerous philosophers agree with it, we can grant it for present purposes.[61] As support for OC2, Olson discusses several parallels, the most informative of which is with Pascal's Wager.[62] Olson writes, 'As is well known, Pascal argued on prudential grounds that we ought to believe in God, even though there is insufficient evidence that God exists' (2014 p. 191). Those prudential grounds can be expressed as a series of conditionals: If God exists and we are observant theists, then we stand to benefit massively by being admitted to heaven. If God does not exist and we are observant theists, we will have lost comparatively little due to our mistaken theism. On the other hand, if God exists and we are atheists, then we stand to incur a catastrophic cost by being condemned to hell. While if God does not exist and we are atheists, we will have gained comparatively little from our atheism. The standard reading of Pascal's Wager is that when we consider these counterfactuals, we will conclude that it is most prudent to be observant theists, since that is the option with the greatest potential rewards and the least downsides.

The idea that we can alter our beliefs in this way, known as doxastic voluntarism, is not as simple as merely choosing what to believe as if we were choosing what clothes to wear.

[61] As well as himself, Olson cites Warnock, Mackie, Joyce, Nolan, Restall & West (2014 p. 180).
[62] For a fuller discussion of Pascal's Wager than space permits here, see Hájek 2018.

Rather, it is a matter of orthopraxy; according to the standard reading of Pascal, if we do the right things – go to church regularly, pray to God, think pious thoughts, and so on – genuine theistic belief will follow. Olson argues that this analogises to the moral case, taking the view that we can bring ourselves to have genuine moral beliefs through developing habits such as making moral assertions and thinking 'moralized thoughts' (2014 p. 191-192).

However, even if we are persuaded by doxastic voluntarism, Olsonian conservationism also requires OC3. Olson argues that the phenomenon OC3 describes is entirely possible, and calls it 'moral compartmentalization' (2014 p. 192). It is worth quoting him at moderate length, as I will discuss several features of the argument as we progress.

> In general, it does not seem impossible simultaneously to have an occurrent belief that *p* and a disposition to believe that *not-p* in certain contexts. Indeed, we can go further and maintain that it is a psychologically familiar fact that we sometimes temporarily believe things we, in more reflective and detached contexts, are disposed to disbelieve. In such cases, the more reflective beliefs are suppressed or not attended to. [...] For instance, someone might say truly the following about a cunning politician: "I knew she was lying, but hearing her speech and the audience's reactions, I really believed what she said." [...] Hence we are sometimes *taken in* by what people say [...] in the sense that we believe what is said, even though we are disposed to believe, upon detached and critical reflection, that it is false. (2014 p. 192-193, emphasis original).

From this, we can see that Olson is not claiming that we ought to have simultaneous, mutually contradictory beliefs (which would threaten to make us irrational). Rather, what Olson is suggesting is what I have summarised above as OC3, that we have mutually contradictory

beliefs *at different times*. And given OC2, Olson seems to think that whether those beliefs are true or false makes no difference to whether or not we can come to have them if we make an appropriate effort.

4.2.1. Criticisms of conservationism

If OC1 - OC3 are true, OC4 seems to follow without issue, so Olson's argument appears to be valid. Unfortunately for Olson, all three of the premises OC1-3 are questionable. Having granted OC1 for the time being for argument's sake, I will take OC2 and OC3 in turn. I do not believe either of these premises is true. Olson's arguments fail to support the controversial account of beliefs on which his conservationist position depends. Therefore Olsonian conservationism is not a viable response to the 'what now?' problem. As we will see, OC2 faces a potentially fatal objection raised by Jussi Suikkanen. And in my view OC3 fares little better, as I will explain.

The main problem with premise OC2 is that Olson's use of Pascal's wager to show that it is possible to have beliefs irrespective of evidence is misleading. One may have misgivings over whether the attitudes produced by the wager are really beliefs at all.[63] But even if we grant that they are beliefs, Olson's analogy only goes so far. According to Olson's reading, Pascal's argument encourages us to 'go the extra mile' when we have insufficient evidence to reach a conclusion and voluntarily adopt beliefs despite the lack of evidence. But remember that accepting the error theory entails that we take there to be sufficient evidence – in this case, in the form of detailed arguments – to conclude that all positive moral beliefs are false. Therefore what Olson seemingly asks is that we ignore the evidence and choose our conclusion in spite of it – that we have moral beliefs *which we judge to be false*. This is a

[63] See Suikkanen 2013, footnote 17.

different matter than having beliefs predicated on insufficient evidence; it is a matter of having beliefs which are *insensitive* to evidence.

In the context of WNP responses, Suikkanen persuasively argues (2013) that the best available accounts of belief support the view that beliefs must be sensitive to evidence.[64] And as is the case with moral error theory and moral beliefs, where we accept that there is a systematic problem with a whole category of beliefs which we take to be sufficient evidence to conclusively contradict all positive beliefs of that type, we have no choice but to give up those beliefs, or else be irrational.[65]

This is compounded in Olson's case because his position as an error theorist is predicated on cognitivism – i.e. the view that one of the most fundamental functions of moral discourse is to express moral beliefs.[66] Thus, if it is definitive of beliefs that they are sensitive to evidence – and the best available accounts of beliefs tell us that it is – and the attitudes which we have towards moral propositions are not sensitive to evidence, then they cannot be beliefs. Therefore cognitivism is wrong, since moral discourse does not serve to express moral beliefs (Suikkanen 2013 pp. 177-178). As a consequence, Olson must provide a theory of beliefs which is more plausible than the most plausible accounts we currently have, and which can account for beliefs being insensitive to evidence. If he cannot, then a) the argument is invalid since OC2 is false, and b) Olson's whole error theory threatens to unravel.[67]

---

[64] See also §2.1.2 of this thesis. I describe Suikkanen's argument as persuasive since even Jaquet & Naar, who can be read as supporting fictionalism, take Suikkanen's arguments to be successful not only against conservationism, but also against their preferred fictionalism (2016 footnote 9).

[65] Holding 'one-off' evidence-insensitive beliefs need not necessarily render an agent irrational – for example otherwise respectably rational parents may believe that their child is the best child in the world, regardless of any evidence, and we would not usually think they had taken leave of their senses as a result. But the kind of systematically evidence-insensitive beliefs Olson's argument requires are a very different matter.

[66] Again, see also §2.1.2 of this thesis.

[67] An argument similar to Suikkanen's is made by Matt Bedke (2014). Very recently, Wouter Kalf has responded to Suikkanen's argument (2019), and Olson to Bedke's (2019). It is debatable how successful these responses are - the dust has yet to settle, and both responses defend accounts of belief which

My own worries centre around OC3. Bluntly, despite what Olson says, the phenomenon captured by OC3 simply is not the psychological commonplace that Olson's argument needs it to be, and I do not believe that the lying politician example Olson gives actually supports OC3. I say this for two main reasons. First, setting aside for a moment the role played by dispositions, the example only covers one 'cycle' of changes in the occurrent beliefs. At the beginning, the subject believes that *not-p* (in this case, that the politician is not telling the truth). When exposed to convincing oratory, the subject shifts to believing that *p* (that the politician is speaking truthfully), and then on reflection reverts to believing that *not-p*. In order to show how Olson's theory could be stable in practise, the example needs to be extended to include re-entering the non-critical context and once again believing that *p*, followed by again re-entering a critical context and believing that *not-p*, and so on. Without this extension, there is no reason not to simply conclude that the subject is gullible. It is much more plausible that with repeated instances of being 'taken in', the subject will become increasingly resistant to manipulation, and eventually come to have a stable belief that *not-p*. That this is a widespread response is summed up by the common saying 'fool me once, shame on you; fool me twice, shame on me'. Having been taken in before, the real psychologically familiar reaction is to resolve not to be fooled again.

In the moral case, we must recall what I wrote in §4.1 - remember that many people will find understanding and accepting a moral error theory difficult, and that the conclusion that there are no true moral judgements may be reached only very reluctantly.[68] So in certain contexts

---

are controversial, to say the least. That being the case, to avoid excessive digression I set the matter aside for now, but note that it may bear returning to in future work.

[68] A prominent example would be Smith, who in various texts has expressed gloom at the thought that something like an error theory might be true, including 'I am […] worried, deep down…' and 'the conclusion I see *looming* is thus wholesale moral skepticism' (1995b p. 278, emphasis added). Possibly most emphatically, see the profoundly dejected tone he uses when describing how analysing evaluative beliefs may lead us to conclude that human life is absurd: 'We dance and drink and have a ball in the sense of acting on the evaluative beliefs and desires that well up. In this way we move forward in the only way we can given that a rational response to our circumstances is impossible. It is rather unsettling to realise that our pursuit of value, if that is indeed what we are doing, is underwritten by such

to temporarily overturn or forget the belief that a moral error theory is true is surely nothing like being temporarily taken in by convincing politicians. I suggest that there is a much more accurate psychological commonplace which captures what it would be like to try to force oneself to believe the opposite of so hard-won a belief as a belief in the truth of a moral error theory: it is summed up neatly by the phrase *living a lie*. It is hard to see how choosing to enter such a psychologically unstable and potentially damaging state can be recommended on prudential grounds. Thus Olson's explanation fails to show that OC3 is true when it comes to moral beliefs in a post-error-theory context.

Second, Olson's example is perhaps better understood as demonstrating a more conventional model of belief than Olson suggests. Olson uses two different phrases for what he has in mind, both of which feature in the quote above. On one hand, he claims that throughout the example, the subject has two fully formed beliefs, but that one of them is 'suppressed or not attended to'. On the other hand, he describes a subject who currently believes that *p* as being *disposed* 'to believe that *not-p* in certain contexts'. It is not entirely clear to me what Olson means by being disposed to believe a proposition in certain contexts but not in others, and thus it is unclear whether it is quite the same thing as having a suppressed belief to which we attend in certain contexts but not in others. Olson seems to be equivocating somewhat. But I suggest that there is a highly credible reading of this which makes the 'disposition model' quite obviously correct, and which avoids Suikkanen's objection discussed above. That reading is that we are disposed to have beliefs which accord with the evidence of which we are aware. This quite standard epistemological view is better supported by Olson's example than the analysis he actually gives us.[69]

unreasoned responses. But unfortunately, […] when it comes to the pursuit of value, that really does seem to be all that there is.' (2006 pp. 105-106.)

[69] For a discussion of several authoritative accounts of belief which agree that beliefs are responsive to evidence, and so supports the claim that this is part of a standard view, see Humberstone 1992.

In the example, the subject begins with a belief that the politician is lying because they have evidence to support this belief (potentially, that the politician in question has been shown to be a liar in the past, that what the politician is saying is initially unconvincing, that all politicians are liars and so on). The subject is then exposed to the politician's oratory, and believes that what is being said is true. There are two ways we might view this, both of which are more plausible than Olson's interpretation. One view is that the subject takes either the apparent authority of the politician or the politician's persuasive way of speaking as evidence that what they are hearing is true. People are taken in by appeals to authority or by 'snake oil salesmen' all the time. In this case, the subject takes this new evidence to be more convincing than the evidence on which they based their former belief (albeit only until they later realise that they were being misled). On this view, the subject quite straight forwardly has an evidence-sensitive belief that *p*, followed by an evidence-sensitive belief that *not-p*. Thus the disposition which is really at work, whatever Olson might claim, is a wholly uncontroversial disposition to form beliefs based on evidence, and to give up beliefs which are contradicted by (more convincing) evidence. Given that acceptance of error theory counts as accepting conclusive evidence that all positive moral beliefs are false, this suggests that when they come to accept error theory, agents' positive moral beliefs that *p* (for example that torture is morally wrong) should change to beliefs that *not-p* (e.g. that torture is not morally wrong). In other words, Olson should be an abolitionist.

The other view is that the subject never actually forms a belief that the politician is telling the truth, but rather experiences a suspension of disbelief akin to the familiar feeling of watching a film or reading a book. In this phenomenon, we suppress those of our beliefs which are incompatible with what we are being told or shown (e.g. that no human can turn green and throw tanks around). But crucially, we do not replace those suspended beliefs with new beliefs, we merely set them to one side for the duration of the story. When the story has

concluded, we return to our original beliefs without any epistemological controversy. This view has the advantage for Olson of fitting neatly with his 'suppression of belief' explanation. But it has the considerable cost in everyday contexts of making moral judgements something which we 'make-believe' or 'entertain' rather than actually *believe*, and thus makes Olson a fictionalist, rather than a conservationist.[70]

Either one of these views fits perfectly well with Olson's example. Further, on either of these views, any considerations of evidence-insensitive beliefs are explanatorily surplus – we do not need any of this controversial, non-standard kind of beliefs to explain the phenomena at hand. Therefore Olson's politician example backfires. Olson needs to provide us with an example which actually does support OC3, because OC3 is a central tenet of his argument, and it depends on doxastic voluntarism - a controversial model of evidence-insensitive beliefs which is not necessarily plausible. But the example he chooses fails, and is actually better explained by standard accounts of belief. Therefore Olson's argument for conservationism fails, and by his own lights he should either be an abolitionist or a fictionalist. Taking this together with the argument which I discussed above that we should not grant OC2 either, I conclude that conservationism cannot currently be regarded as a successful WNP response.

4.3. Abolitionism

Abolitionism, also referred to with varying degrees of appropriateness as eliminativism, moral nihilism, anethicism or amoralism is perhaps the most obvious, immediate response to the acceptance of error theory.[71] Olson calls it a 'natural reaction' (2014 p. 179). Once our moral thought and discourse have been found to be defective in some major way, abolitionists quite

[70] I will discuss the difference between conservationism and fictionalism, and exactly what being a fictionalist amounts to, in §4.4.
[71] See e.g. Nolan, Restall & West 2005, Hinckfuss 1987, Garner 2007, Pigden 2007.

simply claim that we should abandon them. After all, we no longer discuss medicine in terms of the four humours, or maintain the Ptolemaic belief that the earth is at the centre of the universe. History is littered with ways of thinking and talking about the world which were left behind once we realised that they were not accurate representations of the way things really are. Therefore, if we have found our ways of thinking and talking about moral matters to be seriously defective, as acceptance of moral error theory necessarily implies, the same considerations would seem to count in favour of abandoning them as well.

Abolitionism has been defended by, among others, Richard Garner (e.g. 2007, 2011, 2019), John Burgess (2007), and possibly most powerfully and influentially by Ian Hinckfuss (1987 & 2019).[72] And abolitionism is currently the focus of significant debate, following the recent publication of a book entitled *The End of Morality: Taking Moral Abolitionism Seriously* (Garner & Joyce 2019). In the context of WNP responses, there are two main aspects to the abolitionist position. The first main aspect is notable in that, particularly in the discussions by Garner and Hinckfuss, the opposition to moral thought and discourse can be read as separable from error theory. The argument is that on balance, far from beneficially promoting accord or counteracting our limited sympathies, the moral dimension of our lives is in fact responsible for much avoidable harm. Perhaps the clearest expression of this is in chapters three and four of Hinckfuss' *The Moral Society* (1987). In a style reminiscent of Marx, Hinckfuss accuses morality as a whole of, *inter alia*, fostering and cementing elitism (§3.2), of being authoritarian (§3.3), of thwarting altruism (§4.6), of hindering or making impossible the resolution of

[72] A note on Hinckfuss references. Hinckfuss' *The Moral Society: Its Structure and Effects* was published under that title in 1987. This work is not easy to find in print today, hence the refence in my bibliography to an online version, and my occasional use of section numbers rather than page numbers. But a somewhat abridged series of excerpts from the book were published in Garner & Joyce 2019 under the title 'To Hell with Morality' (which we are told in an editor's note was Hinckfuss' originally preferred title for his book). Where more specific references are appropriate (for example quotations or references to the use of specific words), I will refer to this more recent volume, simply because it is more readily available and has easily identifiable page numbers.

disputes (§4.2), of allowing tyrants such as Hitler and Stalin to manipulate citizens (§ 3.6), and of paving the way to war (§4.3).[73]

These accusations may initially sound extravagant to those of us accustomed virtually from birth to thinking and speaking in moral terms, but they are perhaps more plausible than at first they seem. For example, in discussing how wars and major conflicts are justified and represented to the public, Hinckfuss writes,

> Think of any one of these conflicts and think of how the situation would have been if, by a miracle, moral thought could have been eradicated from the minds of all the agents involved. I, for one, find it difficult to conceive of how the conflicts would have proceeded. There would be no sense of duty, no sense of loyalty, no patriotism, no feeling morally obliged to fight for a cause, no sense that the people one is trying to kill or subjugate are less worthy of survival or freedom than oneself or anyone else. (2019 p. 36).[74]

That is not to say that abolitionists believe that phenomena such as war would be impossible if we abolished moral thought and discourse. Rather, Hinckfuss argues that morality makes war a significantly more likely prospect: 'There could be war without morality. But moral propaganda eases the task of those with control of the mass media to get almost all the nation determined to attack, plunder, slaughter and subjugate another group of people' (2019 p. 36).

[73] See Wood 1993 for an analysis of Marx's metaethical views which closely resembles parts of what Hinckfuss has to say. The phrase 'morality as a whole' here is intended in a broad sense which reaches beyond the analysis of morality typical among error theorists. Hinckfuss specifically includes 'the belief in moral obligations, vices, moral virtue, sins and morally good or bad acts or morally good or bad people' (2019 p. 23).

[74] The conflicts Hinckfuss is referring to are 'the situation in Ireland (unresolved after four hundred years of bloody conflict); the situation in the Lebanon (unresolved after about eight hundred years of conflict between Christian and Moslem); the Palestinian Arabs versus the Zionists; the Vietnamese versus the Khmer, the Chinese, the French and the Americans; all the wars of religion and all the blood-letting of the two world wars.' (2019 p. 35).

Alongside the extreme situation of war, another example might be abortion, an issue often conceived of in moral terms which allow scant leeway for compromise – the right to life, the right of women to decide what happens to their bodies and so on – and which is the subject of implacable disagreements which can appear insoluble. In such cases abolitionists argue that if the notion were removed that there are relevant objective, immutable moral facts of the matter, it seems plausible that disputants would find it easier to reach an agreement of some kind based on more pragmatic considerations. For example Garner argues that, while genuinely empathising with others' predicaments rather than leaping to moral judgement is 'capable of unlocking a rich vein of compassion', approaching matters from a moral perspective is much more likely to 'harden our resistance and send us scurrying off in a search for counter-arguments and reasons not to care' (2019 p. 87).[75]

If abolitionists such as Hinckfuss and Garner are right, and assuming that we desire to avoid elitism, war and so on because they are harmful, it may well be that despite the limited benefit which morality may occasionally afford us, we have compelling prudential reasons to discard moral beliefs and discourse before we even consider moral error theory. Indeed, Olson takes this to potentially make any further discussion by abolitionists about whether an error theory is true redundant, asking 'If moral thought and discourse really have the nasty consequences Hinckfuss insists on, why would it matter if there were after all non-natural moral properties and facts in reality? Would not a proponent of anti-elitism, anti-authoritarianism, peace, etc., recommend that we in such a scenario simply ignore the moral properties and facts?' (2014 p. 179, footnote 5).

However, in posing this question, Olson misses an important point. If moral facts were simply part of the fabric of the world then we may have no choice but to persist with the moral

[75] Garner is not arguing specifically about abortion here, but the point is clear.

practices we have. Regardless of the consequences of doing so, we would have authoritative reasons for acting in certain ways – i.e. it would remain the case that we ought to follow moral rules even if it led to our ruin. In moral error theory, however, we have a compelling set of arguments to the conclusion that moral values *are not* part of the fabric of the world. Whatever other ideas abolitionists may choose to discuss, therefore, moral error theory can be seen as a crucial tool to make the abolitionist's case.[76] In simple terms, when abolitionists claim that we would be better off without moral values, error theory allows them to add 'and anyway, thankfully, there's no such thing in the first place!'. [77]

The second aspect of the abolitionists' response to the 'what now?' problem is that, even if we can maintain false beliefs (bearing in mind that this is controversial, as I discussed in §2.1.2 & 4.2.1) they question whether we ought to defend having beliefs which we know not to be true. Quite obviously, the 'ought' here cannot be a moral ought, for this would violate the Normative Circularity Constraint. Nonetheless there are numerous non-moral senses in which there may be something one ought to do. At least two are relevant here. First, absent the specific sort of case made by conservationists, it seems unlikely that we would benefit from a general policy of believing things which we know to be false (if that is even possible). There is therefore a prudential 'ought' which one could apply here. Second, it can be argued (albeit not without controversy) that one ought to have true beliefs as opposed to false beliefs because truth itself is normative – as I put it in §2.1.2, some philosophers would argue that

[76] Olson also comes close to selling short the abolitionist argument when he writes 'One is tempted to say that this critique of what Hinckfuss calls 'the moral society' is *moral* in nature, but we can say more charitably that Hinckfuss claims to identify some consequences of moral thought and discourse that many people are opposed to, and that on this basis he recommends abolishing morality' (2014 p. 179, emphasis original). Hinckfuss in fact requires no such charity. He explicitly frames his criticisms of moral thought and discourse in terms of what people desire, see e.g. the beginning of §4.1. This is why I used the phrase 'assuming that we desire...' near the start of the preceding paragraph.

[77] See also Joel Marks, who despite the 'undoubted utility' of having moral beliefs finds that 'the disutility looms larger', leading to a situation wherein 'the abolitionist sees error theory as presenting an opportunity – to make the world more to our liking – that abolitionism then seizes' (2019 p. 101-102).

beliefs 'aim at truth'.[78] Garner is emphatic: 'What serious philosopher can long recommend that we promote a policy of expressing and supporting, for an uncertain future advantage, beliefs, or even thoughts, that we understand to be totally, completely, and unquestionably false?' (2007 p. 512).[79]

As well as highlighting the potential harm of maintaining false beliefs, here Garner also anticipates a theme to which I shall return in detail shortly - that we cannot know in advance whether retaining known false beliefs (or as I will discuss in §4.4, make-beliefs or 'thoughts'), will be of any benefit at all. While plausibly likely outcomes may be a satisfactory basis for WNP responses in the absence of other complicating factors, when taken along with the above worries, this leads abolitionists to recommend against adopting any WNP response involving false beliefs or anything substantially like them.

To sum abolitionism up, given that accepting the error theory opens up the possibility of discarding moral thought and discourse, difficult though that may be, abolitionists claim that we should do so. They offer several reasons why this would be desirable or unavoidable. Moral beliefs are of questionable benefit to society, and may be pernicious. Even if we could maintain moral beliefs at the same time as accepting a moral error theory, doing so when we know them to be false is tantamount to lying both to ourselves and to others. Even if this kind of lying could somehow be justified in terms of how useful it is to society, in the moral case it is impossible to be sure that its consequences would be beneficial. Garner sums things up well: 'Moral memes have burrowed deep into our brains and our public rhetoric, but we can root them out by reminding ourselves that morality is a human invention based on biology, ignorance, credulity, fear, and a lust for control. If the belief in objective values is as mistaken

[78] For further discussion of this epistemic sense of 'ought', see e.g. Gibbard 2005 & Price 1998.

[79] *Cf.* Joyce 2001, p. 214: 'No policy that encourages the belief in falsehoods, or the promulgation of false belief in others, will be practically stable in the long run'.

as the error theorist argues, and as harmful as the moral abolitionist believes, then we would be well advised to find another star by which to sail.' (Garner 2011).

4.3.1. Abolitionism in practise

It may seem obvious how abolitionists suggest we might respond to the WNP, then: by completely abolishing morality and never making a moral judgement or a positive moral claim again. But this is simplistic. First, it may be objected that eliminating moral judgement may be impossible (Joyce refers to arguments to this effect by Peter Singer and Michael Ruse (2001 p. 171), plus see footnote 47 about Strawson, above), or at least exceedingly, problematically difficult (e.g. Nolan *et al.* 2005 p. 314). But even if we grant that somehow abolishing moral beliefs is feasible, there are several ways this might turn out in practise.

At the weaker end of the spectrum, Russell Blackford advocates what he calls a 'mild or partial' (2019 p. 74) form of abolitionism whereby we continue to employ previously morally loaded terms such as good, bad, should, cruel, kind, and so on, and continue to approve of and abide by previously moral norms such as those relating to debts and promises. According to Blackford, these pieces of 'social technology' (*op. cit*. p. 62) are too useful to abandon entirely, and may be repurposed and redeployed even after we have accepted the error theory. We would simply be mindful as we do so that these terms and norms should henceforth be understood as devoid of objective (moral) authority, and should instead be understood as useful, non-objective conventions. Essentially, on Blackford's proposal, the furniture of traditional morality would remain in use, but any connotations of moral normativity would be replaced by hypothetical norms grounded in the instrumental benefits of practices such as keeping promises and so on.

Presenting a much stronger form of abolitionism, Joel Marks argues (2019) that although the error theory we accept is limited to morality and does not extend to all domains of normativity, we would be best served by treating even non-moral normative terms as posing a psychological threat to drag us back into having objectively prescriptive beliefs. Marks therefore recommends that we refrain from using obviously morally evaluative terms such as moral uses of right, wrong, good and bad. But he goes much further than this and advocates the elimination from our discourse, and eventually our psychology (2019 p. 106) of 'the attitudes of not only morality but also of other value realms to the greatest degree practicable' (2019 p. 102). This includes even the appearance of any value objectification and 'any *hint* of allusion to categoricity' (p. 103 emphasis original). Only then will we be able to truly leave morality behind and move forward as genuine abolitionists.

By contrast, Hinckfuss is compellingly blunt: 'if you want to minimise conflict and you do not want widespread denigration, guilt complexes, elitism, authoritarianism, economic inequality, insecurity and war, then throw morality away and think about how best you can resolve conflict without it.' (2019 p. 38). Hinckfuss' main target is the doctrine of moral desert - i.e. the notion that people may be morally inferior or superior to others – and if we could rid ourselves of that, he suggests that morality in general would quickly disappear (1987 §3.5).

4.3.2. Problems with abolitionism

I mentioned above the issues around the normative circularity constraint and the feasibility of abolitionism which could be thought problems for abolitionists. But even granting that the former is not necessarily a problem for abolitionists who are *moral* error theorists, and granting for the sake of argument that being an abolitionist is not impossible, I believe abolitionism faces insurmountable problems as a response to the WNP. Yet the upshot of these problems may be surprising, as I will shortly explain. I will discuss three flaws of

abolitionism, ultimately concluding that while the abolitionist challenge must be taken seriously, we should not respond to the WNP by becoming abolitionists.

The first problem with abolitionism is that we cannot plausibly do as Marks recommends, because his recommendation is too extreme. As part of his crusade against moral or even moral-seeming thought and discourse, Marks recommends (2019 pp. 102-106) that alongside moral terms and attitudes, we eschew the use of 'ought' and related concepts, forms of argument which invoke practical justification, evaluations of aesthetics or humour, uses of terms such as 'good', 'better' and 'right' even in non-moral contexts such as hypothetical imperatives or in a functional sense, and judgements about whether what others say is true and whether others are rational. Finally, Marks recommends that we alter our understanding of 'thick' conceptual terms, i.e. those with both descriptive and evaluative conceptual content such as 'liar', to purge them of evaluative content while nonetheless continuing to use the terms in a purely descriptive sense.[80]

Marks recommends this laundry list of lexical, practical and conceptual changes despite admitting that 'I myself can attest to the enduring power of my moralist reactions to various actions and traits and states of affairs even though I have striven to suppress them for several years now' (p. 101). Now recall that unless we are to keep the truth of error theory a strict secret known only among metaethicists (which would surely not be stable in the long run), then as I discussed in §4.1.3, WNP responses must be addressable to all moral agents. Yet the degree of conceptual self-vigilance and discipline Marks' proposal would require must surely be beyond all but the most insightful and ruthless handful of prospective abolitionists. In comparison with conservationists' 'just carry on having moral beliefs' or, as we shall see in

[80] One might also wonder whether Marks would recommend getting rid of predictions, too, since they can sometimes be or appear to be normative. The same could also apply to probabilistic expectations, suppositions… The list goes on.

subsequent sections, fictionalism's 'just act as if moral beliefs could be true, even though they can't', Marks' advice, which amounts to 'make myriad changes to a wide range of deeply ingrained patterns of thought and speech, some of which are very subtle, and which even I cannot reliably accomplish after years of trying', is simply not a viable WNP response.[81]

Turning to the second problem for abolitionism, if Marks' proposed form of abolitionism is too extreme in its demands, then perhaps the comparatively less demanding Blackford model will fare better. Blackford's proposal is after all relatively modest – he recommends that we continue to use familiar moral terms, but that we just think of them as relying on hypothetical norms rather than categorical norms of the type which we are error theorists about.[82]

The issue here is that Blackford's model of abolitionism is confused and confusing on a metaethical level. In recommending his general strategy of repurposing moral discourse, he glosses over important distinctions between fundamentally incompatible metaethical views. Yet when the different views are drawn out, they appear unsophisticated and unworkable. For example, concerning moral deliberation, Blackford makes the central question for any agent 'what should I do?' (2019 p. 68). On Blackford's proposal, in the WNP context this would be stripped of any moral sense of 'should', and replaced with a consideration of 'which path is most likely to meet the agent's subjectively weighted ends, to the extent that she can identify them' (*ibid*.). There is nothing necessarily wrong with this on its own, of course, but a metaethical view couched in these terms inherits a classic problem for contractarianism –

[81] A similar objection could potentially be made to my own proposal, i.e. that it would be problematically complicated or difficult to implement. I will discuss this in §7.5.

[82] It may be possible to set up a dilemma for abolitionists here. Even if Marks' proposal is too extreme to implement, arguably there may nonetheless be some truth to his claim that using previously morally normative terms at all threatens to turn error theorists back into moral realists (I do not think this myself, but clearly Marks would argue this). If so, then Blackford's more lax view would pose an existential threat to error theorists as such. This could set up a dilemma: take the strong view and be unable to implement abolitionism, or take a weaker view but undermine error theory. However I leave this to Marks and Blackford to sort out, should they wish to do so, and focus on the arguments in the main text myself.

the free rider problem (see e.g. Shafer-Landau 2010 pp. 190-194). This problem arises because agents whose motivation is to satisfy their own subjective ends are often best served not by being moral themselves, but by getting everyone else to be moral, and taking advantage of them. With no reason to think beyond their own ends, agents are free to act on principles such as 'when you can get away with it, steal' or 'so long as there's no legal or financial penalty, and you don't care personally about them, exploit your employees mercilessly'. This results in terms such as 'good' meaning such wildly different things to different people as to be useless in any coordinative sense, and so means that previously moral discourse cannot carry any useful authority or bring any significant portion of the putative benefits of its pre-error-theory ancestor.

At other times, Blackford appears to propose that societies as a whole consider what would be good for the whole society, and that accordingly, members of a society would then use previously moral language such that e.g. 'good' would come to mean something like 'allows society to function well'. Thus the post-error-theory analogue of moral deliberation would be about what conduces to societal goods, and morality would be replaced by something quite similar to a legal system (2019 §7). This might seem to get around the free rider problem because it moves the focus of practical deliberation away from the individual's ends and towards cooperation among members of a society or group. Yet this faces its own set of classic objections, this time inherited from cultural relativism (see e.g. Shafer-Landau 2010 pp. 278-291). Seen in this way, Blackford's view is something like the relativist view I will propose, but less sophisticated because it offers none of the ways to cope with classic objections which I offer on behalf of my proposal. So anyone tempted by this tendentious reading of Blackford's somewhat confusing proposal would be better directed to the subsequent chapters of this thesis.

The third problem with abolitionism is more fundamental, and goes to the heart of abolitionists' claims about traditional morality. Abolitionists such as Hinckfuss and Garner claim that an amoral society would be less elitist and less authoritarian, and that in the absence of moral thought and discourse we would be better able to resolve conflicts and avoid wars. And I suspect that they may well be right, at least to some extent. But the problem is that there is little theoretical heft to this claim. What we have is primarily a series of empirical claims about the harms of morality, but it is not clear how they could be proven. Therefore it remains perfectly possible to disagree, and simply make the contrary claim that moral thought and discourse are very useful to us, at least on balance - as Olson, Nolan *et al.*, and Joyce do.[83] In other domains, in order to resolve such a disagreement, disputants would typically point to relevant data, usually in the form of real world examples, and explain how those data support their view rather than their opponents' view. But in this case I think it likely that the vast complexity of society makes it impossible to conclude in favour of either party without empirically determining exactly what an amoral society would be like. And such an experiment seems unlikely ever to be performed.[84]

It could be argued that this is simply a failure of imagination on the part of non-abolitionists – perhaps the majority of us, even if we accept a moral error theory, are so wrapped up in a morally informed view of the world that we cannot see the benefits which would accrue to us if only we could step outside our moral conditioning and 'throw morality away'. Garner draws an interesting parallel between abolitionism and 'new' atheism (2011), and it may well be fair to extend that parallel and observe that at one time, many people would have found the idea of life without a religious dimension incomprehensible. Yet today it seems self-evident to huge numbers of people that it is perfectly possible to live without believing in any gods.

[83] See e.g. Olson 2014, p. 180, Nolan *et al.* 2005, p. 307, Joyce 2001, pp. 184-185.

[84] As the reader may have noticed, this has significant implications for the burden of proof here, even beyond those I discuss in the text. For the time being, I will set this matter aside and focus on abolitionism. But I will return to this theme in significant detail in chapter 6.

Likewise, the abolitionist might suggest that in time, even committed moralists would be able to see that we would be better off if we were to embrace abolitionism.

But the parallel with atheism only holds so far. While it may be impossible to prove an ontological negative, we can certainly compare the lives and psychologies of theists and atheists to empirically test whether any gods are involved in beneficial ways, or whether secular explanations are sufficient to explain empirical phenomena relating to material and emotional wellbeing. For example we might test a given religious claim – that prayer to the Christian god is the most beneficial form of prayer, say - by examining and comparing Christian people and cultures with others who adhere to other religions, and to none. But an analogous moral test cannot be performed. The abolitionist makes claims not about distinct moral theories, but about morality itself. Thus despite abolitionists' claim that we may be able to experience the benefits of abolitionism simply by experimenting in our own individual lives (for example, '[t]ry it – you might like it' (Marks 2019 p. 106), or 'give moral abolitionism […] a test drive' (Garner 2011)), the only suitable comparator with our current moralised society would be an entire, entirely amoral society – and no such society exists.

In conclusion, then, the abolitionists' case rests too heavily on practically unprovable empirical claims and too little on provable theory to be convincing to anyone who does not share abolitionists' controversial intuitions. Seen as a challenge, as an attempt to pose questions which might shake us out of complacent assumptions about how wonderful moral beliefs are (or perhaps were), Hinckfuss' critique of morality draws attention to previously unrecognised *potential* negative consequences of moral belief such as authoritarianism, the hindrance of conflict resolution, and even war. On this level, abolitionism makes an invigorating contribution to metaethics, and I will have much more to say about this 'challenge from abolitionism' in later chapters. Yet as a WNP response, absent an unfeasibly large-scale set of experiments, and against the backdrop of a near-ubiquitous intuition that moral thought and

discourse are beneficial or useful to society, the way is abundantly clear for abolitionists' opponents to simply reject their claims out of hand. As such, even if abolitionists may be right about some matters, abolitionism cannot be a convincing response to the WNP. Taking this along with the other problems I discussed relating to sustainable implementation, I conclude that abolitionism should be ruled out as a viable WNP response.

Thus far, then, I have ruled out conservationism and abolitionism as responses to the WNP. I will turn now to the penultimate response I will discuss in detail, revolutionary fictionalism. This is probably the WNP response which has generated the most discussion and found the widest support to date. Yet as I will explain, I believe it too faces insurmountable problems.

4.4. Revolutionary fictionalism

There are various kinds of fictionalism about various domains of discourse.[85] When it comes to metaethical fictionalism, there are two main alternatives: hermeneutic fictionalism and revolutionary fictionalism. I will set aside hermeneutic fictionalism here, as it is a theory about how traditional moral thought and discourse should be understood.[86] Our task here is to decide what we should do after we have accepted a moral error theory, and so arguments about what we should do *rather than* accept an error theory are moot. Revolutionary fictionalism is proposed specifically as a response to the 'what now?' problem, and therefore it assumes the truth of error theory – this is why the term 'revolutionary' is appropriate.[87] It

[85] For an overview of several non-moral domains of fictionalism, see Eklund 2015. For a more detailed discussion of a specific non-moral example of fictionalism (in this case mathematics), see Balaguer 2018.

[86] For a sophisticated view of hermeneutic fictionalism, see Kalderon 2005.

[87] Echoing footnote 59, since I have set other forms of fictionalism aside and will be concentrating solely on revolutionary fictionalism, hereafter I will frequently refer to revolutionary fictionalism as simply fictionalism without further qualification.

has been defended most prominently by Joyce (2001 ch.8, 2005, 2017, 2019) and Nolan *et al.* (2005). It has also been defended more recently by Jaquet & Naar (2016).

Fictionalism shares with conservationism the basic assumption that the practice of engaging in moral thought and discourse is useful – that e.g. having moral beliefs delivers prudential goods unattainable via other means, and we are therefore better off having them than not having them. As we have seen, the claim that traditional morality is beneficial is open to question, especially by abolitionists. But for now I will grant this claim in order to focus on fictionalism itself (though I will return to it later on). While it may be the case that only genuinely having moral beliefs can secure all relevant prudential goods, fictionalists argue that we can secure at least some significant proportion of them by speaking, acting and, importantly, thinking *as if* there were moral facts and properties.

In the WNP context, fictionalists typically see abolitionism as their primary opposition. Joyce, for example, simply dismisses conservationism out of hand (2001 p. 214), and while this might seem hasty, my own discussion of Olson's conservationism shows that he is probably right to do so.[88] And since its adherents claim that fictionalism can deliver at least some of the benefits of genuine moral beliefs, fictionalists argue that it is preferable to abolitionism, which delivers no such benefits and which 'would force large-scale changes to the way we talk, think, and feel that would be extremely difficult to make' (Nolan *et al.* 2005 p. 307). Other WNP responses are largely quite recent, and have yet to provoke proper responses from fictionalists.

Ultimately then, the primary fictionalist claim is that error theorists ought to become fictionalists on prudential grounds. Over the next four sections I will examine whether they

[88] For an alternative discussion of conservationism which finds it similarly implausible (and that fictionalism is preferable), see Jaquet & Naar 2016.

are correct to make this claim. I will begin by explaining what fictionalism is, and then outline how fictionalists claim that their proposal can be useful. Following this, in sections 4.4.3 & 4.4.4 I will explain why I believe fictionalism fails to be an adequate response to the 'what now?' problem.

4.4.1. What fictionalism is

To be a fictionalist about a given domain of discourse involves using its terms in a very similar manner to realists, whilst treating the discourse in question as fictive. Typically, this involves either holding genuine beliefs about fictional subject matter (so-called content fictionalism), or having thoughts with similar content to genuine beliefs, but treating those thoughts as something other than genuine beliefs (so-called force fictionalism). I will explain what these alternatives amount to further below. In the moral case, fictionalists will use terms such as duty, obligation, right and wrong in everyday situations much as realists might. They may even think in moral terms, seemingly arriving at moral judgements in accordance with moral rules, and so on. But, crucially, and in contrast with how error theorists view traditional morality, in doing so they will not be using moral terms to make assertions about states of affairs in the actual world. Accordingly, although in everyday contexts fictionalists will speak and even think as if error theory were not true, in more critical contexts (such as sophisticated metaethical discussions) they will remain disposed to affirm the truth of error theory, and the falseness of their day-to-day moral thoughts and statements.

This is more familiar than it may initially sound. To illustrate, consider that we often discuss fictional individuals and situations using ostensibly the same language we use to discuss real people and events. For example, if I claim that Harry Potter has an unblemished forehead, you might reply that I am mistaken. To an observer with no knowledge of the relevant fiction, this would appear to be a conversation about a real person. Yet the Harry Potter we are

discussing does not exist - there may be individuals with that name of course, but none of them is a young wizard, has attended Hogwarts, was given a scar by an evil sorcerer, etc.

There is a sense, however, in which my claim about Master Potter's faultless complexion would nonetheless be wrong. That is because *according to the fiction* (in this case, the books by J. K. Rowling), Harry Potter has a scarred forehead. When we discuss the character, we know this, and we understand that we are talking in terms of the fiction, rather than the actual world. We might even go a step further and take on the role of the actors in a Harry Potter film, speaking and acting consistently as if the fiction were true. But all along we would not believe the fiction, we would merely *make believe*. And when we were not acting, we would remain disposed to say that Harry Potter is a fictional character, and that no such real person exists.

Fictionalists argue that we can use moral language in much the same way. Recall from chapter 2 that error theorists typically hold that moral discourse is assertoric, and commits us to the existence of moral facts or properties involving phenomena such as moral obligations. Yet they do not believe that moral facts and properties exist, and accordingly they argue that it is never the case that we have moral obligations. Thus, according to error theorists, moral discourse is systematically in error. When we enter a fictional mode of discourse, however, we do not commit ourselves to the existence in the actual world of anything at all. This can happen in two ways.

First, our discourse may remain assertoric in structure but refer to facts and properties in a fictional world as opposed to the actual world. This is what is often referred to as content (or preface) fictionalism. It is analogous to the way in which we might disagree over Harry Potter's scar. In that case, the truth conditions which dictate which one of us is correct are nowhere to be found in the actual world. But they do exist in a fictional world, one which we

understand ourselves to be referring to and which has fixed features, namely the world described in the Harry Potter novels. This cognitive form of fictionalism is what I understand Nolan *et al.* to advocate (2005). According to content fictionalists, moral claims can be read as elliptical, expressing beliefs of the form '(according to *fiction F*), *x* is *y*'. I will have more to say about how we might define 'fiction F' shortly. But since the motivation for fictionalism is the claimed usefulness of moral thought and discourse, and since error theory rules out any *moral* grounds for preferring one fiction over another, we can suppose for the time being that it equates roughly to 'whichever fiction is most useful' or 'whichever fiction brings about the greatest number of things we desire'.

Second, we may speak without assertoric force in order to express our conative attitudes, or perhaps in order to achieve some perlocutionary effect such as persuading or inspiring others. Alternatively, we might do so in order to bolster our own motivation to act in certain ways. This is what is often referred to as force (or attitude) fictionalism, and is analogous to the acting case I mentioned above. Force fictionalists understand moral judgements to have the same content as realists' moral beliefs. However, they consider the attitude we have to that content to be something other than belief. Thus when we judge that 'torture is wrong', our judgement has the content that *torture is wrong* (as opposed to *according to fiction F, torture is wrong*), but rather than believing that content, we merely entertain it or quasi-believe it. Accordingly, on force fictionalism, when we utter 'torture is wrong', we do not assert that there is any fact of the matter about the moral status of torture, we merely quasi-assert it.[89] This non-cognitive fictionalism is what Joyce advocates (2001 chapter 8, and more explicitly, 2017).

[89] This could be phrased in terms of pretending to believe and pretending to assert, but pretence is slightly too weak here. A bad actor who never really engages with their role can pretend to assert or to believe something. But Joyce requires something stronger than that, albeit stopping short of actual assertion. A parallel with what quasi-assertion is meant to capture here might be a more immersive form of 'method' acting.

The following table helps to show the differences between these various ways of analysing moral judgements by highlighting which mental states moral utterances express, according to each type of analysis. In each case, we can see what kind of attitude a moral judgement is, and what the content of that attitude is. I will use a judgement that torture is wrong as an example, and include cognitivist realism and a basic form of (hermeneutic) expressivism as a reminder of the more conventional ways of understanding the issue.

A judgement that torture is wrong consists in:

Realism (cognitivist):	Belief [torture is wrong]
(Basic) expressivism:	Desire [do not torture]
Content fictionalism:	Belief [according to fiction F, torture is wrong]
Force fictionalism:	Entertain [torture is wrong]

4.4.2. How fictionalism might secure the claimed benefits of conventional morality

No matter what variety of fictionalism they defend, all fictionalists argue that by taking a fictionalist stance towards moral thought and discourse, we can secure (at least some of) the same prudential goods as those associated with genuine moral beliefs, without committing ourselves to the existence of moral properties and facts in the actual world. The specific benefits of the moral discourse which fictionalists argue their approach can secure vary according to the theorist(s) concerned. Nolan *et al.* focus largely on social issues, and are concerned primarily with demonstrating that fictionalism is preferable to abolitionism.

Accordingly, they list four advantages which they claim abolitionism cannot provide but which fictionalism can (2005 p. 310*ff.*):[90]

1. Psychological convenience. Moral thought and discourse are embedded into society in general and into individual psychology at a fundamental level. Nolan *et al.* claim that it would be exceedingly difficult and require enormous effort to divest ourselves of them. It may even be impossible. Therefore, continuing to use moral discourse in a fictional mode is of significant practical benefit because it avoids this difficulty while remaining consistent with error theory.

2. Avoiding frequent digressions into complicated metaethics. In everyday contexts, when other people (who may well not be error theorists) raise issues of applied ethics, Nolan *et al.* claim that fictionalists can consistently respond in a way which gets to the point of the discussion without having to explain at length why the issues are in fact illusory etc. Abolitionists, they claim, cannot.[91]

3. Expressive power. Nolan *et al.* argue that it is possible to use moral language to say things about the non-moral world which would be cumbersome or even impossible to say using non-moral language. One example Nolan *et al.* give is 'Suppose that we want to say that the duty to stop and attend to the victims of a car accident is more important than our duty to keep

[90] Nolan *et al.* do not necessarily take the four listed advantages of fictionalism to be the only possible advantages, citing for example potential uses of moral discourse in child rearing and avoiding inconsistency when error theorists are tempted to take moral claims to be true. But the four advantages listed are their primary specific claims.

[91] Nolan *et al.* use the phrase 'when speaking with the vulgar' (p. 311), echoing Berkeley's advice to 'think with the learned but speak with the vulgar'. In a spirit of charity, I am content to read the use of 'vulgar' as an evaluatively neutral reference to non-fictionalists as opposed to some kind of Freudian slip. Though I will raise an objection to fictionalism concerning speaking to others who are not fictionalists – whether they are vulgar or otherwise! – below.

our commitment to meet our friend for lunch. Again, what non-moral features of the world are implied by talk of 'duty' here?' (*op. cit*., p. 312).

4. Coordinating attitudes and regulating interpersonal conflict. Nolan *et al.* claim that moral discourse sets out a set of shared values, and gives us a conceptual and semantic framework of rights, obligations and so on which they deem to be useful in conflict resolution.

Joyce, on the other hand, focuses on the usefulness of the moral thought and discourse in avoiding temptation and weakness of will. The idea here is that at least a majority of moral rules inveigh against practices which we have non-moral reasons to avoid, or promote actions and outcomes which most 'ordinarily situated persons with normal human desires' would want (2001 p. 222). People generally desire not to be tortured, not to have their possessions stolen, not to have promises made to them broken, and so on. Moreover, they normally also desire the stability of a society in which others refrain from those and similar actions. By acting as a bulwark against weakness of will, Joyce argues that taking a fictive stance towards moral thought and discourse can help reduce the number of instances of these pernicious actions, even if people do not believe what they are saying and thinking.

The key to seeing how this might be so is immersion. Ordinarily, when we immerse ourselves in a fiction (by reading a book in a fully engaged manner, for example), we frequently feel emotions, often including very strong emotions. These emotions can sometimes extend beyond the fiction, and form the motivational basis for actions in the real world. For example, when reading *The Color Purple*, we may feel desperately sorry for Celie in the fictional world of the story. But the emotions we experience while immersed in the fiction may also prompt

us e.g. to oppose racism in the real world. Yet this will only happen if we engage fully with the fiction – if we allow ourselves to think and feel as if the fiction were true.[92]

It is by this kind of mechanism that Joyce believes that 'entertaining' propositions which we know not to be literally true can nonetheless motivate us to act in certain ways. Nolan *et al.* seem content to recommend fictionalism as a general policy, without discussing in detail what living a life as a fictionalist might be like. But Joyce goes further (2001 pp. 218-221) and recommends that error theorists should *immerse* themselves in the moral fiction. By this, he means that in most contexts, error theorists should cease to pay attention to the fact that they are error theorists at all. They should think and act entirely as if they were moral realists in all contexts other than their most (metaethically) critical contexts (such as, for example, the philosophy seminar room). For no matter what role their moral thoughts play in their deliberations, Joyce argues that fictionalists may remain consistent error theorists so long as they remain disposed to dissent from their apparent moral beliefs in a least one ultimately critical context. This, says Joyce, is the defining difference between what he calls moral thoughts or images (which he considers compatible and consistent with error theory) and genuine moral beliefs (which he does not) – the former 'may well *seem* to the subject as very much the same as beliefs. The difference between a thought and a belief here is not a phenomenological one; it is a matter of a disposition to dissent' (2001 p. 219).

It is by developing the use of apparently moral thought and discourse into a 'life strategy' – i.e. far more than a simple habit - that Joyce believes error theorists can best protect themselves from akrasia. Only by this level of commitment to the fictive discourse can we be sure that when confronted with temptation, the thought of giving in to it is unlikely to even enter our minds. This is a process which Joyce believes begins long before we might become

[92] Joyce's arguments here touch upon a much more complex set of debates in the philosophy of aesthetics. For an overview and further reading, see Gendler 2013, especially section 5.3.

error theorists, since we are (at least most of us) raised in such a way that we genuinely hold moral beliefs. But Joyce recommends that an agent who goes on to become an error theorist 'continues to use the language of morals, continues with her familiar patterns of thinking, but allows herself to express disbelief in it all when she is placed in highly critical contexts' (2001 pp. 229-230).[93]

4.4.3. Problems with fictionalism: i) Deception

Having laid out what fictionalism consists in and how fictionalists argue it can deliver the benefits they claim, I will now examine whether or not they are correct. Bluntly, I do not believe that they are. In my view, fictionalism cannot deliver the benefits which its defenders claim it can, and brings with it serious problems which fictionalists cannot explain away. Ultimately, we will have to look elsewhere for a viable response to the 'what now?' problem.

There are a number of issues with fictionalism, both in terms of the specific ways fictionalists make their case and with the overall fictionalist strategy itself. Here, I will limit my discussion to two criticisms of the fictionalist strategy in general. I do not consider the objections raised here to be exhaustive of the potential arguments against fictionalism.[94] But taken together, I consider the objections I will raise sufficient to undermine fictionalism as a response to the 'what now?' problem, and to thus necessitate the search for further responses.

[93] Joyce has recently expanded on his view by likening the non-assertoric use of previously moral terms to metaphors such as describing a person as a 'spineless snake' (2019 p. 154). I will omit discussion of this here, since I believe it rests on a misunderstanding of metaphor. When we use slurs which are not literally true – no human *actually is* a spineless snake – I would argue that we really are asserting something nonetheless, just not something which is literally true at a surface level. That being the case, Joyce's recent talk of metaphors threatens to undermine the non-assertoric form of fictionalism he defends elsewhere prior to 2019. Joyce would doubtless have something to say in response to this, but rather than get bogged down in a tangential debate about metaphor, I therefore choose to focus on Joyce's more direct earlier presentations here.

[94] For further criticisms of fictionalism as a post-error theory option, see e.g. Garner 2007, Lenman 2013, Svoboda 2015, Olson 2014 p. 181*ff.*

The first problem with fictionalism is that it necessarily involves deception. To see why this is so, let us begin with a definition:

> To lie is 'to make a believed-false statement to another person with the intention that the other person believe that statement to be true' (Mahon 2016).[95]

Keeping this definition in mind, consider a conversation between Gregor, who is a revolutionary fictionalist, and his sister Grete, who is a moral realist and unaware of the error theory. In what follows, I will proceed as if Gregor is a Joyceian non-assertive fictionalist, but the points will apply to Nolan-style content fictionalists as well.

During the conversation, Gregor says to Grete, 'killing insects is immoral'. As a fictionalist, Gregor is quasi-asserting something which he believes is literally false, and which he is merely 'entertaining'. This fulfils the first half of the definition of lying I just gave, but it is of course to be expected of a fictionalist. But in order for my charge of lying to stick, Gregor must also intend that Grete believe his statement to be true. The problem for fictionalists is that Gregor cannot avoid having this intention. To explain, I will begin with linguistic conventions. One of the background conventions which English speakers share is that sentences with an assertoric surface structure are to be understood as assertions. Joyce himself observes that,

> ...speech acts occur only against a background of conventions shared by a speaker and her audience; a person cannot unilaterally decide that she isn't

[95] This definition is not necessarily entirely platitudinous, and Mahon discusses several objections and possible revisions. But none of the issues raised affect the argument I will make here (not even 'the assertion condition' (see Mahon 2016 §1.5.2), as I will explain below). Thus the definition as given here is respectable for present non-specialist purposes.

> asserting the sentence *S* if she fails to signal this to her audience, all of whom take her to be asserting *S*. (2017 p. 82)

Conventionally in English, the intended mode of speech is typically inferred from the context and subject matter. So, when we discuss the features of the Death Star with others whom we know to be familiar with Star Wars, all parties understand any apparent assertions as expressions of make-beliefs, rather than of genuine beliefs about states of affairs in the real world. Conversely, when we utter a sentence such as 'granite is a type of igneous rock', it is implausible that our audience would typically take us to be to be withholding assertoric force or expect us go on to claim that rocks do not exist when we are in a relevantly more critical context. in the moral case, remember that the vast majority of people are not currently revolutionary fictionalists. Indeed, Joyce's argument for an error theory rests on claims that people are typically moral realists and interpret moral discourse assertorically (see chapter 2 of this thesis).

Returning to Gregor and Grete, this means that as a non-error theorist, under typical circumstances Grete will understand Gregor's apparent assertion that killing insects is immoral as a genuine assertion intended to get her to believe the relevant proposition.[96] And as a competent speaker of English (and a metaethically sophisticated one at that), Gregor knows this. Therefore, if he utters the sentence 'killing beetles is immoral' to Grete without further preamble or explanation, he must intend that she believe him. This fulfils the remainder of the definition of lying given above, and thus Gregor would be lying.

Perhaps Gregor could avoid lying by somehow signalling that his use of moral terms is non-assertoric, or pointing out that he is only quasi-asserting, etc. But this would come at the cost

[96] It is near-platitudinous that assertions are typically intended to produce a belief in an audience. See e.g. Grice 1957 pp. 383-384. (Grice speaks of a speaker 'meaning' something, but for present purposes this is synonymous with their asserting something).

of the all-important immersion which Gregor must maintain if he is to remain a fictionalist (see §4.4.2). For one cannot consistently think and feel as if one is a moral realist while also prefacing one's apparent moral assertions with a metaethical disclaimer about quasi-assertion, entertaining thoughts and so on. Hence if he is to remain a fictionalist, Gregor cannot avoid lying whenever he discusses moral matters with anyone who is not a fictionalist (i.e. currently almost everyone, by Joyce's lights).[97]

Even aside from the idea that truth might have intrinsic value (or, for obvious reasons, the suggestion that lying is morally wrong), this is problematic for two reasons. First of all, most arguments for fictionalism seem to embrace the idea that the fiction which we are to adopt should be essentially a fictional counterpart of the conventional morality. Therefore, lying is likely to be wrong according to the rules of the relevant fiction and thus the fictionalist position threatens to become incoherent. This is because fictionalists recommend that we act and think as if the moral fiction were true, and yet if I am right they simultaneously recommend what amounts to lying – which would be breaking the rules of the moral fiction which they tell us to adopt. And, secondly, it seems likely that most people would agree that there are sound pragmatic reasons for following a general policy of refraining from lying. As Joyce says, 'No policy that encourages belief in falsehoods, or the promulgation of false belief in others, will be practically stable in the long run' (2001 p. 214). Not only is there the likelihood of being found out and facing censure as a result, there is the fact that truth itself is often thought to be instrumentally valuable (see e.g. Lynch 2004 p. 16).[98]

Yet perhaps fictionalists can avoid the deception problem. Although he has not thus far responded to the kind of objection I have raised here, Joyce is sensitive to the possibility of a

[97] Although I am framing the discussion in terms of Joyceian non-assertive fictionalism here, this would also destroy Nolan *et al.*'s advertised benefit of avoiding metaethical digressions (see §4.4.2).
[98] This latter view is admittedly not without controversy – for an opposing view, see Wrenn 2010.

mismatch in background assumptions between speaker and audience, as the quote I gave above shows. And this may point the way to defusing my objection. Joyce observes that recommending fictionalism on an individual basis is rather 'like advising someone to become a rugby team' (2017 p. 82-83), because the fictionalist recommendation to use moral sentences without asserting them requires both speaker and audience to be known fictionalists in order to be intelligible. The proposal is not the kind of thing that makes sense when considered on a per-agent basis. We should therefore interpret Joyce as recommending fictionalism to groups of people, rather than to individuals.

This can be read as forestalling my deception objection because, in recommending fictionalism to groups of people who share an appropriate background of linguistic conventions, Joyce places speakers who speak in the fictionalist mode in the company of others who will understand them to be doing so. Even if what one says would typically be interpreted in a realist sense by the general public, if one is speaking to a group of fictionalists, and all concerned know that they are fictionalists, this misinterpretation would be very unlikely. Thus the deception problem, it could be argued, evaporates.

However, this doesn't actually solve the problem. Rather it simply pushes the problem back one step. For whatever group Joyce expects to take his advice, that group will always come up against outsiders. Given that the community of revolutionary fictionalists is unlikely to be very large, at least initially, the problem which faced the lone fictionalist will quickly return to face the group. When the fictionalist group eventually descends from the metaphorical metaethical mountain, they will immediately be surrounded by, and forced into conversation with, people who typically use moral discourse in a realist fashion. This is, after all, the basis for the moral error theory which gave rise to the WNP in the first place.[99]

[99] See chapter 2 of this thesis.

One way I can see around this is for Joyce to stipulate that the group he intends to make his recommendation to is very large – at least a majority of the population of an entire society. This is clearly not feasible, since it would require an unprecedented and implausibly widespread revolution among huge numbers of people, many of whom may be incapable of understanding fictionalism in the first place, or at least unwilling to make the effort to do so.[100] Thus fictionalism remains either dishonest or unintelligible in the way I have described. Alternatively, Joyce could recommend fictionalism only to small groups, for use exclusively within those groups. Yet unless those groups secrete themselves in monastery-like conditions, which is unlikely to be appealing to many people, the problem will return as soon as they encounter anyone who is not a member of the group.

Moreover, returning to Nolan *et al.*, if they were tempted to try to avoid the deception objection I raised above by adopting a similar 'group orientation', they would in so doing undermine another of their claimed benefits, psychological convenience. For it can hardly be convenient, psychologically or otherwise, to demand that everyone with whom one converses should have a sufficiently sophisticated grasp of metaethics that they are a specific variety of post-error theory fictionalist. And if fictionalists converse knowingly with other fictionalists, the claimed advantage of avoiding metaethical digressions is also undermined – it is not an advantage to avoid explanatory digressions where no digression is needed in the first place.

I conclude that fictionalism necessarily involves deception. This renders the proposal that we respond to the truth of error theory by becoming fictionalists problematic in several ways, and I do not believe that error theorists will be willing to embrace a response to the 'what now?' problem which forces them to become liars. This may be enough by itself to convince

[100] I will respond to a similar challenge to my own proposal in §7.4.

us that we should look to other WNP responses as the way forward. But just to be sure, I will now discuss a further problem which further undermines fictionalism as a WNP response.

4.4.4. Problems with fictionalism: ii) Which Morals?

The second problem with fictionalism I will discuss is one I mentioned earlier (§4.4.1), namely the issue of deciding which moral quasi-beliefs we should have (hereafter I will refer to this as the 'which morals?' question). This is a problem for Joyce more than for Nolan *et al.*, and he acknowledges that it something he needs to address, writing,

> There are different ways of understanding the claim that X is useful to a group, even before we get to more specific questions raised by replacing "X" with "morality." Let us suppose that we settle on one such way. If a group is motivated to adopt morality as a fiction because doing so is useful (in the manner settled upon), then when faced with the choice of *which* moral fiction to adopt (from an infinite range of possibilities), the answer is simply "The most useful one." It is important to remember that the fiction is being maintained for practical purposes; it is entirely possible that a group might adopt the wrong moral fiction. (2017 p. 83)

There are a number of problems here. First, there seems no clear way to distinguish *moral* fictions from any other fictions which may be deemed useful. This is important, because fictionalism demands that we treat moral quasi-beliefs very seriously, immersing ourselves almost entirely in the thought that they are true and incorporating them into our practical deliberations. But which fictions do we adopt this involved, dedicated attitude towards, and which do we treat as mundane fictions which require no deeper immersion on our part, like the rules of a fictional game in a novel? Perhaps it might be suggested that the only issues

which would feature in our moral quasi-beliefs would be those about which we formerly (i.e. before we came to accept error theory) had moral beliefs. So we could sort through our prior moral beliefs, and retain in fictional form those which would be most useful.

But that seems as though it would be too restrictive. Society is continually changing, yet if fictionalism was restricted to some subset of whatever moral beliefs we had before we became error theorists, there would be no way for the fictive morality to evolve or to reflect novel moral questions. For example, just as two hundred years ago there could have been no concept of the moral issues around, say, online communication, as 'fossilised' fictionalists we would have a hard time accounting for new situations which obviously had a moral dimension. Yet without some way of picking out what makes a fiction a *moral* fiction, how could we sort out moral fictions from other fictional rules which were merely useful?

It seems potentially useful to institute a fictional rule whereby every individual in a group who can afford to do so must buy a tool – a hammer or a screwdriver or somesuch – for another randomly chosen member of the group. Tools are very useful things, so having a group-wide, tool-based secret Santa scheme should be useful, too. Thus the relevant rule would be very like a fictionalist 'thought'. We may act as if it were binding upon us, it might nudge us into action when otherwise we might have been insufficiently motivated, and we might react poorly to those who fail to act according to the rule. We could do all of this while knowing, when we were thinking critically about it, that the rule was a fiction - it was not really a rule which was binding upon us in any objective sense. But would we really think of such a fictional rule as a *moral* rule? Surely it would be going too far to claim that we would. Joyce owes us an account of what counts as a moral fiction, with the attendant requirement that we take it seriously and act at almost all times as if it were true, as opposed to any other kind of fiction.

A reasonable response from fictionalists here might be to say that we should consider 'moral' whichever fictional rules play a particular 'moral' role in our deliberations, in whichever way that role might be specified – for example in terms of reactive attitudes such as blame, guilt, censure and so on.[101] But I would reply that we would still need an account of this (or whatever alternative fictionalists might suggest). For example, if lots of people have an intense 'moral-style' reaction to something which is not useful, where does that leave us? There are numerous instances of 'moral panics' in which some phenomenon triggers widespread, apparently moral reactions which frequently seem highly unlikely to be useful in any obvious way to society or to the individuals involved.[102]

The second problem, and I believe a far more significant problem, is that Joyce gives the impression that an answer to the 'which morals?' question can be found. Yet I believe there may be no satisfactory answer. I find Joyce's formulation in the quote above a little unclear, since it seems to conflate usefulness *for accomplishing a goal* with the question of *which* goal(s) it would be useful to accomplish. The reader may have no such issue with Joyce's formulation, but in an attempt to be a little clearer (even if only for my own benefit), I suggest that we can break the 'which morals?' question down into the following two sub-questions without changing what Joyce is driving at:

A) What do we want to get out of adopting morality as a fiction?

B) Which moral fiction would be the best means to that end, if we adopted it?

The reason I am not optimistic that there can be a satisfactory answer to the 'which morals?' question is that I see no satisfactory way of answering A) or B). Let us consider A) first. In §4.4.2 I noted that Joyce focuses on overcoming weakness of will as the primary thing we will

[101] I will return to the theme of reactive attitudes in §5.4.3 when describing my own proposal.
[102] The classic text on moral panics is often thought to be Cohen 1972. For an insight into more recent views on moral panics, see e.g. Cree 2016.

get out of adopting fictionalism, but this is a model – the above quote from 2017 clearly shows that this is not the only way in which adopting fictionalism is supposed to be useful to the group of fictionalists. A much fuller answer to A) is required here. It initially seems that a satisfying answer to A) might involve such things as a more harmonious society, a reduction in how frequently people do things which harm others, and other things along similar lines. But why should we want those things?

To illustrate, imagine someone (let's call her Caroline) who is of a certain Nietzschean bent and believes that humanity in general would be best served by denouncing pity and revelling in conflict, in order to move towards a stronger, more glorious future. Caroline's answer to A) reflects these sentiments, and so what she wants out of adopting morality as a fiction is more strife, more adversity, and more consequent opportunities for triumph. On what grounds can we tell Caroline that her answer to A) is wrong? Certainly error theorists can offer no *moral* grounds for criticising Caroline's position (as most people who have conventional moral beliefs might). And on a pragmatic level, there seem to be few grounds on which to say that Caroline is conclusively wrong either. It seems to me that whatever answer we come up with for A), there will be a corresponding character we can come up with to play Caroline's role and undermine it.[103]

Recall from §3.3.7 that Joyce is committed to the claim that there is no single desire which is shared by all people. Given our disparate desiderative starting points, there is nothing which we all want to achieve. This means that there can be no particularly specific answer to A). And Caroline's example shows us that even broad answers to A) will be unsatisfactory to someone. Joyce considers something like this problem, and replies

[103] My point here is that no list of 'Carolines' would be exhaustive. But to flesh the suggestion out a little, further examples might include some kinds of communist who would want to abolish private property, and some neo-Luddites who would place environmental concerns above all others.

> Perhaps taking a fictional stance toward morality will recoup costs for one person but not for another. Even the best advice is unlikely to be good for anyone in any circumstances. In light of this, the revolutionary fictionalist should be permitted a degree of modesty and a dose of vagueness: The position is reasonable if it's good advice generally for most people. (2017 p. 82)

This philosophical shrugging of the shoulders is inadequate. The implication is that whatever a majority of people think is good should be pursued, regardless of whether a minority of people consider it harmful. Clearly Joyce's 'advice' is premised on a form of consequentialism – good advice is good insofar as it promotes a net improvement in outcomes for a majority of people. But why should we grant Joyce's claim that adopting revolutionary fictionalism is good advice if it's good for most people, most of the time? The claim seems to run contrary to one of the plausible benefits of morality - that it defends those who would otherwise be subject to the tyranny of the majority. In fact, Joyce's recommendation seems to be premised on a form of consequentialism, and it is precisely an ethical consequentialist who made one of the most famous moral arguments against the tyranny of the majority – J.S. Mill in *On Liberty*. This is not the place for a discussion of consequentialism in ethics, but note that it is a significant disadvantage of Joyce's view that it strongly threatens to inherit some or all of the issues around consequentialism, many of which remain very controversial.[104]

One way Joyce could respond to this is to argue that a group of fictionalists could stipulate that one of the fictions to which they will commit is that minorities should not be harmed by the tyranny of the majority. But this just further complicates the search for an answer to A). Are we to seek answers to A) which are optimally useful to a majority within a group, as Joyce

[104] For an overview of consequentialism in ethics and the problems it raises, see Sinnott-Armstrong 2019.

explicitly says, or not? If so, is optimal usefulness to such a majority the same as not harming minorities? It certainly seems unlikely that it is, and it seems equally unlikely that this rather *ad hoc* response will lead to a satisfactory answer to A).

Turning to B), there seem to be important empirical questions with no clear answers. Even if we were to be able to overcome the above problems with A) and settle on an answer to it, how are we to know which moral fiction will contribute best to bringing it about? Again, there are some initially plausible answers to B) which, on closer inspection, turn out to be problematic. For example, consider an answer to A) along the lines that, as far as possible, people should not be harmed by the actions of others. This seems to have some quite obvious implications for our answer to B), such as implying that it should include a rule against, say, stealing. But does such a rule really contribute to minimising harm to others? This is an empirical question, and although we will be inclined to answer in the affirmative, there seems to be no way of knowing for sure that it does not in fact merely promote wealth inequality, which breeds conflict and disadvantages the poor. It seems likely that an analogous example can be found for virtually any proposed moral rule. As I discussed in §4.3, this is the foundation of the arguments put forward by abolitionists. While it may be a disadvantage of abolitionism that it ultimately rests on an empirical claim which may not be provable, we can now see that fictionalism rests on just such a claim as well.

Underlying all of this is a sense in which I believe Joyce falls foul of *status quo* bias. Despite his insistence that the moral beliefs we should take a fictive stance towards should be whichever are most useful, Joyce often speaks of the fictionalist habit in terms of continuing our pre-error theory moral practice or of adopting 'morality' (as opposed to 'parts of morality' or 'certain moral beliefs') as a fiction without reference to any further discrimination (see e.g. 2001 chapter 8 *passim*). Yet it seems impossible that the most useful moral beliefs are the ones we had before error theory came along. Even if there were a credible evolutionary basis

for some traditional moral beliefs, we do not have to exert very much effort to come up with widely held yet mutually exclusive moral beliefs, hence for example the intractability of conflicts around abortion. And any attempt to sort out the evolutionarily justifiable moral beliefs from the rest inevitably leads straight back to the issues with the search for an answer to A).

Moreover, entirely apart from whether they could be mustered into anything approaching a coherent set with no contradictions, abolitionists have shown us that the moral beliefs the majority of people have may well be responsible for disastrous events. Not only would it be question begging against abolitionists to dismiss the possibility that traditional morality is actively pernicious, doing so cannot be supported by evidence. While fictionalists may be correct if they were to argue as I have done that abolitionists cannot empirically prove their case, neither can opponents of abolitionism empirically prove abolitionists wrong.

Surely part of the point of error theory is that morally speaking, all bets are off. One might suspect that mere usefulness is not the only criterion by which we should judge responses to the 'what now?' problem.[105] But even ignoring that question, for now, if we follow the fictionalists in their insistence that usefulness is enough, then that leads inexorably to questions of what is and is not useful which cannot be answered.

Taking §4.4.3 and §4.4.4 together, I conclude that fictionalism fails because as a general strategy it cannot plausibly be described in satisfactory detail or implemented. Once again, we will have to look elsewhere for a convincing WNP response. In the next section, I will

[105] For example, perhaps it might be argued that experiencing things we find beautiful is psychologically important to us, and therefore something we have prudential reasons (which are compatible with the truth of a moral error theory) to want. Yet usefulness and beauty are not obviously the same thing, and nor is it obviously *useful* to us to contemplate beauty, even if it is psychologically important.

discuss a handful of further responses to the 'what now?' problem, including one which has emerged as a noteworthy candidate for answering the WNP only comparatively recently.

4.5. Revolutionary expressivism

This is the last of my sections discussing why error theorists should reject the WNP responses which have appeared to date. Other than conservationism, abolitionism and fictionalism, few responses to the 'what now?' problem have emerged as yet. Despite its appearance in some literature (e.g. Joyce 2001 p. 214, Cuneo & Christy 2011 pp. 93-94, Husi 2014), I will set one possible response aside immediately. So-called propagandism, according to which a metaethically learned elite would keep the truth of error theory a secret, and let the moral 'proletariat' carry on unaware of their profound errors, is unviable for a couple of reasons.[106] One, despite several philosophers mentioning propagandism, none has yet made a serious attempt to defend it. Until someone does so, it would be unfair and too much of a distraction from the task at hand to concoct a straw man form of the view and put it into the mouth of an as-yet fictional opponent. Two, the 'secret' of moral error theory is already out, and has been since at least 1977. Moral error theory may not (yet) be the talk of taverns across the land, but the attention it has already attracted makes it impossible to keep it secret from anyone interested in the topic.[107]

---

[106] An analogy may be drawn between propagandism and 'Government House Utilitarianism' in the works of Sidgwick and Hare (see e.g. Williams 1985 pp. 108-109 for explanation and criticisms). Joyce mentions this in passing (2001 p. 214) and Blackford suggests that Mackie may have had something like G. H. U. in mind (2019 p. 62), though declines to elaborate.

[107] Variations of a further view have been suggested by Matt Lutz (2014), who calls it substitutionism, and Stan Husi (2014), who calls it revisionism. The idea is that we might respond to the WNP by reforming our moral thought and discourse to remove the erroneous features while retaining the rest. I omit further discussion here because Lutz and Husi describe variations on a strategy according to which completely different WNP responses as disparate as my own revolutionary relativism and revisionary expressivism (see below) would both count as substitutionist/revisionist. I therefore consider 'substitutionist' and 'revisionist' labels for certain kinds of WNP response rather than fully fledged responses in their own right, in a similar way to how 'realist' can be used to describe very different and mutually exclusive forms of e.g. naturalism and non-naturalism. As such, I will set Lutz and Husi aside and concentrate on more contentful views.

Of the notable WNP responses to date, this leaves just one to discuss in detail here. I will examine a recent paper, the author of which suggests that if we accept error theory, we should become expressivists.[108] In my view, the attempt to recommend expressivism is either

[108] A similar proposal was also made by Köhler and Ridge (2013). Despite their insistence to the contrary, I believe it is self-undermining, and Svoboda's discussion is therefore more useful here. Despite lacking the space for a proper discussion, I will attempt to give an exceedingly brief reconstruction of Köhler and Ridge's (complex) argument, and show why I either see no reason for *moral* error theorists to adopt it, or see it as potentially self-defeating. Their argument, which is about a general error theory about all normativity, rather than just morality, runs roughly thus:

If a Normative Error Theory (NET) is true, we should become revolutionary expressivists. This suggestion could be seen as self-defeating (p. 429). This is because it contains a 'should' clause which appears to rely on a practical form of normativity ruled out by the NET. The authors argue that this pitfall can be avoided via the following analysis of the situation:
1. Practical normative thought and discourse serve important functions.
2. Either
2a. We have no choice but to care about preserving those functions.
Or
2b. It is in our interests to preserve those functions (p. 435).
3. Of the available theoretical options, revolutionary expressivism preserves those functions best.
4. Therefore, of the available options, we should adopt revolutionary expressivism and reform our definitions of normative terms accordingly.
5. This 'should' must be read in a functional sense, similar to '*x* is required in order to allow *y* to perform its function'. This is roughly analogous to the way in which one might say that blunt knives should be sharpened in order to allow them to perform their definitive function of cutting (p. 434). This avoids reliance on the perhaps more intuitive practical sense of 'should' in '...should become expressivists', and so avoids self-defeat.

My problems with this strategy concern 2a, 2b and 5. If 2a is true, then surely the best way to preserve the functions in question is, where possible, to leave practical normative thought and discourse as they already are. This is not an option for error theorists about practical normativity in general, which is how the authors characterise their NET. But it is an option for *moral* error theorists who have no issues with accommodating non-moral 'oughts' which are predicated on our desires or interests and which can perform the desired functions. If 2b is true, and adopting revolutionary expressivism is something we would need to *do* (which surely it must be), then 2b constitutes a claim about a practical sense of normativity ruled out by Köhler and Ridge's NET. The argument is therefore self-defeating.
This flows into the issue I have with 5, which is that reforming the definitions of normative terms, just like knife sharpening, does not happen in a vacuum or by itself. Even if something is required in a functional sense, some agent is still required to bring it about. This is clear if we continue with the knife analogy. Given that we as agents have no definitive function (i.e. there is nothing that we are *for doing* in the way that knives are *for cutting*), there is nothing we should – in a functional sense – do. Therefore, even if there is a functional sense in which knives should be sharpened, there is no *functional* sense in which *we ought to sharpen them*. I find it plausible to suggest that, consistent with the arguments for error theory discussed in previous sections, moral error theorists would agree that we have an authoritative reason to sharpen a blunt knife only insofar as e.g. we desire to cut something, and believe that a sharp knife will help us do so, while a blunt knife will not. This is clearly an instance of practical normativity, and by analogy, once again Köhler and Ridge's argument turns out to be self-defeating.
I do not take the above to necessarily demolish Köhler and Ridge's view. But it should suffice to illustrate that their view is not without controversy, and that given the space constraints on this project, a less contentious revolutionary expressivist view is a better subject for discussion here.

inappropriate, given the commitments of contemporary error theorists, or inadequate as a response to the problem at hand. That being the case, I argue that we should not adopt revolutionary expressivism. In the next chapter, I will begin to outline my own proposal, revolutionary relativism, and then move on to show why error theorists should adopt it in the light of the criticisms of the foregoing theses.

In 'Why moral error theorists should become revisionary moral expressivists' (2015), Toby Svoboda argues on prudential grounds for a non-cognitivist, expressivist response to the WNP.[109] That is, he proposes that if we accept the error theory, we should nonetheless continue to use moral discourse because it is useful. But we should revise our understanding of moral judgements and discourse and view them not as beliefs and expressions thereof, but as desires or desire-like attitudes (such as approval or expectation) and expressions thereof instead. This, he claims, will allow us to gain the benefits of moral thought and discourse while avoiding error.

Svoboda argues that revolutionary expressivism is superior to abolitionism, conservationism and fictionalism 'in three different contexts: how these positions fare in avoiding moral error, how they fare in securing intrapersonal benefits of morality, and how they fare in securing interpersonal benefits of morality' (2015 p. 8). If he is right, he takes this to show that moral error theorists should adopt revolutionary expressivism on the grounds that it will better deliver the stated benefits (which are things that most people desire) than competing theories.

[109] I will refer to Svoboda's view as revolutionary as opposed to his preferred term, revisionary. This is because 1) it is more consistent with the foregoing sections of this project and thus avoids introducing a new term for little gain, and 2) because revolutionary better captures the idea that this is a post-error theory proposal as opposed to a form of expressivism about conventional morality. Nothing turns on the distinction here, and anyone more sympathetic with Svoboda's phrasing can just replace revolutionary with revisionary where appropriate.

I take the arguments I gave in previous sections to count decisively against Svoboda's chosen opponents, and will therefore forego discussion of his comparisons between the various theses. Rather, I will appraise whether revolutionary expressivism can deliver the benefits Svoboda claims. Setting aside the criterion of avoiding moral error (since avoiding moral error would seem to be a requirement of all post-error theory recommendations), I will discuss the purported intrapersonal and interpersonal benefits of revolutionary expressivism.[110] I will show that Svoboda's proposal fails in both areas. I believe that Svoboda's lack of success in those contexts means his response to the 'what now?' problem fails on its own terms, making it unnecessary to consider any further contexts which might be thought appropriate. Before getting into the arguments though, I would like to pause briefly to correct an inaccuracy in Svoboda's terminology.

4.5.1. A note on terminology

When he claims that moral error theorists should become 'revisionary moral expressivists', the latter instance of the word 'moral' is inappropriate. Recall from chapter 2 that moral error theorists typically assume the truth of cognitivism, i.e. the view that moral judgements are beliefs, as a description of pre-error theory moral practice. Expressivism, being a non-cognitivist view, entails that moral judgements are not beliefs. Therefore, according to the typical commitments of error theorists, the phrase 'moral expressivism' is oxymoronic, since 'moral' precludes expressivism.[111]

[110] With the possible exception of conservationism. But recall that one of the problems with conservationism is precisely that it implausibly relies on the viability of continuing to have erroneous moral beliefs (see §4.2.1).

[111] It might be suggested that this raises the interesting question of whether cognitivism could be seen as another non-negotiable commitment of moral discourse, alongside objective prescriptivity/categoricity. But beyond my observations in the above text, I leave it to others to consider that question.

Svoboda acknowledges that this accusation could be levelled against him (*op. cit*. p. 18), but claims that being a revolutionary moral expressivist 'is consistent with supposing that moral error theory is descriptively true, for that truth may simply reflect contingent features of our current morality rather than necessary ones' (p. 19). This is a misunderstanding of the situation. It is true that in metaethics in general, cognitivists must begin by accepting that moral judgements may not be beliefs, in order to avoid begging the question against non-cognitivists. Thus 'being beliefs' cannot be said to be a necessary feature of moral judgements at the start of debates between cognitivists and non-cognitivists.

But moral error theory is not a position in any such debate. Rather, the best developed forms of moral error theory currently available proceed from an assumption that the debate between cognitivists and non-cognitivists has already been settled in favour of cognitivism. Indeed, one might think that moral error theory as defended to date depends on cognitivism for its coherence – if moral judgements were anything other than truth-apt beliefs, how could they be systematically erroneous? Therefore unless someone comes along with a radically novel moral error theory which is compatible with the truth of non-cognitivism, one cannot consistently be both a moral error theorist and a *moral* expressivist.

This does not necessarily derail Svoboda's thesis, however. It remains open to him to simply drop the latter 'moral' from his title and/or substitute some other appropriately defined term (e.g. 'schmoral' or moral*). Revolutionary expressivists could then recommend that we replace e.g. moral discourse with schmoral discourse, which is to be understood in expressivist terms, and which delivers the same purported benefits as moral discourse. Thus even if, properly speaking, Svoboda's recommendation is a form of abolitionism (in that it recommends replacing moral discourse with something which is not *moral* discourse, given the commitments of typical error theorists), an expressivist understanding of putatively moral

vocabulary may still deliver the benefits he claims for revolutionary expressivism. Therefore I will move on to consider what he says despite the terminological issue.

4.5.2. Problems with revolutionary expressivism: i) Intrapersonal

In the intrapersonal (i.e. motivational) context, Svoboda holds that conventional morality is beneficial because 'it bolsters one's commitment to act for certain ends, increases one's self-control, and helps overcome weakness of will' (p. 20). Revolutionary expressivism, he claims, can deliver these benefits because, like expressivism in general, it allows us to

> understand moral judgements as desire-like attitudes that have inherent motivational force. For example, if the moral judgement that lying is wrong is understood as a desire-like attitude with respect to lying (e.g., disapproval of it), it is easy to see why the person making the judgement would be motivated to some degree not to lie. (*ibid*.)

We might say that in expressivist terms, an agent who has judged that lying is morally wrong has thereby made a sincere, considered judgement about lying, to the effect that they disapprove of it. Their judgement just is this sentiment. Giving in to weakness of will represents doing something which runs contrary to that sincere, considered judgement, usually for short term gain (and often despite, or without considering, the long term costs). An expressivist account of moral judgements can help to show how we overcome such instances of weakness of will, because on expressivism, moral judgements – being desire-like attitudes – automatically include or bring with them motivations to act in accordance with the

judgement made.[112] Simply put, if you don't like the idea of anyone lying, you won't want to lie yourself. This, I take it, is the essence of Svoboda's argument here.

My response is that yes, we can do as revolutionary expressivists urge and resolve henceforth to interpret 'moral' judgements as desire-like attitudes (hereafter DLAs). While expressivism may be incompatible with error theory as typically defended for the reasons I noted above, in the context of responses to the truth of error theory, all previous bets are off. This is because in the latter context, metaethical theories do not necessarily have to account for the features of conventional moral thought and discourse, but can instead concentrate on how moral thought and discourse might be reshaped in the light of error theory. However, it is highly controversial whether non-cognitivists of any stripe can help themselves to the notion that DLAs have inherent motivational force.[113]

But let us be charitable and grant Svoboda that when we make a moral judgement we are automatically relevantly motivated at least to some degree. Let us also grant that moral judgements are considered judgements which we make after proper deliberation, whereas temptation simply 'crops up'.[114] I still maintain that Svoboda's proposal fails to secure the benefits he claims.

In the case of weakness of will, if the sentence 'stealing is wrong' is understood as expressing a DLA such as a desire that people do not steal, consider what happens when one is tempted to steal. It seems to me that being tempted to steal is most clearly understood as a *desire* to

[112] See e.g. Blackburn 1998a (especially chapter 3), Gibbard 1990. For further discussion & analysis, see Toppinen 2015.

[113] See for example Smith 1998. Among other things, Smith points out that it is possible to judge that one ought to do something, but at the same time to feel no motivation to do it, for example when one is depressed (p. 161).

[114] I do not necessarily consider this latter concession plausible, not least since it seems to me that moral judgements are often made in the heat of the moment. But it strengthens Svoboda's argument against a possible line of criticism according to which moral judgement seems to require something more significant than a snap decision about how we feel (*cf.* Miller 2013 §3.6).

steal. It may not be a desire to steal which is regularly felt, and it may not be the product of the kind of thorough deliberative process whereby sincere, considered moral judgements might be reached. It may not be a rational desire. It may not be a desire which we desire to have. But temptation in a situation like this is nonetheless most clearly understood as a desire *of some sort*.

If that is correct, then being tempted to steal when one has previously judged (according to the expressivist account) that stealing is wrong represents at best having two DLAs which contradict one another.[115] This results in a motivational conflict. We have two DLAs pulling us in opposite directions, each with an attendant motivational force. While there may be interesting reasons to divide DLAs into various categories (moral versus non-moral is one which springs to mind), in this case of weakness of will, it is only the motivational force of a DLA which matters.

Remember that by Svoboda's own lights, moral judgements are supposed to help with weakness of will because they include motivations to act in what we judge to be morally appropriate ways, rather than give in to temptation. Yet this will only happen if the motivation involved in a judgement that stealing is wrong is somehow stronger than the motivation involved in being tempted to steal. If that is the case, then the moral judgement will mean that we do not give in to temptation, because our strongest relevant motivation is against stealing. But if the motivation involved in the temptation to steal is stronger than the motivation involved in judging that stealing is wrong, then that judgement will not prevent us stealing.

[115] I say 'at best' because at worst, temptations threaten to partially or even entirely reverse our previously held DLAs and thus our moral judgements, especially in cases where we eventually give in to temptation. I leave this as a footnote because the scenario I go on to describe in the main text is slightly more charitable yet still, I argue, undermines Svoboda's proposal.

Perhaps if there were some way to rank the motivational force of different types of DLAs, Svoboda might be able to show that moral judgements somehow come out on top in motivational terms, while the desire represented by an instance of weakness of will always loses out. But I am skeptical whether any such ranking system would be plausible. I suspect that the only criterion by which we can judge the motivational power of a DLA is the strength of the DLA in question, at least in cases without complicating factors such as depression. Rather, the motivational efficacy of a DLA and the strength with which it is held seem to be intertwined, or even to be the same phenomenon.

Revolutionary expressivists could reply to this by clarifying their proposal: just as hermeneutic expressivists may claim that our moral DLAs can be very strong (Blackburn's 1998 book is, after all, entitled 'Ruling Passions'), revolutionary expressivists could argue that in the WNP context, we should come to have very strong DLAs about traditionally moral matters. This will not help, however. We cannot simply choose to strongly desire anything. For example, most people would likely agree that it would be prudentially highly beneficial to have an overriding desire to eat healthily. But we cannot just start having such a desire, that simply is not how desire works.

Thus revolutionary expressivists face a dilemma. Either our post-WNP moral judgements are to be DLAs of arbitrary motivational efficacy, and thus potentially be of little or no help in resisting weakness of will, or our moral DLAs are to be very strong, in which case we cannot decide to have them (and thus very possibly cannot choose to become revolutionary expressivists at all).

Svoboda might object that I am framing the discussion as if revolutionary expressivism has to make submitting to weakness of will *impossible*, when all it needs to do to be worth recommending is give us an analysis of 'moral' judgements such that they *count against* giving

in to temptation. This, he might claim, is sufficient for post-error theory 'moral' discourse to serve the same regulatory purpose as conventional morality. Or perhaps all that is needed to recommend revolutionary expressivism is that it serves the relevant purpose *to some extent*, albeit less well than conventional morality did.

I do not accept this potential objection. What I am arguing is that sufficiently strong temptations will *always* win out over moral judgements. This falls a long way short of substantiating Svoboda's claim that revolutionary expressivism can deliver the intrapersonal benefits of conventional morality.[116] But worse than that, Svoboda's view threatens to ensure that agents' moral judgements coincide with what they most desire (and are therefore motivated) to do. One of the issues with expressivism as a descriptive thesis about our current moral practice is that it must account for the fact that moral judgements seem to be somehow more significant than merely expressions of our preferences.[117] But this may or may not be the case in a post-error theory context - perhaps 'moral' judgements in this context need to be more significant than expressions of our preferences, perhaps they do not. It is incumbent upon Svoboda to provide reasons for deciding the matter one way or another. But Svoboda does not address this issue. In the absence of an account of why 'moral' judgements should be anything more than preferences, revolutionary expressivism threatens to reduce 'moral' judgements to the kind of arbitrary judgements which would be the opposite of a bulwark

[116] While I do not wish to make a full-blooded argument here, one might point to Blackburn 1998a, p. 191, where Blackburn argues that acting contrary to a prior (expressivist) moral judgement can make us realise what our motivations actually were, contrary to what we thought they were, when we made the judgement. Thus an argument could be made that on expressivism, succumbing to temptation may actually tell us retrospectively that we should have judged it right to do as we were tempted. It seems possible that many people would find this unsatisfying in comparison with the 'bulwark effect' provided by traditional morality. It would be an excessive digression to investigate here whether an argument along these lines could succeed, however, and I believe the points made in the main text above will suffice for present purposes.

[117] See e.g. Miller's discussion of what he calls the 'moral attitude problem', Miller 2013, p. 39ff, especially footnote 11.

against weakness of will. Rather, it is in danger of becoming a justificatory framework for giving in to temptation.

4.5.3. Problems with revolutionary expressivism: ii) Interpersonal

In the interpersonal context, Svoboda cites the ability to track normative disagreement as conventional morality's main benefit (p. 21).[118] He claims that it is an advantage of revolutionary expressivism that it 'preserves the useful feature of accounting for normative and evaluative disagreement, because it can track these as attitudinal divergences' (p. 22). I question whether the ability to track normative disagreement is a benefit in the sense required here.[119] It is a *theoretic* benefit if a given moral theory can account for moral disagreement. For any theory which offers no way of understanding what is going on when people disagree about first-order moral matters (which they certainly seem to do) will be at a serious disadvantage in comparison with competing theories. Otherwise, there will be an obvious phenomenon which the theory is at a loss to explain. There will therefore be a natural motivation for expressivists in general to show how expressivism can give an account of moral

[118] Svoboda also cites morality's ability to allow for moral reasoning. I omit this facet of his argument for three reasons. First, because he concedes that his opponents can also account for moral reasoning, and that conservationism probably fares better in this regard. It is therefore not so much a benefit of his theory, as a cost his theory avoids incurring. Second, I have doubts that the ability of revolutionary expressivism to facilitate moral reasoning is of genuine benefit on similar grounds to my doubts about the tracking of normative disagreement (see below), and my argument about that can readily be applied to the moral reasoning case. And third, Svoboda's argument concerning moral reasoning rests on there being a satisfactory solution to the infamous Frege-Geach problem. While it would be question-begging to assume that there is no such solution, Svoboda still arguably owes us a proper demonstration that a solution which supports his moral reasoning argument *has been* found. Without such a demonstration, certain aspects of the arguments for revolutionary expressivism must be somewhat provisional, pending a satisfactory solution to the Frege-Geach problem. Given that he does not provide a demonstration (which would surely be impossible within the scope of his paper, given the complexity of the issue) other than pointing to Blackburn's (still highly controversial) view, I find Svoboda's moral reasoning argument to be a theoretical possibility, rather than a compelling argument in its own right.

[119] It is also controversial whether expressivists can make sense of moral disagreement in the way Svoboda suggests. For examples, see Merli 2008 or Ridge 2013. Both papers criticise previously influential expressivist accounts of disagreement. Ridge offers a new analysis of disagreement which is compatible with expressivism, but we should be wary of considering him to have thereby 'solved the disagreement problem'.

disagreement. But that does not mean that an expressivist account of moral disagreement is necessarily of *instrumental* use or benefit in a post-error theory context.

In the WNP context, the usefulness or otherwise of apparently moral thought and discourse must be understood in terms of what contribution they make to our wellbeing and to securing goods which we desire or are in our interests. Yet seen from this angle, all revolutionary expressivism allows is that two people who obviously appear to disagree can be said to disagree in an intelligible way. Svoboda owes us an account of why such a feature is *useful* – why it is conducive to our wellbeing or serves our ends.

It might be suggested that what Svoboda has in mind is the idea that once we have made the substance of our moral disagreements explicit – once we can get a clear picture of what disagreements are *about* - we will be better equipped to find solutions to them. Conversely, if we have no intelligible account of 'moral' disagreement, we are unlikely to make progress in solving disagreements because we will lack a clear picture of what we are even disagreeing about, yet alone how we might resolve our differences. If true, this would indeed seem to be beneficial, since resolving moral disagreements seems to serve our interests by facilitating better coordination between members of society. If we can resolve our disagreements rather than talk past one another, we can get on with more productive tasks. So it seems that being able to account for moral disagreement is a useful feature of any response to the 'what now?' problem in the required instrumental sense.

The problem with this is that an expressivist understanding of moral disagreement, while possibly beneficial in comparison to having no analysis at all, falls short of delivering the same extent of benefit that a (cognitivist) realist understanding does. The key here is to understand that there is more than one account of disagreement at work. If two people disagree about whether e.g. torture is morally wrong, then according to a realist analysis of the disagreement,

there is a fact of the matter about the moral status of torture, and only one party can be correct about it. That being the case, the dispute will be resolved when both parties come to agreement in what they believe the fact of the matter to be.

On an expressivist account of the same disagreement, the disagreement is not about having mutually contradictory beliefs. Rather, the disagreement is in attitudes – one (e.g.) approves of torture, while the other disapproves. It is up for debate whether this counts as disagreement at all. But to claim that it does not count as disagreement here, without considerable digression, would be to beg the question against expressivists.[120] So I will grant that both disagreements about facts and disagreements about attitudes are forms of genuine disagreement.

Nonetheless, this shows that when Svoboda claims that revolutionary expressivism 'preserves the useful feature of accounting for normative and evaluative disagreement' (2015 p. 22), this is slightly misleading. The phenomenon of moral disagreement which occurs on a cognitivist account (i.e. the account of conventional morality assumed by error theorists) is factual disagreement. This factual sense of disagreement is not preserved by a revolutionary expressivist account, rather it is replaced by an attitudinal sense of disagreement which, presumably, Svoboda takes to be equivalent to factual disagreement in terms of how useful it is as an analysis of moral disagreement.

But just because both factual and attitudinal disagreements are forms of genuine disagreement, it does not follow that both analyses of moral disagreement are on a par when it comes to how useful they are – i.e. what degree of instrumental benefit they would yield in a post-error theory context. Moral realism is useful in this sense because when moral

[120] See Jackson 2008 for a discussion of the differences between factual and attitudinal types of disagreement, and an explanation of why some non-cognitivists are committed to defending the idea that attitudinal disagreements *really are* disagreements.

disagreements are understood as differences in beliefs about matters of fact, there is the possibility that the process of disagreement might help parties agree on what the facts are, and so resolve their disagreement. This result is plausibly beneficial because after resolving their disagreement, it seems likely that the parties will be more willing to cooperate, or at least cease to hinder one another.

On the other hand, if we view 'moral' disagreements as parties having differing DLAs, this implies that there is no fact of the matter for them to discover. True, it may be that appreciating the sentimental nature of the disagreement might pave the way for the parties coming to sympathise with one another, and so resolving the disagreement. But it is not plausible that this offers the same degree of benefit as factual disagreement. Factual disagreement offers no guarantees – some people still claim that the world is flat. But it seems reasonable to think that someone offering you what you take to be convincing epistemic reasons to hold a certain belief will be more able to get you to change your mind than someone trying to convince you to feel a DLA you are not otherwise inclined or disposed to feel.

The upshot of this is that revolutionary expressivism fails to preserve the sense of moral disagreement which conventional morality allows for. Rather, revolutionary expressivism involves replacing the conventional model of moral disagreement with a different model, which it is less certain can yield the instrumental benefit of disagreement resolution which Svoboda claims (or at least implies) it can. I do not therefore conclude that revolutionary expressivism yields *no* relevant kind of benefit, but that it is plausibly of significantly less benefit than Svoboda advertises. I will return to this later in §6.5.1 and §7.1, and so for the time being will leave the matter at that.

Given the above, and along with the necessarily provisional nature of the position, I reject Svoboda's proposal that error theorists should become revolutionary (moral) expressivists.[121] And since the issues with Svoboda's proposal I have discussed stem from the expressivist nature of the proposal rather than from small details which could easily be tweaked, I doubt that any other expressivist recommendations will fare much better as responses to the 'what now?' problem.

4.6. Conclusions, and the lessons to be learned so far

I began this chapter by arguing that accepting a moral error theory means that we face a problem about what to do next. Quite how pressing a problem this is for error theorists has been somewhat overlooked, but in section 4.1 I outlined the problem in more detail than has appeared in most of the relevant literature, and underlined how important it is that we find a satisfactory solution to the WNP. The past few sections have concentrated on analysing the existing responses to the WNP, and in each case I have argued against the position in question. Many of these arguments were, I believe, original. One reason for this is that at this point in my project, we are running up against the limits of the debates around this issue to date – broadly speaking, whatever else needs to be said on the matter has yet to be written. In this section I will consider where all this leaves us. If my arguments in sections 4.2-4.5 were right, then in the WNP we still face an important problem and yet we lack an adequate response to it. Furthermore, the problems I identified with previous WNP responses were often systematic – i.e. an improved version of any of those responses is still unlikely to be convincing, because the problems in each case are 'built in' to the fabric of the response. This means that in order to overcome the WNP, we will need something radically different.

[121] As I touched upon in footnote 118, if they are to avoid at least some facets of their arguments being rather provisional, revolutionary expressivists also need to find DLAs with the required logical features to avoid the Frege-Geach problem, which is a very difficult thing to do – see e.g. Schroeder 2008 for discussion of such issues.

Accordingly, in the next and subsequent chapters I will present and defend a new variety of response which I believe can succeed where previous responses have failed. But before embarking on that task, we can learn a lot about what a more satisfactory answer to the WNP will need to look like by taking on board the lessons to be learned from the failure of the previous responses. Therefore in this concluding section I will try to draw out some of those lessons, in order that we can bear them in mind as we move forward to consider a new WNP response.

Three related points should be noted as preliminaries to this section. First, the lessons I draw from the failures of previous WNP responses do not constitute an exhaustive list of the required features of any future WNP responses, including my own. For example, it seems largely uncontentious that, being metaethical theories, responses to the WNP must include some account of the relationship between certain putatively moral judgements and reasons for action (even if that account is one according to which those judgements *do not* give us reasons to act). Yet nothing in this section shows that this *must* be a feature of future WNP responses – the reasons why WNP responses must include such features are independent of the lessons I draw here. Second, the lessons drawn are not an exhaustive list of the features a WNP response may include. For example, I will go on to present a response to the WNP which involves having beliefs. But nothing in this section should be read as arguing that WNP responses *must* involve having beliefs. It is up to anyone presenting any kind of WNP response to argue for the specific features of their response. There are numerous other features which WNP responses may also have than those discussed here, provided they are supported by appropriate arguments. And third, this section may not constitute an exhaustive list of all of the lessons we might draw from the foregoing sections. I will simply highlight those points I feel are most salient.

In section 4.2 I discussed conservationism. I rejected it principally because it is not plausible that we can knowingly have and maintain false moral beliefs in the stable way conservationism requires. Conservationism is therefore not something which we could plausibly implement. At first blush, then, the lesson to be drawn here seems quite clear – error theory rules out the possibility of having true (affirmative, atomic) moral beliefs, and conservationism failed precisely because of the known falsity of the beliefs involved. It could therefore be suggested that any successful future WNP response must avoid reliance on us having beliefs at all.

I do not think, however, that this is necessarily the case. Rather I believe a more subtle lesson must be drawn from conservationism's example: my response to the WNP must be such that, if we are to have anything substantially similar to moral beliefs, they must be (potentially) true. As we will see in chapters 5 and 6, I believe it is possible to construct a successful WNP response which relies on agents having something quite similar to moral beliefs. But bearing in mind the nature of the arguments for error theory and the reasons for the failure of conservationism as a WNP response, I will describe a view which supports having true beliefs which can play a role very close to that played by traditional moral beliefs, but which are consistent with error theorists' commitments.

Following conservationism, in section 4.3 I examined abolitionism. I argued that abolitionism fails as a WNP response principally because it rests on an unprovable claim that conventional morality is inherently pernicious. Absent an unfeasible empirical experiment, it remains too easy for opponents of abolitionism to simply reject that central claim of abolitionism and to insist instead that conventional morality is a force for good in the world. What I want to highlight here, however, is that rejecting abolitionism out of hand in this manner is a mistake. It is a mistake because there is a big difference between rejecting a claim and showing why a claim is false. In this case, while it is unlikely that it would ever be possible to perform the

empirical experiment required to substantiate the abolitionist claim that conventional morality is pernicious, it is just as unlikely that the required experiment could ever be performed which might prove that it is not. That being the case, and unless opponents of abolitionism come forward with something more concrete, there seem to be no grounds for deciding the matter one way or another, other than whether we share an intuition with either abolitionists or their opponents. This is hardly the gold standard of philosophical argument.

While this lack of provability undermines abolitionism as a WNP response, however, I believe it draws vital attention to some risks. Even if we cannot be certain whether traditional morality has on balance been harmful, abolitionists have described the potential harms we must be careful not to risk inflicting via our chosen WNP response. In developing my own proposal, I will therefore argue that we must demand of WNP responses that, rather than dismissing it, they are able to cope with the possibility that abolitionists are right that traditional morality is harmful. To put it bluntly, despite the recent emergence of a volume subtitled 'taking abolitionism seriously' (Garner & Joyce 2019), I worry that in not making this demand, even some abolitionists fail to take abolitionism seriously *enough*.

I will seek to avoid this error by treating abolitionism as presenting a challenge which must be answered. Of any WNP response, we can ask whether and how it can cope with changes in how sure we are that any moral or moral-like norms we adopt are really for the best. Since I will be advocating the adoption of a form of moral relativism, which will indeed involve retaining something like moral beliefs, then I must be open to the charge that those beliefs are not as beneficial as we might have assumed. That being the case, there must be some thought given to how we might reassess our new judgements as we learn more about their real, rather than expected or hoped-for, effects and the consequences of adopting my proposal. This in turn seems likely to raise issues around the authority of our relevant judgements – if they are open to reappraisal, how can they have the authority required to

play anything like the role played by traditional moral judgements? I will discuss these worries, and show how my view can cope with them, in chapters 5 and 6.

Added to this, since my own proposal will be based on a form of relativism, I will need to be doubly careful - one of the traditional criticisms of moral relativism is that it cannot provide sufficiently authoritative moral reasons. For example, it is often objected that moral relativists lack grounds on which to claim that the agents who were involved had any authoritative reasons not to commit the atrocities perpetrated by the Nazis, and that this undermines relativism (see e.g. Baghramian & Carter 2018 §4.5). I will not digress here into the debates around this topic, but the lesson that I will need to pay close attention to the matter of normative authority is clear.

I would also like to highlight two instructive features of the discussion of revolutionary expressivism in section 4.5.[122] First, I argued that expressivism cannot give a satisfactory account of the tension between 'moral' judgements and temptation in the WNP context. It seems to me that temptation, understood as some kind of desire to act contrary to a prior 'moral' judgement, will always be able to override 'moral' judgements in motivational power because according to expressivists, those judgements are themselves simply desires or desire-like attitudes. It remains open to expressivists to polish their proposal and to introduce a way to distinguish specifically 'moral' judgements from other desire-like attitudes in such a way that 'moral' judgements are somehow reliably more motivationally efficacious. They have not yet done so, but the lesson which I can take away from the failure of expressivism is that my WNP response must include a way of showing that the relevant judgements can effectively counteract temptation.

[122] As in section 4.5, unless noted otherwise, all references to expressivism here should be read as referring to *revolutionary* expressivism, i.e. expressivism as a response to the WNP, rather than to more standard hermeneutic forms of expressivism.

Second, I noted towards the end of section 4.5.2 that on closer examination it was not clear that expressivism delivers the benefits claimed, specifically for its account of moral disagreement. This reflects the general requirement outlined in section 4.1 that there must be a clear and explicit link between adopting a given strategy and securing the benefits claimed for doing so. The example of the failure of expressivism in this regard serves to highlight that requirement.

A final lesson is that we must bear in mind the implementation costs of any given WNP response. For any WNP response, we can ask whether it would be necessary for the general public to understand all of the metaethical intricacies involved in order for the response in question to deliver its benefits. If it would, it seems exceedingly unlikely that the proposal could ever be implemented – some problems in metaethics are simply too complicated for many people to understand, and it seems unlikely that many people would ever be willing to devote the time required to do so even if they were capable of understanding them. That is not to say that all WNP responses must be simple enough for every member of the general public to quickly and easily understand their every detail, but it does mean that responses requiring everyone to suddenly become metaethical geniuses should be regarded with a measure of suspicion.

Given the failure of previous attempts to propose a satisfactory response to the WNP, and armed with the lessons drawn here, in the next chapter I will begin to propose my own, new response, which I call revolutionary relativism. In subsequent chapters I will argue that revolutionary relativism is superior to the other WNP responses to date, and then defend the proposal against a number of objections. Ultimately, I will argue that having accepted a moral error theory, we should eschew the WNP responses defended to date, and become revolutionary relativists.

Chapter 5. Revolutionary Relativism

In this chapter, I will introduce and explain my own proposal for how we should respond to the 'what now?' problem. I call this proposal revolutionary relativism. I use the term relativism because, as we will see, I propose that we adopt a practice of making apparently moral judgements and statements which include what Boghossian has called (2006) a 'relativising parameter'. And I use the term revolutionary for two reasons. Firstly (and obviously in the current context) because the proposal comes into play only after we have come to accept a moral error theory, which is surely a revolutionary metaethical event. And secondly because, as will become clear, the model of relativism which I propose is one which, in line with the demands of the WNP context, is considerably different from hermeneutic forms of moral relativism. In comparison with the existing literature at the time of writing, my proposal is unique among responses to the WNP in suggesting that we should adopt any form of relativism.

I will begin in §5.1 with a short summary of where we are in the overall structure of this thesis. I will then lay out some preliminary issues, including some constraints on good WNP responses in §5.2. And then in §5.3 I will give the details of my proposal, showing how the various parts of my proposal are each required and fit together, and showing how my proposal respects the constraints discussed in §5.2. Finally, I will summarise where this chapter has taken us, and how it feeds into the subsequent chapters.

Looking ahead, this chapter will form the first of two broad parts of the positive section of my thesis. Here, I will lay out my own positive proposal. Then the second of these broad parts comes in the following two chapters where I will go on to defend the proposal I offer here. I will do this in two distinct ways. In chapter 6, I will argue that revolutionary relativism is a

better response to the WNP than those offered by others to date. This chapter will be one of the most important sections of this project, and hence will be in considerable detail. Alongside the explicit aim of showing that my proposal is the best response to the WNP, this will allow me to draw out some important facets of how revolutionary relativism would work in practise. And then in chapter 7 I will defend revolutionary relativism against a number of criticisms which confront relativist metaethical theories in general, revolutionary or otherwise, in order to show that my proposal can stand on its own two feet, philosophically speaking. I will also defend revolutionary relativism against certain objections which crop up specifically in the post-error-theory context.

5.1. Summary thus far & preliminary considerations

In this section I will give a brief reminder of where we are in the overall structure of this thesis. I will then discuss part of Crispin Wright's objection to error theory (1996). Despite its failure to convince error theorists that they are wrong, I will suggest that Wright's argument highlights two important considerations which will inform my proposed response to the WNP.

If we become convinced by the arguments for error theory which I discussed in chapters 2 and 3, then as I argued in §4.1, we will inevitably and immediately face a new metaethical problem. I call this the 'what now?' problem, or WNP. There are various reasons why this problem arises – we might worry that something must fill the void in our normative lives left behind by traditional morality in order to avert catastrophe, for example. But one of the most compelling (and often overlooked) reasons why we face the WNP arises out of a fundamentally practical matter of philosophy itself. Regardless of our feelings about error theory, if we accept it, we will do *something* next. If we respond to error theory by abandoning moral discourse, as some commentators have suggested we might most naturally

do, this represents just one response among several possible responses.[123] The same is true if we seek to retain our moral beliefs, or anything substantially similar to them. And so long as there are multiple possible responses, the adoption of each of which may have different results from a variety of theoretical and practical viewpoints, then surely we fail as philosophers if we do not carefully consider the options.

Accordingly, a number of philosophers have sought, whether they would see it in the terms I am using here or not, to respond to the 'what now?' problem. In sections 4.2-4.5 I grouped these responses into four categories: conservationism, revolutionary fictionalism, abolitionism and revolutionary expressivism.[124] In each case I argued that the responses offered to date face serious problems. The extent and severity of these problems force us, I argued, to look for new responses to the WNP, and just such a new response is what I aim to provide here.

When considering how to respond to the WNP, we effectively have a moral *tabula rasa*. If we assume the truth of a moral error theory, and assume the commitments typical of error theorists along with that truth, then the way we have understood morality until now is not merely *somewhat* mistaken. It is entirely insufficient to think that, say, *some* of our moral judgements might be unreliable or incorrect. Rather, we are forced to accept the profound conclusion that morality is *systematically* in error – it is radically mistaken, root and branch. There is not, nor can there ever be, anything which any agent is ever morally required,

[123] See e.g. Wright 1996, p. 2: 'Whatever we may once have thought, as soon as philosophy has taught us that the world is unsuited to confer truth on any of our claims about what is right, or wrong, or obligatory, etc., the reasonable response ought surely to be to forgo making any such claims'.

[124] In what follows, I will often omit the *revolutionary* tag and refer only to fictionalism or expressivism *simpliciter*. This is because, as I have iterated above, at this point in this project we are now in a post-error-theory context. It therefore becomes appropriate to assume the revolutionary nature of the fictionalism and expressivism under discussion, and draw a distinction instead by using the label *hermeneutic* when *non*-revolutionary forms of these views are discussed. For reasons of readability or emphasis I will still use the revolutionary tag from time to time. But where fictionalism or expressivism are mentioned without further qualification, the revolutionary form of each should be understood to be the intended target.

permitted or forbidden to do. All bets, morally speaking, are off. And so we plot our course forwards by beginning with what appears to be a blank slate.

But there are things which might point us to certain kinds of response to the WNP over others. And in formulating my own response, I am to some extent guided – perhaps ironically – by one of Wright's arguments *against* error theory. In 'Truth in Ethics' (1996), Wright's concern is that, if an error theory is true, then it would have 'calamitous' (p. 2) consequences, for 'how are we supposed to take ourselves seriously in thinking the way we do about any issue which we regard as of major moral importance?' (*ibid*.). Wright then goes on to make an argument that error theorists are mistaken about their commitments as regards what counts as truth for moral judgements. This argument has not proved convincing to error theorists, but it can inform WNP responses nonetheless, as I will explain.

Briefly, writing before Joyce's *The Myth of Morality* had been published, Wright argues against a Mackieian error theory (see chapter 3 of this thesis). His argument is that error theorists understand truth for moral discourse in terms of correspondence with moral facts which, on reflection, do not exist – hence the systematic error in moral thought and discourse. However, Wright suggests that it may be possible to locate some other element of moral thought and discourse in virtue of which we might say that moral judgements or sentences could be true or false. For example, we might determine which moral judgements were useful or prudentially beneficial, and take those judgements to be true. If we were to do so, Wright argues, we would make available a new success theory of morality which can withstand error theorists' arguments concerning traditional morality, and therefore avoid a metaethical error theory. Wright then argues that error theorists will not be able to respond effectively to this strategy, writing, '[t]he error-theorist may be able to argue that the superstition that he finds in ordinary moral thought goes too deep to permit any construction of moral truth which avoids [construing moral truth in terms of the satisfaction of some weaker standard than

correspondence with metaphysically outlandish moral facts] to be acceptable as an account of *moral* truth. But I do not know of promising argument in that direction' (1996 p. 3).

Unfortunately for Wright c.1996, as we saw in chapter 3, this was just the kind of argument Joyce made a few years later. To put it in compatible terms, we might paraphrase Joyce's principal 'non-negotiable commitment' (Joyce 2001 *passim*, beginning p. 17) as the claim that no judgement can be *moral* unless it ascribes a non-institutional categorical obligation. Put in terms of truth, Joyce's argument for an error theory becomes roughly that no sense can be made of such obligations, and so no moral statement can be true.

Wright's argument may fail to convince today's error theorists, but we can still draw out some useful considerations regarding how we might respond to the WNP. First, Wright's worry about whether we will be able to take seriously the things which we considered hugely morally important before we accepted error theory highlights the fact that error theory itself will be much more palatable to skeptics if we can respond to the WNP in a way which preserves the profound seriousness with which we previously regarded matters such as torture, murder and suchlike. Indeed, a WNP response which can accomplish this may have an advantage over competing responses in virtue of its being psychologically easier to accept than, say, abolitionism.[125] Alongside this, it is plausible that seriousness is important in an instrumental sense – judgements which we take seriously seem much more likely to be able to affect our behaviour than those we do not. Such is the sense of importance with which we regard many of our moral judgements that we may consider good WNP responses to be under a *Seriousness Constraint* – i.e. any good WNP response must be able to preserve the seriousness with which we regard moral judgements about murder and similarly weighty matters. While it may be possible to construct a WNP response which fails to respect this Seriousness Constraint, a

[125] This point is similar to one raised against abolitionism by Nolan *et al*. (2005 p. 310-311).

response which *does* respect it will be a more satisfying response and will head off possible objections along the lines of Wright's worry here. As we go on, some further constraints on good WNP responses will also become clear.

Second, Wright's argument highlights the possibility of responding to the WNP in such a way that apparently moral judgements may be thought true or false in respect of some other standard than correspondence with a kind of moral fact which, having accepted a moral error theory, we do not believe exist. In keeping with Joyce's 'non-negotiable commitment', we would not be able to call these judgements *moral* in the proper sense. But where Wright sees a reason to change our definition of the term moral and otherwise keep moral thought and discourse largely as they are – and therefore abandon the error theory - I suggest that we can consistently remain error theorists, yet avoid any 'calamitous' consequences of error theory by keeping as much as possible of moral thought and discourse intact, and simply dropping the full-blooded use of the term moral. In order to draw the appropriate distinction while emphasising that my proposal retains much of what error theorists would claim constitutes traditional morality, I will use the term moral*.[126]

[126] A note on the use of the terms moral and moral*. It may seem somewhat cumbersome to continually use the term moral* in the remainder of this thesis. But, as I pointed out in my discussion of revolutionary expressivism in the previous chapter, to label the obligations, reasons, judgements etc. under discussion here *moral* would be incorrect by error theorists' lights (see §4.5.1). The commitments of typical error theorists will include the commitment that moral obligations, reasons, judgements etc. involve some non-institutional categorical element without which those obligations etc. cannot properly be called *moral*. And since the obligations etc. revolutionary relativism deals with do not have this categorical element, revolutionary relativists cannot therefore consistently call them *moral*, even if the phenomena in question are in other respects very similar to their properly moral equivalents a lot of the time.
It may seem appropriate to insert a general disclaimer explaining that this remains the case, but that I will henceforth gloss over the matter and use the term moral regardless for the sake of fluency or readability. But this risks confusing the reader (and alarming any error theorists who might have skipped past the disclaimer). On a slightly more sophisticated level, in principle I see no reason why revolutionary relativists could not defend a localised form of fictionalism about the term moral. Even though the reasons, obligations etc. revolutionary relativism involves are not categorical and therefore are not actually moral by error theorists' lights, the argument could be made that revolutionary relativists are free to speak and even deliberate *as if* the term moral could properly be applied to these phenomena. At least, they could do so in all but unusually metaethically critical contexts. Again, however, in the present context I fear this might confuse less careful readers and generally muddy the

If this suggestion is right, then we have a prudential reason to formulate a WNP response which preserves important elements of traditional moral practice, yet does so in a way which respects the commitments of error theory. This is, of course, part of the reasoning behind conservationism as I discussed in chapter 4. But I believe we can respond to the WNP in a way which preserves key features of moral thought and discourse such as interpreting judgements as beliefs, yet which avoids the problems which overwhelm Olson's proposal.

5.2. Groundwork for the metaethics of revolutionary relativism

I will begin to outline my proposal with a couple of preliminary notes which necessarily underpin the proposal, and may underpin other proposals as well. The first two of these are 'ground rules' which can be read as constraints upon good WNP responses, which go alongside the Seriousness Constraint mentioned in §5.1. After explaining these constraints, I will move on to a discussion of prudential reasons which will flesh out the 'should' in my proposal that error theorists should become revolutionary relativists. In §5.3 I will then present my proposal, and explain how it works and how the various aspects of the proposal relate to and follow from one another.

5.2.1. Constraints upon good WNP responses: i) The ROBET Constraint

The first thing to establish is that my proposal that error theorists should become revolutionary relativists must not, and does not, rely on anything which is incompatible with the truth of error theory. For example, to suggest that error theorists have a moral obligation to respond to the WNP in one way versus another is a non-starter. The *should* in 'should become revolutionary relativists' must not be read in a moral sense, or any other sense which

metaethical waters. Therefore I will continue to use moral to refer only to what might also be called 'traditional' morality, and to use moral* as a distinct term despite its slight unwieldiness.

relies on any non-institutional categorical notion of normativity – otherwise my proposal would itself be ruled out by error theory, and the whole aim of this project would unravel. For ease of reference, we can call this the '*Ruled-Out-By-Error-Theory*' (hereafter *ROBET*) *Constraint*.

5.2.2. Constraints upon good WNP responses: ii) The RC Constraint

Second, my proposal must, and does, respect the commitments incurred by error theorists *en route* to their error theories. Imagine, for example, that there was a theory, 'red-theory', to the effect that humans can only see shades of red. In the course of their supporting arguments, 'red-theorists' make the claim that colour cannot be understood other than as a visual phenomenon. It would be inconsistent for someone to then include in their attempt to respond to 'red-theory' the claim that humans can perceive the colour green by smell. Such a claim would undermine the previous claim about visual phenomena, and so undermine 'red-theory'. Naturally this is an absurd example, but the structural issue of consistency is clear – the commitments borne by the arguments for a given conclusion must be respected by theories about how we might build on that conclusion. For ease of reference we can call this the '*Respect Commitments*' (hereafter *RC*) *Constraint*.[127]

The RC Constraint applies generally across the supporting arguments for error theory. In this section, the particular commitment typically borne by error theorists in virtue of their arguments for the error theory which I wish to draw attention to is aptly captured by Joyce's theory of practical reasons. Joyce's arguments in this area are influential, and arguments made by others to date have often either recommended or gone along with them (see e.g.

[127] The ROBET and RC constraints are a broadened and clarified version of the Normative Circularity Constraint I discussed in §4.1.3. There, it was sufficient to use a more hastily expressed constraint. But here I have expanded on the NCC to bring out relevant details as is more appropriate for this more sophisticated stage in the thesis, and to include e.g. proscribing the attribution of true moral (as opposed to moral*) beliefs alongside the issues discussed in the earlier section.

Robertson 2008, Nolan *et al.* 2005) or offered similar arguments themselves (see e.g. Olson 2014 chapter 6). Therefore, it seems fair to consider Joyce's theory as typical among error theorists. As we saw in chapter 3, Joyce argues persuasively, and as a key part of his argument for an error theory, that it is practically rational for an agent to act only on those reasons which she believes conduce to the satisfaction of some desire or end she has.[128]

Accordingly, it would seem to fail to respect both the RC and ROBET constraints if I proposed that agents who come to accept the truth of a moral error theory *should* do anything which would fail to satisfy any of their desires (or which promotes an outcome which is not among the agent's ends).[129] It seems that there would, by Joyce's lights, be no authoritative reason why agents *should* do anything which would not conduce to their current desires or ends, and thus agents may even be practically irrational to adopt my proposal unless they antecedently had appropriate desires or ends. This may seem to dictate that the form of my proposal must be something like this:

1. Adopting my proposal and becoming revolutionary relativists will result in error theorists' lives going better than they would otherwise.
2. This means that error theorists have prudential reasons to adopt my proposal.

[128] Or, possibly, reasons which and agent may have in virtue of some institution in which they are participating – because of the rules of a game they are playing, say, or because they are playing a certain role (see also Olson 2014 §6.1). I set these 'institutional' reasons aside here because I am discussing reasons for adopting one WNP response rather than another, and there are no roles or games which require the adoption of specific WNP responses.

[129] There is a degree of overlap between the constraints here. Suggesting that agents should do something which does not relate to their desires/ends violates the commitment, specifically part of Joyce's theory of practical reasons, that an agent's authoritative practical reasons must relate to their desires or ends. The same manoeuvre also implies that there can be authoritative forms of obligation which do not relate to agents' desires/ends, which is ruled out by the error theory.

The idea here is that error theorists' lives will go better if they adopt my proposal, and therefore *to the extent that they want their lives to go better* (which surely most error theorists do, if not all), error theorists should, prudentially speaking, adopt my proposal. This avoids reliance on any kinds of categorical practical reasons, and thus respects the RC and ROBET Constraints.

This falls somewhat short of what I need here, however, as a number of questions may remain. For example, what does it mean for an agent's life to go better? Or what if an agent has come to accept error theory, but for some reason has only frivolous or even harmful desires at the moment (because they are drunk, say), which would not be served by adopting my proposal? Or what if error theorists antecedently desire to retain some of the coordinative benefits of traditional morality (which I am claiming that adopting my proposal can deliver), but hold false beliefs about the course of action most likely to bring this about? Do they still have authoritative reasons to adopt my proposal? And if so, what does this imply about practical reasons, regardless of what I have said above – are the RC and/or ROBET Constraints under threat?

What is needed here is a more concrete understanding of the sense of prudential reasons I have in mind. Therefore, in the next subsection I will offer an account of prudential reasons and why agents have them, and thus show more precisely what I mean when I claim that error theorists *should*, prudentially speaking, adopt my proposal.

5.2.3. Prudential reasons

Without a sufficiently nuanced and sophisticated view of prudential reasons, vexing questions about the detail of my proposal remain, and the proposal may seem to risk violating one or more of the constraints which I have only just introduced. Just such a view is famously

defended by Bernard Williams.[130] Williams' paper discussing the issue, 'Internal and External Reasons' (1981), is both important and nuanced, and the view he describes is not necessarily one about prudential reasons, being rather a view about normativity in general and internal versus external reasons in particular.[131] But I believe that sufficient detail concerning prudential reasons can be extracted from Williams' wider view to be useful in the context of this thesis. Joyce also discusses Williams' view at some length (2001, especially chapter 5), but I differ in emphasis here by focusing specifically on the implications for prudential reasons. The classic example of the view is this:

> The agent believes that this stuff is gin, when it is in fact petrol. He wants a gin and tonic. Has he reason, or a reason, to mix this stuff with tonic and drink it? […] On the one hand, it is just very odd to say that he has a reason to drink this stuff, and natural to say that he has no reason to drink it, although he thinks that he has. On the other hand, if he does drink it, we not only have an explanation of his doing so (a reason why he did it), but we have such an explanation which is of the reason-for-action form. (Williams 1981 p. 102)

One way of interpreting Williams here is as offering a definition of prudential reasons. For simplicity, I will set aside the tonic, and focus on the gin. We can see that in his current state of ignorance about the petrol in the glass, it would be practically irrational by Joyce's lights for Williams' agent to avoid drinking what he believed to be gin, given that his current desiderative set includes the desire to drink gin (and not to drink petrol - or so we are clearly

[130] A not dissimilar view is discussed in another well-known paper by Peter Railton, 'Moral Realism' (1986). Discussions and refinements of this kind of view can be found in Smith, Lewis & Johnston 1989 and Smith 1995a.

[131] To give a sense of how important and nuanced, a google scholar search at the time of writing showed over 1500 citations of Williams' paper. For an overview of the relevant issues, including at least two different readings of Williams' overall argument, see Finlay & Schroeder 2017, especially §2.1.1 & 2.1.2.

to infer from the example). Yet, if the agent had all the relevant information, and no false information, it seems that given his desiderative set he would refrain from drinking.

So here we have a suggested kind of reason which depends only partially on an agent's current desiderative set, and partly on a counterfactual idealisation, i.e. what the same agent would desire *if they had better relevant information*. To make this picture of prudential reasons more precise, we may also add to the counterfactual idealisation that alongside having all the relevant information and no relevant false beliefs, and that the agent deliberates clearly (i.e. is not irrational). Finally, we should also add that the agent's desires are made into an appropriately coherent set. For example, the 'petrol-ignorant' agent may have a *de re* desire to consume the contents of the specific glass in front of him. Given the rest of the agent's desiderative set, this desire would not survive clear deliberation in light of the relevant facts and in the absence of relevant false beliefs, and so making the agent's desires more coherent would involve his ceasing to have this *de re* desire. Alternatively, and recalling Joyce's story about Molly (see §3.3.4-5), an agent may have an occurrent desire which runs contrary to another more deeply considered and subjectively more highly valued desire or end. Were the agent to deliberate carefully on the matter, they may come to see that these desires were incompatible, and so come to have a more coherent set of desires by ceasing to be motivated to act on the occurrent desire. Note that this process involves only such changes to the agent's desires as are needed in order to make the agent's desiderative set more coherent in light of the relevant facts, and no wider changes are suggested.[132]

[132] One thing I am specifically not suggesting is that there would be any necessary convergence among the desires of multiple idealised agents, as Smith argues (e.g. Smith 2015). I mention this explicitly both because Joyce discusses and rejects Smith's argument (2001 §3.8), and because allowing this convergence claim would risk violating the RC Constraint, since Smith builds from the convergence claim towards a refutation of error theory (e.g. 1995b §5, 2010). In the current context, we have accepted error theory, so that battle is over, but the RC Constraint must be respected.

On this definition, then, a prudential reason is a reason for an agent to act which that agent would be practically rational to act upon, given their current set of desires, if they were aware of the relevant information, had no relevant false beliefs, were not irrational, and had appropriately coherent desires.[133] This is the kind of reason I mean to imply when I say that error theorists *should* become revolutionary relativists. And when I claim that error theorists' lives will go better if they adopt my proposal, what I mean for an agent's life to go better is for more of the desires of their counterfactually idealised selves to be fulfilled. Roughly speaking, an agent's life will generally go better for them if their prudential reasons are acted upon. Accordingly, a prudential benefit is a situation or outcome which satisfies or conduces to the satisfaction of the desires for current agents of those agents' counterfactually idealised counterparts.

Now, error theorists could currently be in a position analogous to Williams' petrol-ignorant agent – for example they may believe that adopting a given competing proposal will bring about a desirable situation, when in fact the consequences of adopting that proposal would be catastrophic.[134] If that were so, respecting the RC constraint in the flat-footed manner described in §5.2.2 would imply that the only thing I or anyone else could claim that those error theorists *should* do would have disastrous consequences.

[133] This may be a controversial use of the term prudential reason. For example it could be suggested that what I am describing should more properly be called a subset of internal reasons, and/or that prudential reasons have different features to the features I claim here (see e.g. Hubin 1980 for a different definition of prudence and prudential reasons). I will not engage with these suggestions here. My use of the term is clearly defined above, does not contradict an intuitive sense of the meanings of the words involved, and most importantly does not violate the ROBET constraint by, for example, positing authoritative reasons which do not relate to agents' desires or ends. I leave it to other work to consider whether this is the proper, or only, relevant definition of prudential reasons.

[134] Note that I am simply describing a logical possibility here - I am not claiming here that adopting any of the specific competing WNP responses defended to date actually *would* result in catastrophe. I gave what I take to be compelling reasons to reject other WNP responses in the previous chapter, and I will discuss comparisons between the other WNP responses and my own in the next chapter. Note also that although I will not make an argument to this effect in this project, it may well be that other WNP respondents should bear the reasoning given here more closely in mind when formulating their own proposals.

However bearing in mind the more sophisticated view of prudential reasons I just described, we can see that the intention in this project is to give such error theorists the best available information, demonstrate that some of their current beliefs may be false and may therefore be discarded, and show that the clearest deliberative route supports my proposal. Essentially, this thesis is aimed at turning current error theorists into their counterfactually idealised counterparts.

Given this structure of argument, prefacing my proposal that as error theorists we should become revolutionary relativists with 'if we want our lives to go better…' is redundant. The definition of prudential reasons I am using here means that to claim that an agent has a prudential reason to ϕ *just is* to claim that that agent already has relevant desires, even if they don't realise it.

It could be objected that this assumes too much about some error theorists and their current desires. Perhaps there are those who do not want to retain any of the apparent benefits of traditional morality, and who would not want to do so no matter how much extra information they had or no matter how much their deliberative process were improved. Such a stance might be taken by abolitionists, who urge us to reject moral thought and discourse because they cause us harm, at least on balance. To them I respond by begging their indulgence until the next chapter, where I will show that it is possible to retain traditional morality's benefits while avoiding the harms which abolitionists diagnose. A similar stance might be taken by error theorists who believe we (or they) would be better off in a world without cooperation, coordination and so on. I will return to this issue in chapter seven. For the time being it will

suffice to respond by claiming that it is highly plausible that *most* error theorists have desires such that – if my arguments in this thesis are right - they should adopt my proposal.[135]

With the above established and borne in mind, I am now in a position to introduce my proposal. The next section will contain my proposal along with some arguments supporting the inclusion of its various stages. I will then break down and explain the proposal in more detail in the subsequent sections. Since my reasoning here is specific to my proposal and therefore distinct from the equivalent stages of other WNP responses, I will begin by building up from some quite basic considerations.

5.3. Formulating revolutionary relativism

This section will really be the heart of my thesis, because I will explain in detail how exactly I propose we should respond to the 'what now?' problem.[136] I will begin in §5.3.1 by fixing the scope of my proposal – something I raised as a criticism of Joyce's proposal in the last chapter. In §5.3.2-5.3.4 I will then lay out in detail the attitudes and commitments of the judgements I propose we should make in the post-error-theory world, and why they can help deliver the prudential benefits we desire.

[135] The view of prudential reasons described may raise another potential problem in the WNP context, even if that problem does not arise for Williams, Smith *et al.* For even if prudential reasons as described here exist, we might wonder how we could come to know what they are. After all, we cannot be sure that we know what our idealised counterparts would desire for our current selves, because by definition they may have access to information which we do not, and which we cannot know that we currently lack. This is not a particularly serious problem for my proposal, however. I am not proposing that we must know with utter certainty what our prudential reasons are sub specie aeternitatis. Rather, I believe that we will have a clearer, albeit potentially imperfect, idea of what our reasons are if we carefully think things through along the lines I have described, and make committed efforts to get the best handle we can on what the desires of our idealised counterparts would plausibly be. In Joyce's terms we can thus bring our subjective reasons into a closer alignment with that subset of our objective reasons which I am calling our prudential reasons than we could if we failed to make the effort.

[136] In case it needs repeating, remember that I am assuming the truth of error theory here – in the current context, we are all error theorists.

5.3.1. The scope of the proposal

The most basic aspect of my proposal is its scope – to whom is my proposal addressed? In §4.4.3 I criticised Joyce's competing WNP response partly because of questions about its intended scope - a proposal which applies to just one person, or even to groups, can run into difficulties. Therefore let me be clear about my own WNP response: I propose that all capable agents become revolutionary relativists.[137]

The reason for this universal scope is simple. If error theorists are right, as we are assuming here, then all agents who wish to avoid systematic errors in their beliefs should accept error theory. This means that everyone, everywhere faces the 'what now?' problem. This whole thesis is my attempt to show that the best way of responding to the WNP is to adopt revolutionary relativism. In the broadest possible terms, this is because, out of the available ways of responding to the WNP, revolutionary relativism most plausibly delivers the greatest level of prudential benefits. If I am right about that, especially about the comparisons between my proposal and competing ways of responding to the WNP, then we – that is, all capable agents - should become revolutionary relativists.[138] This is one of the senses in which my proposal is revolutionary – whatever may happen in practise, from a theoretical point of view, we are talking about changing the way everyone, everywhere understands morality.

[137] I use the word capable here simply because there may be individuals who count as agents, but who cannot become revolutionary relativists in any meaningful sense. For example children may fall into this category because they cannot understand virtually any metaethical view. I do not see this as a problem, since 'everyone who can do so should ϕ' is an acceptably universal scope in the current context. But it is better to include capability here since if I left it out, it could leave open an objection to my proposal on these spurious, albeit technically correct, grounds. In what follows I will sometimes simplify this formulation to 'everyone, everywhere' or 'everyone'.

[138] This is a subtly different issue to the question of whether or not every individual agent's all-things-considered strongest relevant practical reason is to adopt my proposal. At this stage however, this distinction may seem somewhat confusing, and so I will defer discussion of this subtle point until chapter 7.

There are two further related reasons for the wide scope of my proposal. First, one of the putative benefits of traditional morality is that it facilitates coordination among agents, for example by furnishing agents with moral reasons to keep promises, stick to agreements, and so on. As I will argue, we can preserve this benefit, even though in the WNP context we no longer believe that moral reasons exist, by adopting my proposal. That being the case, the more people who adopt revolutionary relativism (or at least the more people to whom the advice to do so can be addressed), the more coordination among agents there will be.

Second, there can be a tendency to dismiss niche views, regardless of how well-founded those views are. This means that if my proposal were addressed to only a small group of people, even if the proposal were totally successful and convinced every member of its target audience, it would still run the risk of failing to respect the Seriousness Constraint I discussed in §5.1. Naturally, having a wide scope does not guarantee wide adoption, but a wide scope does mean that this potential problem is not built into my proposal from the start.

The final issue with the scope of my proposal is what counts as adopting my proposal or becoming a revolutionary relativist. One could interpret this as a demand that agents should be very metaethically sensitive and reflective indeed. Or it could be that simply acting roughly in accordance with my proposal is sufficient, even if agents do not actively engage with all of the metaethical details. What I am proposing is somewhere between these two interpretations – I am suggesting that it is in everyone's interests to act in accordance with my proposal, and this would count as 'adopting' the proposal. But beyond this, it seems likely that a more conscious and deeper understanding of why they are doing what they are doing will be helpful for people who do act in accordance with my proposal. Thus I am not suggesting that in order to adopt my proposal, everyone must do a PhD in philosophy. Simply acting in accordance with the proposal will be sufficient to count as adopting it and being a

revolutionary relativist. But being sensitive to and reflective upon the metaethical details as well surely won't hurt!

5.3.2. Core belief & truth conditions

Having established to whom my proposal is addressed, the question now turns to what exactly it is that I am proposing all capable agents should do. Remember that we are seeking something which can amend or replace traditional morality, and which will have a number of desiderata. As far as possible, adopting the proposal will preserve features of traditional morality which are widely held to be prudentially beneficial. For example it will facilitate coordination among agents, and will provide a framework for understanding and resolving disagreements.[139] Given Wright's worries, we also need to ensure that the WNP response we adopt can accommodate and offer the means to articulate certain judgements and actions which are very important to us, by giving us an appropriately serious way of thinking about and discussing some of the most difficult topics we can think of like rape, murder, torture and so on. And bearing in mind Mackie's point about different ways of life, i.e. systems of first-order moral beliefs being optimal for people in different 'concrete circumstances' (1977 p. 37), the proposal should admit of locally distinct variations. For example, if there were a prudentially desirable outcome in terms of resource distribution – that everyone has enough water, say – then there could be local variations in the prudentially optimal norms which regulate people's behaviour around water, dependent on the local variations in the availability of water in arid vs. verdant parts of the world.

[139] This may seem to beg the question against abolitionists who would claim that morality, and anything substantially similar to morality which we might adopt in response to the WNP, necessarily hinders conflict resolution. But it does not. In the next chapter I will be at pains to engage with abolitionism far more thoroughly than other WNP respondents typically do.

I submit that the best way of achieving these desiderata is to retain as far as possible the language and patterns of thinking which go along with morality, albeit changed in order to remain consistent with error theory and the commitments of the arguments for error theory.

Therefore, I propose that we respond to accepting the truth of error theory, and to the resulting 'what now?' problem thus: on the same occasions when we would previously have made moral judgements, we make moral* judgements instead. The same would apply, *mutatis mutandis*, to moral vs moral* thought and discourse in general. These moral* judgements, exemplified here by the token judgement that 'it is morally* wrong to ϕ', are defined as follows:

For an agent, *S*, to judge that it is morally* wrong to ϕ, is for it to be the case that:

> BELIEF: *S* believes that ϕing is in contravention of one or more practical norm(s), acceptance of which is a condition of participation in the moral* community in which *S* is a participant.[140]

This will require some unpacking (for example I will need to explain precisely what I mean by acceptance of a norm, and how moral* judgements differ from moral judgements), and I will lay out what it amounts to in more detail shortly. Initially, however, note that for the purposes of assigning truth conditions to *S*'s belief, this is all that I propose. Thus S's belief that it is morally* wrong to ϕ is true if and only if ϕing is in contravention of one or more practical norm(s), acceptance of which is a condition of participation in the moral* community in which *S* is a participant.

[140] I use 'wrong' here simply because it follows the terminology I have used in previous sections. The same proposal would generalise, *mutatis mutandis*, to obligatory, permissible and so on, just as is the case with the traditional moral terms wrong, obligatory etc.

Note also that facts about the norms a community accepts are descriptive, non-normative facts about the moral* community in question. Therefore revolutionary relativism does not posit the existence of any objectively prescriptive facts or properties, and neither does it invoke any non-institutional categorical forms of normativity. Thus, revolutionary relativism respects the RC and ROBET constraints. On a terminological level, and recalling §4.5.1, this means that my proposal cannot involve e.g. having *moral* beliefs, since in the WNP context, the term *moral* necessarily implies categoricity. Yet since the beliefs I am describing here are so closely related to moral beliefs, I find the term moral* appropriate (along with related terms such as morally*, morality* etc.).

The final point to note at this early stage is that having a mental state with the structure of BELIEF is not by itself the full extent of making a moral* judgement according to my proposal. In order for revolutionary relativism to be a comprehensive proposal which works as intended in the WNP context, BELIEF will need to be supplemented somewhat. Therefore in order for *S*'s belief to count as a moral* belief in the required sense, I propose that *S* must have certain further commitments. By the term commitment, what I have in mind is similar to Valerie Tiberius' notion of commitment, which she describes thus: 'I use the term "commitment" to indicate the attitude or set of attitudes we have toward the things we take to have (at least some) normative significance. Commitments are, at least in part, constituted by passions and sentiments that motivate us to act' (2002 p. 166). A commitment in this sense, then, involves a set of attitudes. Thus my proposal is that for *S* to have a moral* belief that it is morally* wrong to ϕ, is for *S* to hold BELIEF plus certain further attitudes, which I will specify as we go on.

This heads off a potential objection. It could be argued that we commonly have non-moral beliefs similar to BELIEF (for example in matters of etiquette or style), and so proposing only that we cease to have moral beliefs and instead have beliefs similar to BELIEF amounts to a

form of abolitionism. However, my proposal is not that simple – the moral* judgements I am proposing we make comprise not only BELIEF but also the further attitudes I will describe as we go on. Therefore, my proposal is not that we jettison the moral category of judgements and 'make do' with our other pre-existing categories of judgement (which *would* make me an abolitionist). Rather, I propose that we adopt a new kind of judgement (which makes my proposal non-abolitionist) which also includes the further attitudes I will shortly describe.[141] This raises a matter which must be cleared up before moving on, however – given what I just said about the truth conditions of moral* beliefs, how are we to determine what counts as a moral* vs a non-moral* belief, and how do the further commitments of moral* beliefs bear on the truth or falseness of moral* beliefs?

5.3.3. Moral* vs non-moral* and the truth conditions of moral* beliefs

I am proposing that if *S* judges that it is wrong to ϕ, yet lacks the further attitudes I am about to describe, then *S*'s belief is not a moral* belief. In hermeneutic theories, it is a controversial matter whether moral beliefs necessarily imply motivation or desires which go along with the propositional content of the moral belief in question. In the present context, however, there are no pre-existing phenomena of moral thought and discourse which require explanation. These phenomena have been explained, and found to be subject to an error theory. It is therefore an advantage of the WNP context that it is up to me to specify the nature of my proposal, and I am specifying that in order for a belief to count as a moral* belief, further attitudes are required alongside BELIEF. Accordingly, an agent who has a belief with the

[141] This means that technically, although I am using the name revolutionary relativism, since the view I am proposing we adopt includes more than just one (kind of) mental state, its full name would be Revolutionary Hybrid Cognitivist Cultural Relativism. Revolutionary relativism is a much snappier title, though!

content BELIEF but who lacks the further attitudes I will describe does not have a moral* belief, but instead an anthropological or some other non-moral* type of belief.

This means that according to my proposal, there could be two beliefs with identical content of the form of BELIEF which could both be true, one of which is a moral* belief because the agent who holds the belief also holds other attitudes as specified here, and the other of which is not a moral* belief because the agent who has that belief lacks those further commitments.[142] This may seem a little confusing at first, but we already think and speak in this manner often and without difficulty or confusion. One example of a commonplace phenomenon with similar features is conventional implicature, where a proposition can be said to conventionally implicate further commitments beyond its truth conditions.[143] Compare the following propositions:

i) Peter is rich and kind.

ii) Peter is rich but kind.

Both i) and ii) share the same truth conditions – both are true if and only if Peter has two properties simultaneously – richness and kindness. This is straightforwardly conveyed by i), and an agent who has a belief with the content that i) need not accept that they are thereby committed to any further attitudes regarding Peter, richness or kindness. Yet the word *but* in ii) conventionally implicates something more. This means that holding a belief that ii) requires believing that Peter has certain properties, but it also requires that the agent who holds the belief feels there is some kind of contrast between richness and kindness; that rich people are

[142] I draw here on Hare's discussion of the various ways in which the term 'good' may be used, especially 1952, p. 124-126.

[143] I will present only a very brief sketch of conventional implicature here, as I am merely using it as a model rather than trying to defend or attack it. For more on conventional (and other kinds of) implicature, and how the term has been used and developed over time, see e.g. Grice 1989, Copp 2001 & 2009 & Finlay 2017.

seldom kind. Holding a belief that ii) means the agent has a further commitment that such a contrast exists, and if an agent is not committed to the existence of that contrast, the agent does not hold a belief that ii). This is very similar to how I intend the term moral* to be understood – if an agent has a belief with the form of BELIEF, they must also have certain further attitudes in order for that belief to be a moral* belief. Those further attitudes do not alter the truth conditions of the belief in question, but the belief cannot be a moral* belief unless the further commitments are present among the agent's mental states. Having established this, I can now lay out what the further commitments I have in mind are.

5.3.4. Further commitments of moral* judgements

The first of the commitments which go along with BELIEF in moral* judgement is ACCEPTANCE, as referred to in the definition of BELIEF in §5.3.2. I will give a definition of ACCEPTANCE, and then begin to unpack it by picking out three features of moral* judgements which are promoted by ACCEPTANCE, and which I will argue help to maximise the prudential utility of morality* in the post-error-theory world.

Alongside having a belief with the content BELIEF, for an agent, *S*, to judge that it is morally* wrong to ϕ is for *S* to be committed as follows:

> ACCEPTANCE: *S* i) is disposed to be motivated or plans to refrain from ϕing, ii) feels appropriate reactive attitudes (e.g. blame) towards agents who ϕ, and iii) endorses a general (i.e. community-wide) policy of acting in accordance with a norm against ϕing because she believes that so acting is prudentially beneficial.

Each of the three elements of ACCEPTANCE plays a key role in helping to make it the case that adopting a practice of making moral* judgements would be plausibly beneficial. I will explain

each of these features in turn. The first element's role is to make it clear that, absent complicating factors such as depression (hence including 'is disposed'), moral* judgements should imply a motivation to act in accordance with the judgement.

It is easy to see why this feature of revolutionary relativism would be beneficial - what we are aiming for is a theory which *actually does* conduce to prudentially beneficial outcomes, rather than a theory which merely involves *S* believing that prudentially beneficial outcomes will likely be produced. ACCEPTANCE facilitates this by establishing a direct link between what it is to make a moral* judgement and being motivated to act in accordance with moral* judgements.

It is a matter of some controversy in hermeneutic metaethics whether moral judgements entail motivations, but on my proposal the matter is settled – in order for a judgement for be a moral* judgement, I stipulate that under typical circumstances, the judging agent is thereby committed to a motivation or plan to act in accordance with the norm around which the moral* judgement is based. This feature of moral* judgement cannot rule out akrasia or weakness of will, since agents may have other, stronger motivations, or they may be subject to depression, etc. But this feature nonetheless counts against giving into temptation by ensuring that judging agents have at least some motivation to act accordingly.

The second element ensures that moral* judgements should involve reactive attitudes such as praise and blame. That is, a moral* judgement that e.g. it is wrong to ϕ should retain the feature of the equivalent moral judgement that agents who ϕ thereby lay themselves open to blame and possible censure by others in respect of their having ϕed.[144] This is a very basic part of the meaning of traditional moral terms, and many metaethical theories assume

[144] I use the formulation 'in respect of their having ϕed' simply to relate the censure to the act of ϕing such that a person who e.g. tortures others is not as a result censured for some other thing, say, not having washed up (though they may be further censured for the latter).

something like this, even if they do not spell it out. For example, if torture is morally wrong, and someone tortures others, moral realists will virtually universally agree that other things being equal, that person thereby becomes a fit object of blame. Depending on the species of moral realism in question, blame may also be accompanied by other deserts such as outrage, correction, restraint or imprisonment.[145]

This feature of moral judgement plausibly aids coordination among agents by offering a mechanism for guiding agents towards certain kinds of behaviour and away from others, and is thus prudentially beneficial. A further benefit of this feature of moral judgement is that it also counts towards making agents motivated to perform acts evaluated as good, and to refrain from performing acts evaluated as bad or wrong. Put simply, the reactive element of moral judgement doesn't just tell agents who to blame or what they might be blamed for, it also makes agents want to act appropriately in order to avoid blame and censure, and (albeit possibly less strongly) to seek approval. This is not indefeasible – agents may say 'to hell with what others think', but it seems undeniable that the reactive element of moral judgement counts at least to some extent in favour of agents being motivated to act in accordance with moral judgements. By including reactive attitudes among the commitments of moral* judgements as a component of ACCEPTANCE, revolutionary relativism can retain these features of traditional morality and the benefits it brings.

A potential worry here is whether revolutionary relativists can choose, based on theoretical considerations, to feel appropriate reactive attitudes towards agents who act in certain ways.[146] For example, what if I believe it is in everyone's interests that people do not steal, but I do not blame those who do so? This is not a significant issue, however – whatever the

[145] The *locus classicus* here is Strawson 1962. For further discussion, see Eshleman 2016, especially §2.
[146] This parallels a worry I expressed about revolutionary expressivism in §4.5.3. There, it was problematic. Here, it is not, as I explain in the text.

status of what might be called 'desire-voluntarism', we are talking here about actions which likely promote or threaten other agents' wellbeing. Such actions uncontroversially arouse reactive attitudes – we standardly allot praise or blame to agents who act in ways which we feel promote or threaten our wellbeing. If you cut us, not only do we bleed, we also typically hold you responsible for cutting us! I would therefore argue that consistent with the definition of prudential reasons I gave in §5.2.3, proper reflection on the stealing-but-who-cares example would reveal that one of two unproblematic things is actually the case. Either the details of the theft in question mean that I do not actually judge that it was wrong, all things considered (for example, perhaps it was to feed starving children, and thus actually produced an aggregate prudential benefit), or I actually do feel some reactive attitudes about it after all (e.g. I may not have initially realised that the theft was harmful, but come to realise this and so feel aggrieved when I think more carefully about it).

The role of the third element of ACCEPTANCE is to support the motivational feature mentioned above by including endorsement. Endorsement here is intended to imply a sense in which the judgement in question is something the agent would 'stand behind' on a lasting basis. Naturally this does not mean that agents cannot change their minds about moral* judgements, or cannot make moral* judgements in the heat of the moment, but it carries an assumption that moral* judgements will be durable and appropriately serious judgements rather than snapshots of whatever the agent feels like at any given time. Thus iii) adds a degree of stability which the motivation element discussed above may lack.

Alongside this, the generalised aspect of iii) makes moral* judgements community-focused. Without element iii), all of the attitudes involved in moral* judgement so far could apply to purely personal matters which have no effect on other agents at all. Including the 'general policy' part of ACCEPTANCE allows moral* judgements to keep other agents firmly in view, and so emphasises the coordinative benefits revolutionary relativism can help to deliver.

There are at least two ways in which this community focus can be beneficial. One, extending the focus to other people beyond the individual agent can help emphasise that members of a moral* community may and prudentially should educate younger generations about the moral* code the community accepts. This aids the stability aspect I mentioned a moment ago, and also facilitates coordination among agents. And two, the community focus helps make the moral* norms accepted by a moral* community public. This need not mean that the norms are strictly codified and published, but it seems likely to increase general awareness of the relevant norms in comparison with a community in which behavioural norms were considered a private matter. Again, this is plausibly beneficial because it aids coordination among agents.

All of the above gives rise to a problem, however. Imagine that there is a moral* community which shares the moral* belief that it is wrong to *squanch*.[147] This means that agents in the community will typically have the BELIEF and ACCEPTANCE attitudes as regards squanching, and thus believe that it is prudentially beneficial to refrain from doing so. But suppose that in fact, given the local circumstances, they are mistaken - it is actually prudentially harmful to refrain from squanching, and everyone would be much better off if squanching were widespread.

In such a situation, the moral* attitudes I have recommended thus far could threaten to trap the community in their mistake. After all, so long as agents believe that their moral* practices are prudentially beneficial, neither BELIEF nor ACCEPTANCE suggests that agents must check to make sure the relevant policies *actually are* beneficial. Worse, features such as element ii) of acceptance may seem to count against agents changing their minds about which moral*

[147] I use a nonsense verb here so that the discussion is not coloured by attitudes we may have, but which the community in question may not.

beliefs might be prudentially beneficial, and nothing I have said so far concerns what might happen if agents are indeed mistaken about what to do for the best.

This is why my proposal includes one more commitment of moral* judgement. I propose that alongside having a belief with the content BELIEF, and being further committed to ACCEPTANCE, for an agent, *S*, to judge that it is morally* wrong to ϕ is for *S* to be committed that:

> GOOD CITIZEN: Should it become known that a policy of adhering to the practical norms accepted by her moral* community is prudentially suboptimal, and that a policy of adhering to different practical norms would be more beneficial, *S* will attempt to engage with her community in order to facilitate the acceptance of the new, prudentially optimal norms.

One of the purposes of GOOD CITIZEN is to highlight that acceptance of practical norms by members of moral communities is not a permanent phenomenon. While some features of moral* judgement lend themselves to doxastic stability – the community acceptance and reactive attitudes elements in particular – that stability must always be in the service of delivering prudential benefits. If a community comes to accept a moral* norm which is actually prudentially harmful, there needs to be some mechanism within revolutionary relativism which can change that situation. In concert with the conditionality aspect of ACCEPTANCE, GOOD CITIZEN means that revolutionary relativism provides for, and can indeed encourage, breaking down and replacing widely held first-order moral* beliefs which are found to be prudentially suboptimal.

In the current context, my desire to highlight this grows out of the need to avoid begging the question against abolitionists such as Hinckfuss. But the same concern has roots at least a

century older. I am sensitive to the concerns of those who sympathise with – or at least suspect there might be some merit in - Nietzsche's view that widely held traditional first-order moral beliefs may be very far from the optimal beliefs for encouraging human flourishing (e.g. Nietzsche 2007 p. 20). I believe we would be wise to share Nietzsche's suspicion of promoting a first-order moral (or in this case moral*) status-quo even as we become second-order moral error theorists. Rather, we must be wary of allowing ourselves to be merely, as he put it, 'wily spokesmen for [our] prejudices' (1998 p. 8). That is, we should be careful not to fall into the trap of insufficiently critically agreeing with those who make what Nietzsche and Hinckfuss would both see as a cosy and erroneous assumption: that traditional moral beliefs are, for all that they may be in error, still somewhere close to prudentially optimal.

I do not seek to accuse any of the philosophers under discussion here of anything specific or malicious,[148] but these thoughts ground the way in which I want, and feel well philosophically justified in doing so, to treat Hinckfuss' and wider abolitionist arguments more seriously than other WNP respondents frequently seem to. This is something I will elaborate on in the next chapter. For the time being, my explanation of revolutionary relativism is now complete. I will close out this chapter by offering some concluding remarks, and looking towards what will follow in the next two chapters.

5.4. Conclusion

To recap what happened in this chapter, in §5.1 I gave a summary of the overall picture so far, including highlighting the 'what now?' problem I had introduced in the previous chapter. I then discussed some preliminary considerations about how we might formulate a good

[148] Nietzsche may well accuse me of being too timid in this (see e.g. Nietzsche 1998 §1, *inter alia*), but I am expressing sympathy with a particular idea in his work here, not confessing agreement with the tone and content of everything he said!

response to the 'what now?' problem. In §5.2 I laid some groundwork for my response to the WNP by introducing two constraints on good WNP responses, and by defining in some detail what prudential reasons are and why we have them. Then in §5.3 I gave the formulation of my proposed response to the WNP, revolutionary relativism, and explained how adopting it could secure prudential benefits in a post-error-theory world. This involved fixing the scope of my proposal in §5.3.1, and then in §5.3.2-5.3.4 spelling out my proposal in detail – the difference between moral and moral* beliefs and the bearing this has upon truth conditions, the commitments involved in moral* judgement, and the role played by each of those commitments in delivering prudential benefits if my proposal were to be adopted. This completes my explanation of revolutionary relativism.

A slightly more detailed note on the use of the term 'revolutionary' than I gave at the very start of the chapter is appropriate at this point.[149] On one level, my proposal is revolutionary in that it applies only as a response to the WNP, which in turn arises as a result of accepting error theory. I am not proposing a hermeneutic view intended to explain traditional moral thought and discourse, but rather something which comes into play only after moral thought and discourse have been found to be systematically in error for something like the reasons I discussed in chapter 3. Such a finding is surely a revolutionary event, and using the term revolutionary to signal the context in which my proposal operates and to differentiate my view from hermeneutic forms of relativism seems entirely appropriate.

However, there are two further senses in which my proposal is revolutionary. One of these senses I mentioned above – that I am proposing that all agents everywhere alter their understanding and use of moral thought and discourse. Any change with such broad scope

[149] I am mindful here of Miller's admonishment that Joyce's revolutionary fictionalism is much less radical than claimed (2013 p. 121). I add these extra remarks here because I want to leave no doubt that my relativism is indeed revolutionary.

must merit the term revolutionary, even if it were subtle – and in theoretical terms, error theory and the possible responses to accepting it are far from subtle.

The other reason why my proposal is revolutionary is because I am not proposing the adoption of an 'off the peg' form of relativism. Rather I am proposing a novel form of relativism which has not been defended to date. This is because the revolutionary (i.e. post-error-theory) context of my proposal makes it possible and indeed appropriate for me to stack the deck in my favour, so to speak. That is to say, there are certain problems which traditional, hermeneutic forms of relativism are often thought to face which I can hope to avoid because the task here is not one of explanation of how things are, but rather suggestion of how things might be. As such, I have provided an entirely new – i.e. revolutionary – form of relativism.

Hoping to avoid objections is not the same as actually avoiding them, though, and I will still have to defend my proposal against the 'inherited' principal objections to hermeneutic forms of relativism. One of the best known examples is disagreement – if moral judgements are relative to an individual or a community, how can they be of any use in situations involving members of other communities? This objection confronts revolutionary relativism just as much as it does hermeneutic varieties of relativism. I will offer a defence against this objection and other 'inherited' objections on behalf of my proposal in chapter 7. There are also certain issues which arise for my proposal which may not arise for hermeneutic forms of relativism, either because of the post-error-theory context or because of the way I have constructed the view. Again, I will offer a defence against these context-specific objections in chapter 7.

None of that will matter, though, if I cannot make a strong case for why error theorists should adopt my view rather than one of the competing WNP responses – after all, who cares if there are interesting defences of my proposal if it is not one we have any reason to adopt? Therefore, before getting into the potential objections to my proposal, in chapter 6 I will offer

extensive arguments that revolutionary relativism is a better response to the WNP than the responses defended by others to date.

Chapter 6. Why Revolutionary Relativism Is the Best WNP Response

In this chapter and the next I will show why my proposal is a better response to the 'what now?' problem than the responses offered by others to date. I will begin by fleshing out the claim I made in §4.6 that WNP respondents to date have failed to give sufficient weight to the challenge from abolitionists. Thus §6.1 will contain my argument for taking abolitionism more seriously. Briefly, I believe that all WNP responses must be compared under two scenarios: one in which abolitionists are wrong, and morality is prudentially beneficial (at least on balance), and another in which abolitionists are right, and traditional morality is harmful (at least on balance). It was not necessary to discuss this in chapter 4, as my focus there was on arguing that we should reject previous WNP responses on their own merits. All non-abolitionist WNP responses assume that the abolitionists are wrong about traditional morality, at least on balance, and so it was appropriate in chapter 4 to go along with this. But, for reasons I will explain in §6.1, to give a full account of why my proposal is superior to others, I will argue here that a twin-track case must be made, explicitly including the possibility that abolitionists are right and traditional morality is indeed harmful.

I will then turn in sections 6.2-6.5 to showing why revolutionary relativism is a better response to the 'what now?' problem than the available alternatives in both of the scenarios laid out in §6.1. I will argue that revolutionary relativism is superior to competing proposals because it can either avoid or cope better with my own objections to other proposals (which I presented

in chapter 4) and also with the key objections others have raised in the literature. My strategy here will be to offer a version of a dominance argument.[150]

Each of the sections 6.2-6.5 will tackle one of the WNP responses I rejected in chapter seven. In each section, I will give a reminder of the proposal itself. I will then provide a brief precis of each reason I gave for rejecting the proposal in question, and an argument that revolutionary relativism either avoids the criticism, or copes better with it than the proposal in question can. I will also discuss the implications of the distinction drawn in §6.1 for the proposal in question. Thus in each section we will see that revolutionary relativism is a better response to the WNP than the competing response under discussion. And when taking the chapter as a whole, even if there may seem to remain a glimmer of hope for the other proposals after my arguments in chapter 4, we will see that regardless of whether traditional morality is harmful, and whichever existing WNP response we might consider adopting, we will always be better off if we adopt revolutionary relativism. As a result, I will conclude that revolutionary relativism is the best response to the WNP overall.

In the overall structure of my project, this chapter will constitute the first stage of my defence of revolutionary relativism. In the next chapter, I will present the second stage of my defence, and show how my proposal can also cope with the most significant problems which confront traditional relativist theories, and can therefore stand on its own two feet, philosophically speaking. I will also discuss some problems which arise for revolutionary relativism specifically

[150] Dominance arguments are familiar in game theory, but for examples of variations on dominance arguments being used in philosophy, see e.g. Sayre-McCord (2013) or, famously, Pascal's Wager. Very briefly, dominance arguments can apply when an agent makes a choice between possible courses of action without knowing which of two or more situations they are in (or will imminently be in). Where one course is preferable to the other(s) regardless of which situation obtains, then that choice is said to superdominate the other(s). For example in Pascal's Wager, the unknown situation is that either 1) God exists or 2) God does not exist. Pascal argues that regardless of God's existence, we are always better off believing that God does exist (and acting accordingly). Thus Pascal's conclusion can be expressed as the claim that belief in God superdominates atheism.

in the WNP context, and show that they cannot derail my proposal. This will complete my case for revolutionary relativism, and will be followed by the concluding chapter of the thesis.

6.1. The challenge from abolitionism and 'moralbad'

Recall from section 4.3.2 that one of the main problems I raised with abolitionism is that it rests on a claim which cannot plausibly be proven: that traditional morality is, on balance at least, harmful to our wellbeing. If we are not persuaded by this claim, it seems we can simply set abolitionism aside. If the motivation for abolitionism is the harmfulness of traditional morality, and traditional morality is in fact not harmful, then the motivation for abolitionism is wholly undermined.

This is what most philosophers who have responded to the WNP have argued.[151] Olson's short dismissal of abolitionism centres largely around one sentence: 'My suspicion, though, is that moral discourse is at least potentially more beneficial than detrimental to human and non-human well-being' (2014 p. 180). Joyce spends a little more time discussing abolitionism, but ultimately comes to a similar conclusion – moral beliefs may have occasionally resulted in tragedy, but they are nonetheless useful (2001 p. 184-5). Nolan *et al.* go so far as to doubt that jettisoning moral thought and discourse is something we could easily do, if at all (2005 p. 307).[152]

[151] An anomaly among WNP respondents here is Svoboda, who characterises abolitionism as being motivated primarily by epistemological worries about error theorists having false beliefs, rather than by the worry that traditional morality is prudentially harmful (2015 p. 9). Abolitionists do of course discuss this dimension of abolitionism (e.g. Garner 2007 p. 508), but it is hardly the only point they raise. Thus rather than presenting a view of the issues involved which competes with the one I have offered in the text (and which I would therefore need to argue against), Svoboda's characterisation seems to me to rather miss the point, or at least one of the salient points, of abolitionists' arguments. After all, other responses offer freedom from error *plus* other benefits, so why be an abolitionist unless all the alternatives are untenable, including Svoboda's own proposal? That being so, I merely mention the matter and move on, rather than digressing further.

[152] Some philosophers claim that it may actually be impossible to abolish moral thought & discourse, see e.g. Streumer 2013a & Strawson 1962, particularly §4.

Far from being a cynical whitewash on the part of non-abolitionist WNP respondents, the above is quite credible. For example, imagine a survey where participants were asked whether it was more likely that traditional morality a) beneficially helps us combat weakness of will (*cf.* Joyce 2001 p,184) or b) inexorably leads to harmful elitism, authoritarianism and war (*cf.* Hinckfuss 1987 especially chapter 3). It is entirely believable that a majority would answer a).

In light of this, the fact that I have proposed a response to the WNP which includes retaining (anything substantially similar to) elements of traditional morality, such as making apparently moral judgements, might make it appear that I intend to make a similar case for the superiority of my proposal over abolitionism. First, I could reject the claim that traditional morality is harmful, at least on aggregate. And I could then argue that by definition, abolitionism cannot retain any of the benefits of traditional morality, because eliminating moral practice must obviously entail eliminating the beneficial effects of moral practice. Thus I might argue that my proposal is superior to abolitionism to the extent that it can deliver (any of) the benefits of traditional morality. This line of argument is just as convincing when deployed in support of revolutionary relativism as it is when others use it (*cf.* Olson 2014 p. 181).

But I will go further, and I believe that other non-abolitionist WNP respondents should similarly respond to abolitionism more thoroughly than they frequently do. This is because the unprovability problem with abolitionism cuts both ways – it is implausible that we could prove whether traditional morality leads to harm *or not,* since we have no comparison world in which no agent ever makes a moral judgement or engages in moral discourse. We can assume, since our history includes morality and the current world is still here and we are still in it, that any harmful effects of morality have not proven *catastrophic* on a global scale.

However we cannot be certain whether we would be significantly better off now, had humans never made a moral judgement or expressed a moral belief.

And this issue persists into the WNP context. Hinckfuss' central point in his 1987 was surely that many of the moral beliefs which we previously thought were beneficial *turned out after critical reflection* to have harmful effects (again, see Hinckfuss 1987 chapters 2 & 3). This is not an *a priori* matter or an outcome which could easily have been foreseen. The harmful effects of traditional morality need to be pointed out to us after the fact, otherwise there would be no need for Hinckfuss' book.

Therefore let us put aside for a moment the intuition that traditional morality is beneficial (at least on balance), and consider what would follow if the claim that traditional morality is harmful were true. For ease of exposition, I will label this situation 'moralbad', defined more precisely as the epistemic possibility that our conventional moral beliefs and discourse could be such that our lives would go better without them. The converse situation in which traditional moral beliefs are beneficial (i.e. as is presupposed by the other WNP responses) I will label 'moralgood'. I would argue that WNP respondents must do one of two things. Either i) they must provide a convincing argument that it is implausible that we currently are, or ever could be, in moralbad, or ii) they must provide an account of how their view would fare in moralbad, should it be the case. Simply disagreeing that traditional morality is harmful is inadequate. Despite the cursory examinations of abolitionism mentioned above, none of the proposals under discussion adequately include either of these two elements.

What follows from this is that non-abolitionist responses to date have failed to answer the challenge from abolitionism adequately. In order to give proper weight to this context, I will need to consider the superiority of revolutionary relativism in moralbad as well as the more usual moralgood situation.

What I want to highlight is that abolitionism is only one response among several to the possibility that we are currently in moralbad. The key to seeing this is to note that abolitionists such as Garner do not advocate eliminating all normative reasons. Rather, abolitionists recommend that we abolish traditional morality for prudential reasons. They argue that we would be better off if we didn't have the moral beliefs we have, and being better off is something most people want (see e.g. Garner 2007 & my footnote 76, above). Thus it is at least possible that we might do away with the kinds of moral beliefs which both a) are false according to error theorists, and b) lead to the negative outcomes abolitionists suggest, and yet still retain practical rules by which we regulate our behaviour. Any view which can accomplish both a) and b) can respond to both the WNP and moralbad.

As I said at the beginning of this chapter, I intend to offer a dominance argument. I believe that regardless of whether moralgood or moralbad obtains, when compared with any competing WNP response, revolutionary relativism is always the better choice. To defend this claim, I will discuss the available alternatives, conservationism, revolutionary fictionalism, abolitionism and revolutionary expressivism in sections 6.2-6.5 respectively. In each case I will explain why revolutionary relativism is a better response to the WNP both in moralbad *and* the more usual moralgood situation.

6.2. Revolutionary relativism versus conservationism

In section 4.2, I discussed and rejected conservationism, principally defended by Jonas Olson, as a response to the WNP. To recap, Olson argued that we can and should respond to the truth of a moral error theory by choosing to retain our moral beliefs, even though we now know them to be systematically false. We should do so, Olson claimed, for prudential reasons, i.e. because we are likely to be better off in a society in which people have genuine moral beliefs, and act in accordance with them, than we would be in a society in which no one had

such beliefs.[153] Olson argued that, despite certain intuitions we might have to the contrary, we can consistently have genuine moral beliefs and remain moral error theorists. This is because, Olson argued, we can compartmentalise our beliefs – we can genuinely hold certain beliefs at certain times whilst remaining disposed to dissent from those beliefs at other times.

In section 6.2.1 I will assume that moralgood obtains, and show why revolutionary relativism avoids the objections to conservationism I discussed previously, and is therefore preferable to conservationism in moralgood. In section 6.2.2 I will explain why revolutionary relativism is also preferable to conservationism if moralbad obtains.

6.2.1. Revolutionary relativism vs. conservationism in moralgood

I discussed two main objections to Olson's proposal. The first, drawing on Suikkanen 2013, was that the best accounts of belief available to us from the philosophical debates concerning the nature of propositional attitudes suggest that beliefs must by definition be sensitive to evidence.[154] But remember that we are considering how we *as error theorists* should respond to the WNP. Error theorists as such will take themselves to have conclusive evidence that our traditional moral beliefs are systematically false. This puts Olson in a very awkward position. Either he must offer a highly unconventional theory of propositional attitudes which can accommodate evidence-insensitive beliefs, and which is more plausible than the best theories of beliefs we currently have, or he must abandon the claim that we can consistently and rationally retain genuine moral beliefs after we have become convinced that a moral error

[153] In the terminology I am employing here, and given remarks such as '…moral discourse is at least potentially more beneficial than detrimental to human and non-human well-being' (Olson 2014 p. 180)), we can infer that conservationism assumes that moralgood currently obtains. See also §6.2.2 below.

[154] Throughout this subsection, it will be helpful to recall the discussion of the nature of beliefs versus other propositional attitudes I gave in §2.1.2, and the development of these issues in §4.2.1. For quick references on the evidence-sensitivity of beliefs, see e.g. Smith 1994, especially §4.6, pp. 111-116 & Schwitzgebel 2019, especially §1.4.

theory is true. Since Olson does not offer a satisfactory theory of what beliefs are which is compatible with his proposal, it seems that he must abandon conservationism.[155]

Revolutionary relativism is not vulnerable to Suikkanen's line of criticism, because my proposal requires only a standard account of beliefs.[156] To see why, consider the differences between our moral beliefs *before* we accept an error theory and the beliefs Olson and I propose we should have *after* we accept an error theory. We may assume that typical error theorists would agree that before we become error theorists, our moral beliefs conform to the standard account of beliefs. I say this because in arguing for an error theory, when error theorists describe our traditional moral beliefs, they typically use terms such as 'beliefs' without presenting arguments that they intend a non-standard reading of the term. There is also evidence for the claim that traditional moral beliefs are indeed evidence-sensitive. For example, we sometimes change our moral beliefs in response to convincing arguments or other evidence. Thus before we accept an error theory our moral beliefs have a mind-to-world direction of fit, and so are sensitive to evidence.

In summary, then, error theorists typically hold that before we accepted error theory, moral judgement consisted in having an attitude of belief towards propositions which ascribed mysterious non-existent properties to certain actions. And according to the best accounts of

[155] As I noted in §4.2.1, Olson has recently responded to this kind of criticism (2019). Olson claims that we can suppress beliefs by 'advertently not attending to evidence supporting moral error theory' (p. 310). Yet even if we can do so, I suspect this fundamentally threatens the stability of error theory itself (for reasons very like those I gave in §4.2.1). Olson also claims (p. 309) that that there can be degrees of belief, and so we can believe in error theory partially, but also have partial moral beliefs. I find it plausibly more accurate to say that we can have varying degrees of certainty that a belief is correct. If so, what Olson describes is merely a matter of not being sure what to believe. A familiar predicament, sure enough, but this fails to be a picture of how we can rationally maintain genuine moral beliefs at the same time as genuinely accepting a moral error theory. As such, I do not believe Olson's arguments work. But in the present context it would not be especially helpful to digress into a full discussion here, only to end up no further forward. So I set Olson's recent response aside for the time being, though I note that the issue may bear revisiting, especially if he is able to expand on the matter further in future.

[156] For a further overview and discussion of beliefs, which includes discussion of various authorities in the field (hence the claim above that it is a standard view), see Humberstone 1992.

propositional attitudes we have, if an attitude we have towards some proposition is a belief, then insofar as we are rational we tend to give up that belief if we encounter convincing evidence that the belief is false. Now, in the 'post-revolution' WNP context, I am proposing that we make moral* judgements which consist in having an attitude of belief towards propositions about norms which our moral* communities accept (and so on, as described in detail in §5.3). This depends on an entirely standard account of belief-attitudes – most saliently, I am proposing having attitudes which are sensitive to evidence, and which we give up if we encounter convincing evidence that they are false. One could even say that the beliefs I propose we have are roughly the same as some beliefs we already have, given that there is evidence that some of our traditional moral beliefs are relativistic.[157] So the tension between the evidence-sensitivity of pre- and post-error-theory models of belief which Suikkanen objects to in Olson's case simply does not exist on revolutionary relativism.

The second objection to conservationism I discussed in chapter 4 was my own objection that Olson failed to make a successful case for the strand of his argument which I summed up thus:

> OC3. For a proposition, *p*, we can have an occurrent belief that *p* in one context, while simultaneously being disposed to believe that *not-p* in other contexts.

My concern was that, for beliefs of the kind under discussion, this is simply not plausible (see §4.2). There is *a sense* in which something like OC3 is true. We are indeed sometimes disposed to change the beliefs we have in context *A* upon entering context *B if context* B *exposes us to evidence that the belief we had in context* A *was false*. The evidence need not be particularly good evidence – we may be persuaded by convincing oratory or by strong

[157] I refer here to Goodwin and Darley (2008), whose experiments show that some 'lay people' have relativistic as well as objectivist intuitions about morality. I will return to this in more detail in §6.3.1 and §7.4. Note that this is entirely consistent with error theorists' typical understanding of traditional morality – the 'lay' people could simply be wrong.

emotions to believe things which we cease to believe upon the most cursory reflection later on. But we do nonetheless sometimes find ourselves temporarily believing things which we do not believe most of the rest of the time. Thus it could be said that although we may believe that *p* while we are in context *A*, we are simultaneously disposed to believe that *not-p* in context *B* in this way. This, I argued, is what is going on in the examples Olson gives to demonstrate the plausibility of his view.

But Olson needs more than this, for two reasons. One, Olson needs us to be able to change from believing that *p* to believing that *not-p* not because of a change in evidence between different contexts, but because we choose to feel more reflective in one context than another.[158] And two, when Olson talks of a disposition to believe different things in different contexts, he does not mean having distinct beliefs at different times in the manner I just described. He means something much closer to having two contradictory beliefs at once, but only *paying attention to* one of them at a time: 'In such cases, the more reflective beliefs are suppressed or not attended to' (2014 p. 192). This is an entirely different phenomenon that that captured by OC3.[159] I argued that neither of these further steps is plausible, nor are they supported by the examples Olson provides.

By contrast, I do not need anything like OC3 in order for my proposal to work. Revolutionary relativism requires only that agents' beliefs track what agents take to be the best evidence available to them. Having begun as moral realists, we were exposed to the arguments for error theory, and thus to what we took to be convincing evidence that our moral beliefs were

[158] I take it that the volitional aspect is clear here. Granted, we may be guided by external influences to become more reflective in some contexts (such as the philosophy seminar room). But if we can move from believing that *p* to believing that *not-p* as a result of entering 'more reflective and detached contexts' (2014 p. 192), then surely we can also unilaterally shift our context in the relevant way by simply *choosing* to think carefully about the matter.

[159] Lest it be thought that I have mischaracterised Olson's view in OC3 in order to make it easier to object to, I offer the following quote: 'In general, it does not seem impossible simultaneously to have an occurrent belief that p and a disposition to believe not-p in certain contexts' (2014 p. 192).

false. That being the case, as revolutionary relativists we would cease to believe as we previously did, and instead we would form new beliefs which are in line with the available evidence. Nothing about the beliefs on which revolutionary relativism relies has been shown to be false by the arguments for error theory, and so no compartmentalisation or maintenance of false beliefs is required for an error theorist to become a revolutionary relativist. Revolutionary relativism does not require that we develop a previously unmanifested ability to hold genuine beliefs at will. Neither does it require us to be able to simultaneously hold mutually contradictory beliefs, yet to only pay heed to one of them in any given context. In this way, revolutionary relativism is preferable to conservationism because it avoids reliance on implausible models of belief.

In summary, conservationism assumes moralgood, and in moralgood it is vulnerable to at least two criticisms described above and in §4.2.[160] Revolutionary relativism avoids both these criticisms. Thus I conclude that, even if we grant Olson's claim that traditional morality is beneficial on balance, revolutionary relativism avoids the criticisms which render conservationism untenable and is therefore the preferable WNP response.

6.2.2. Revolutionary relativism vs. conservationism in moralbad

Turning to the situation I have labelled moralbad, my argument is very straightforward indeed. It is impossible for conservationism to be prudentially beneficial if moralbad obtains. This is because retaining our traditional moral beliefs means that we will also retain any negative consequences of regulating our behaviour in accordance with those beliefs. Recall that the motivation for conservationism is that retaining traditional moral beliefs is supposedly

[160] For further criticism of conservationism, see Jaquet & Naar 2016. I have omitted discussion of whether revolutionary relativism can cope with their criticisms of conservationism in the text above because their argument takes place only after accepting something like OC3. Since I reject Olson's argument before it gets that far, Jaquet & Naar's criticisms of conservationism do not apply to my proposal.

prudentially beneficial. Olson's argument is that if we accept a moral error theory, our lives will go better if we continue to have traditional moral beliefs – even though we know they are false - than they would if we had any other (or no) kind of moral beliefs. In moralbad, it is understood that having traditional moral beliefs is harmful. Therefore retaining those moral beliefs will also be harmful, and the prudential motivation for conservationism is entirely undermined.

Thus conservationism cannot be of net benefit in moralbad, and therefore depends entirely on the claim that traditional morality is of net benefit. Unless we are given much better reasons to rule out the possibility of moralbad than have been offered to date – and as I argued in §6.1, it is unlikely that sufficiently convincing reasons could be offered – conservationism fails in this scenario. Alternatively, conservationists could try to come up with some kind of selective variant of their position, according to which only prudentially beneficial moral beliefs were retained. However, this would likely be a fraught affair since it would result in a hybrid of conservationism and abolitionism. And even if this were attractive to conservationists, there would be significant difficulties involved in showing which moral beliefs were beneficial and which were not – as I argued in §6.1 above, this latter issue is precisely why we must consider moralbad as I am doing here.

To illustrate this, consider Garner's claim that 'morality inflames disputes and makes compromise difficult, it preserves unfair arrangements and facilitates the misuse of power, and it makes global war possible' (2007 p. 502).[161] Remember that we are now comparing conservationism and revolutionary relativism *in moralbad*, and so we are granting for the time

[161] Garner is not alone here. In the quoted text he is overtly drawing on Mackie. And as I discussed in chapter 4, Hinckfuss also argues at length for a similarly dismal view of morality (1987). As a more recent example, Goodwin & Darley (who are neither error theorists nor abolitionists, as far as I am aware) suspect that '(c)ulture wars that are fought over fundamental ethical values may become more intractable to the extent that each side of the debate harbors an objective view of the truth of its own beliefs' (2008 p. 1361).

being that having traditional moral beliefs does cause these prudentially harmful effects. That being the case, then responding to the WNP by advocating the wholesale retention of the same first order moral beliefs we had before we became error theorists must necessarily lead to the same negative effects.

By contrast, revolutionary relativism is largely neutral on which actions should be forbidden or required by the moral* standards of a given moral* community, and hence which moral* beliefs are true or false. I am proposing only that the moral* beliefs we should adopt are beliefs about which actions are forbidden or required by the moral* standards accepted by a moral* community. The only constraint imposed by revolutionary relativism on what those standards are is that acting in accordance with them should plausibly be of net prudential benefit to the community. To be as explicit as possible: on revolutionary relativism, the ascription of moral* beliefs to agents presupposes or requires that the agents in question have certain commitments. These include a commitment to the principle that a generalised policy of acting in accordance with the relevant moral* standard will prove to be prudentially beneficial.[162] Accordingly, it is a requirement of sincerity that utterances which express moral* beliefs can be understood to pragmatically express this commitment.[163]

This, of course, provokes questions about what exactly the relevant commitment is, and what follows from agents being so committed. As I laid out in chapter 5, a commitment in the sense I have in mind involves a set of attitudes. In the present context there are two salient features of the attitudes involved. First, there is element i) of ACCEPTANCE: that agents who make moral* judgements thereby commit themselves to being disposed to be motivated or planning to act in accordance with the judgement in question. My proposal must include this in order to account for agents' beliefs actually resulting in action – without something like

[162] See §5.3.4.
[163] See §5.3.2.

this, it may be controversial whether moralbad could obtain in the first place because it would be unclear that agents' moral beliefs could result in prudential harms. Including element i) of ACCEPTANCE in my proposal puts this beyond doubt. And second, there is GOOD CITIZEN: a plan or intention that, should it become known that there is an available, more prudentially beneficial norm, agents will be 'good moral* citizens'. This means that agents will engage with one another to reappraise the norms they accept, in an effort to come to accept only those norms which plausibly conduce to the optimal prudential outcomes.

With the above established, I am now in a position to offer a story of what would follow if we were to adopt revolutionary relativism and then realised that we were in moralbad. Discovering that holding or acting in accordance with a given moral* belief leads to harm would make it impossible for us to keep the moral* code (i.e. the system of moral* norms the community accepts) of our community unchanged, specifically because of our commitments to ACCEPTANCE and GOOD CITIZEN. According to my proposal, this would trigger a process of reflection aimed at restoring equilibrium between the moral* norms the moral* community accepts, the beliefs of members of the moral* community, and the commitments implicated by those beliefs. The possible outcome of this process which is most relevant to the present argument can be laid out as follows: 1) the moral* community would cease to accept the standard in question, acting in accordance with which had been shown to be harmful, 2) the relevant beliefs would therefore go from being true to being false, 3) the moral* community would accept a new standard, acting in accordance with which *was* prudentially beneficial, 4) new, true beliefs would be adopted.[164]

[164] I say that new true beliefs *would* be adopted (as opposed to *could* be) because I take it that the considerations around direction of fit I discussed §2.1.2 (and revisited above in §6.2.1) mean there is a natural tendency towards having true beliefs about matters on which evidence is available. To analyse this prediction in more detail would involve too great a digression here, but I consider the point sufficiently intuitively plausible to assume it here.

To see how this might work in practice, let us take a concrete example, and consider the morality* of private property. Consider the principle that *owning private property is morally permissible*. Pre-error theory, according to typical error theorists, to accept this principle was to have a belief, call it B:

> B [*owning private property is morally permissible*]

After we come to accept error theory and adopt revolutionary relativism, to accept the same principle would involve having a slightly different belief:

> B* [*owning private property is permissible according to the moral* standards accepted by our moral* community*]

Alongside having this belief, accepting the principle would also involve having the further attitudes described by ACCEPTANCE and GOOD CITIZEN. For present purposes, the relevant features of these further commitments can be summed up by saying that that agents who sincerely believe that B* are committed to the principle, C, that:

> C [*acting in accordance with the standards referred to in B* is plausibly prudentially beneficial, and if we find that it is not, we will accept new, prudentially optimal norms instead*]

Note that the phrasing here is not a carbon copy of the terminology I used in §5.3. The individual, *S*, is replaced with a community, the action involved is deemed permissible as opposed to wrong, and elements of both ACCEPTANCE and GOOD CITIZEN are brought together in C. This is intentional, in order to show how revolutionary relativism would work in practise, rather than in terms of verbatim, textbook-style definitions. But it should be clear that B* and C amount to a description of the practical application of the attitudes laid out in

§5.3.

Now imagine that it is shown to be the case that abolitionists are right about the morality of private property. In this situation, the way most people currently think about property and theft actually entrenches inequality and leads to deprivation for the many while ensuring that disproportionate resources find their way only to the powerful few.[165] I will assume that this is an uncontroversially negative outcome in prudential terms. What happens next? Remember that conservationism is weak here because it makes no provision for doing anything other than persisting with false moral beliefs inherited from our pre-error-theory selves, and so perpetuating any associated harms. However, were this harm to be discovered within a revolutionary relativist community, it would not of itself render B* false. But it would violate the relevant commitment, C. This in turn would lead to a reappraisal of which moral* standards it was prudentially beneficial to accept, and to the accepting of new standards which *were* beneficial. Following this reappraisal, the belief B* would become false, and a new true belief, B** would be adopted:

> B** [*owning private property is impermissible according to the moral* standards accepted by our moral* community*]

Alongside B**, accepting the relevant principle would involve having further commitments which we can sum up as:

> C* [*acting in accordance with the standards referred to in B** is plausibly prudentially beneficial, and if we find that it is not, we will accept new, prudentially optimal norms*

[165] I take this to be close to what Garner has in mind when he mentions the preservation of inequality and the misuse of power (2007 p. 502). Hinckfuss takes a similar point further, seeing the same mechanism as potentially resulting in revolution (1987 especially §2.5 and §3.5). The history of Russia in the early 20th century would seem to go some way towards supporting his view.

instead]

Note the near-identical nature of C and C*. As a result of the process described, equilibrium is restored between the standard accepted by the moral community, the true beliefs of members of that community, and the relevant further commitments of the members of the community.

In summary, we can see that in moralbad conservationism condemns us to prudentially negative outcomes. Whereas revolutionary relativism offers scope for the revision of our beliefs in order to avoid harm and to promote prudentially beneficial outcomes. I therefore conclude that in moralbad, revolutionary relativism proves superior to conservationism as defended to date, and seems likely to remain so. Taking sections 6.2.1 and 6.2.2 together, I conclude that revolutionary relativism is a better response to the WNP than conservationism.

6.3. Revolutionary relativism versus revolutionary fictionalism

In §4.4 I discussed and rejected the proposal that error theorists should adopt revolutionary moral fictionalism. Joyce proposes that error theorists retain aspects of traditional morality because they are useful, principally the feature of traditional morality which he argues helps us combat weakness of will. This can be done while avoiding error, Joyce argues, by taking a fictive stance towards morality – that is, by continuing to speak and think largely as moral realists do. But instead of holding and expressing genuine moral beliefs, Joyce proposes that we should *entertain* moral *thoughts* (a.k.a. make-beliefs or quasi-beliefs), and that our apparently moral utterances should not be assertions, but rather quasi-assertions. For Joyce, these moral thoughts are largely the same as genuine moral beliefs, excepting that we will remain disposed to dissent from them in sufficiently critical contexts. This latter exception means, according to Joyce, that our moral 'thoughts' would not be moral beliefs, and

therefore we would avoid the error of having false beliefs. So long as we thoroughly immerse ourselves in this practice, thinking and speaking as if moral realism was true in all but our most critical context(s), Joyce believes that this will mean that when confronted by weakness of will, we will nonetheless be motivated to do the right (prudentially speaking) thing.

As was the case with conservationism, Joyce assumes that we are currently in moralgood. As I argued in §6.1, I believe this cannot be assumed. Thus I will argue that revolutionary relativism is a better response to the WNP than revolutionary fictionalism in both moralgood (§6.3.1) *and* moralbad (§6.3.2).

6.3.1 Revolutionary relativism vs. revolutionary fictionalism in moralgood

There were several reasons why I rejected Joyce's proposal: i) revolutionary fictionalism necessarily involves dishonesty in problematic ways, ii) fictionalists cannot satisfactorily specify which moral quasi-beliefs we should adopt, and iii) error theorists cannot adopt the required fictive stance towards moral beliefs in a way which preserves the immersion on which revolutionary fictionalism depends. In this subsection I will tackle each of these objections in turn, first giving a brief recap of the details of my objection, and then showing why revolutionary relativism fares better than revolutionary fictionalism does.

On dishonesty, I argued that we might see Joyce as addressing his advice to individuals, to small groups, or to everyone, everywhere. In each case, either dishonesty follows (which Joyce himself says is unacceptable (2001 p. 214)) or the proposal becomes self-undermining. Remember from chapter 2 that it is a commitment of Joyce's error theory that most people intend and interpret typical moral utterances assertorically. Remember also that error theorists typically agree that most people typically use moral discourse to express beliefs. Yet fictionalism is defined by the non-assertoric use of apparently moral language, which in turn

means that fictionalists do not express beliefs when they participate in apparently moral discourse. This has the upshot that the lone fictionalist will intend to withhold assertoric force from their apparently moral utterances, but their audiences will typically interpret those utterances as assertions about the fictionalist's beliefs. According to a quite standard definition, this means that fictionalists would unavoidably be lying.[166] Thus the source of my dishonesty worry was Joyce's recommendations as regards assertoricity and the genuine expression of beliefs.

Moreover, since meaningful conversation requires both parties to be able to understand what is being said in the sense in which it is intended, individuals cannot unilaterally withhold assertoric force (as Joyce observes himself, 2017 p. 82). And since revolutionary fictionalists would know this, on an individual level, Joyce's proposal seems unintelligible. Turning to groups, this may make fictionalism tenable within the group, but I observed that groups will always come up against outsiders, and so the problem faced by the lone fictionalist will always loom. And if Joyce proposes that everyone, everywhere becomes a fictionalist, this simply cannot be done – not everyone will be able to grasp the required material, and even if they could, the inculcation costs could well be so enormous as to outweigh the prudential benefits which were the motivation for Joyce's proposal.

Virtually the same dishonesty objection could be made to my theory. Theoretically, if error theory is true, then the ideal outcome from the point of view of my proposal is that everyone, everywhere should become a revolutionary relativist. But even if the error theory is true and my proposal is inarguably the best response to that truth, this will not happen instantly. So for the foreseeable future anyone who adopts my proposal will frequently come up against

[166] For ease of reference, the definition of lying I used in §4.4.3 was 'to make a believed-false statement to another person with the intention that the other person believe that statement to be true' (Mahon 2016).

people who are not (yet, at least) revolutionary relativists. It is a commitment of error theory that most people are not currently moral relativists, since arguments for the error theory typically target the supposed objectivist features of moral facts or moral normativity.[167] Thus even if we set assertoricity and genuine beliefs aside, it could still be objected that when speaking about moral matters with others, revolutionary relativists will mean one thing while their audience takes them to mean another. To make matters worse, it is likely that the revolutionary relativists would know that this was the case, while their audience would not. Therefore, the objection runs, dishonesty is still a problem for my proposal just as it was for fictionalism.

To make this objection explicit, my opponent could say that when a revolutionary relativist, let's call her Cassandra, utters 'torture is wrong', she expresses a belief, B_{RRev}, with roughly the content that:

> B_{RRev} [*torture is forbidden because it contravenes the practical norms accepted by my moral* community*]

When others who are not revolutionary relativists utter the same sentence, my opponent could observe that according to error theorists they will typically be expressing a different belief, B_{FOLK}, which can be understood as having roughly the content that:

> B_{FOLK} [*there are non-institutional categorical reasons to refrain from torture*]

Thus, when Cassandra says 'torture is wrong' to Helen, who is not a revolutionary relativist, it is clear that there is a mismatch between Cassandra's intended meaning and the meaning

[167] See chapter 2. Revolutionary expressivists would also face a similar problem, and I suspect it would be even trickier for them to respond to it. However, since I reject revolutionary expressivism for other reasons, I set aside further discussion of this point.

Helen infers from Cassandra's utterance. Plus, as a self-aware revolutionary relativist, Cassandra is likely to know this. So there arguably *is* a dishonesty worry here, and Cassandra may be lying, or at least talking past her audience. Indeed, it may seem that Cassandra cannot help but intend Helen to come to have a belief which Cassandra deems false. The challenge for me is therefore whether I can successfully argue that when talking with outsiders, despite Cassandra expressing beliefs with different contents to those inferred by her audience, there is sufficient agreement between the parties about what is being said to support the claim that the parties are not lying or talking past one another in a problematic way. And in order for this answer to be satisfying in the present context, it must also be an argument which is unavailable to fictionalists. I believe I can respond in a way which does both of these things.

The first thing to note is that I see no way around conceding that there are differences between Cassandra's intended meaning and the meaning inferred by others, and this means that to some extent, Cassandra is talking past her audience. My argument is instead that the relevant differences are so slight as to be unproblematic, and may in fact occur routinely in traditional moral discourse without incident.

Several features of moral* discourse according to my proposal support my claim that the differences in meaning are only slight. On my proposal, moral* discourse is typically used to sincerely express genuinely held beliefs. Moral* discourse is also typically assertoric. And while B_{RRev} and B_{FOLK} have different contents, they are alike in one important respect – they both have practical import. That is to say, they are both beliefs concerning what is to be done.

This is one reason why I can claim that Cassandra is not lying when she says to Helen 'torture is wrong' – Cassandra does not intend Helen to believe B_{FOLK} (which would make Cassandra a liar). Rather, in everyday contexts, Cassandra is not interested in whether Helen comes to have a belief which could be technically analysed as B_{RRev} or B_{FOLK}. She cares only about what

Helen's practical stance towards torture is (or the practical stance of people in general). Therefore Helen and Cassandra may have technically different beliefs, but their beliefs are very similar in that they have the same practical import.

Now it could be replied to this that fictionalists can claim virtually the same thing – when they utter apparently realist sentences, they don't care whether their audience comes to have realist beliefs, so long as their audience responds appropriately on a practical level. And to an extent, this would likely be correct. But under scrutiny, this reply falls apart. On my proposal, Cassandra's view can be made more precise by saying that although on an everyday level she is neutral as regards the technical details of Helen's beliefs, on a metaethical level, she would prefer Helen to have B_{RRev} than B_{FOLK}. That being the case, Cassandra is free – and perhaps is even motivated to some extent – to have metaethically critical discussions and explain this to anyone willing to engage with her, without in any way threatening the stability of her own position. By comparison, fictionalists must guard against doing so in everyday contexts for the sake of preserving their vital immersion in the fiction. Therefore they cannot help but intend that their audience will have B_{FOLK} most of the time.

The similarities I have described make it the case that when specifying exactly what she means by utterances such as 'torture is wrong', Cassandra can 'go most of the way' along with the non-revolutionary relativist before encountering a difference. And this kind of specification crops up much more frequently than highly critical metaethical discussions in philosophy seminar rooms.[168] For example Cassandra can respond to questions which are problematic for fictionalists without issue – when asked 'Do you really believe that? Is it really wrong to do that?', she can happily and truthfully reply 'Yes, I really do believe it and yes, it really is wrong'. My point here is that in everyday terms, Cassandra and Helen actually do mean the same thing

[168] This point is similar to Olson's argument that fundamental moral disagreements are in fact quite prevalent and frequent (2010 §4.1).

to the extent that most people think about such matters, most of the time. And those everyday contexts include quite natural questions to which Cassandra can respond freely, yet fictionalists cannot.

It only becomes apparent that Cassandra and Helen are talking past one another once we ask whether the concept of wrongness they are each using is an objective concept or a relative one. Yet we learn from experimental philosophy that talking past one another on this metaethical level is likely a commonplace phenomenon which most people do not even realise is happening, and which poses no threat to the intelligibility of moral discourse in everyday contexts.

Geoffrey Goodwin and John Darley conducted a series of experiments to determine whether 'lay individuals' (2008 p. 1339) are intuitive metaethical objectivists or relativists. Their results show that most subjects were actually objectivists about some of their moral judgements, and relativists about others.[169] This may not be the metaethical bombshell it appears, because of course this finding is entirely compatible with the truth of both objectivist and relativist metaethical theories, since the subjects may simply be mistaken in either direction.[170] But it must be noted that this finding does not cause a metaethical crisis by rendering traditional moral discourse unintelligible at various unpredictable times. Rather, it shows that in most contexts, people can get on with moral discourse just fine without knowing whether their interlocutors are relativists or objectivists, or even knowing which they are themselves.

[169] Note that although Goodwin and Darley drawn their distinction between objectivism and subjectivism, which is usually understood as an agent-centric form or relativism, the way they frame subjectivism is also compatible with group-focused forms of hermeneutic relativism.

[170] Lest it be worried that I am in violation of the RC constraint here (see §5.2.2), note that this is perfectly compatible with error theorists' commitments that relativism is false as a hermeneutic theory. I am not claiming that most people are *right* to be hermeneutic relativists about some moral matters.

In highly metaethically critical contexts, the differences between B_{RRev} and B_{FOLK} may be more important than the similarities I have discussed here. In such contexts it would often be appropriate to further clarify utterances of 'torture is wrong' with further statements such as 'by which I mean to say…' even without being prompted to do so by others. But the vast majority of moral discourse takes place in less metaethically critical contexts than the philosophy seminar room. Therefore I suggest that when Cassandra and outsiders use sentences such as 'torture is wrong' in everyday contexts, the similarities are more important – without any further clarification being required, both parties understand one another to be making assertions, to be expressing sincerely held beliefs, and that those beliefs include the same practical import. I submit that these similarities are sufficient to support the claim that in most contexts the parties would not be talking past one another to an extent which derails my proposal.

This account is unavailable to revolutionary fictionalists for the reasons I discussed in §4.4.3. When uttering the sentence 'torture is wrong', fictionalists in principle do not make an assertion, and they do not express a belief at all. Where my proposal offers sufficient similarity with traditional morality to at least make the above argument against the dishonesty objection, revolutionary fictionalists have no such response. Furthermore, should any confusion arise, Cassandra can happily discuss with outsiders every aspect of what she means by moral sentences. It may be quite time consuming, but doing so does not imperil her metaethical position in any way. In contrast, revolutionary fictionalists must guard against spending too much time discussing their metaethical position in case this threatens the immersion on which revolutionary fictionalism relies.

Finally, on the points I raised in my dishonesty objection, if we read Joyce as recommending his proposal to everyone, everywhere, I objected that the huge educational resources required to implement this could undermine any prudential benefit which becoming

revolutionary fictionalists could provide. To an extent this problem is again shared by revolutionary relativism. But the problem is not as daunting for revolutionary relativism as it is for fictionalism. As I mentioned above, there is convincing evidence from experimental philosophers that most people already have a basic understanding of moral relativism, and may already have some relativist moral beliefs (see e.g. Goodwin & Darley 2008). That being the case, my proposal may even begin to look somewhat more modest – I am simply suggesting that we make all of our relevant beliefs work in a way that some of our beliefs already do.

Admittedly revolutionary relativism is a distinctive and novel kind of relativism, and no one yet has beliefs which conform entirely with the view I laid out in chapter 5. Thus the inculcation costs of revolutionary relativism cannot be ignored. But contrast revolutionary relativism with revolutionary fictionalism on this point - it is not believable that most people already have an understanding of any variety of moral fictionalism, assertoricity and so on. Yet the proposal is that *all* of our moral beliefs should be replaced with some kind of mental state which is thus far quite unlike any moral beliefs we currently have. Thus the task of educating everyone, everywhere about revolutionary relativism would plausibly be significantly simpler than teaching everyone, everywhere about fictionalism.[171]

Overall, then, while revolutionary fictionalism and revolutionary relativism both face issues around potential dishonesty and cost of implementation, these issues are much less problematic for revolutionary relativism than they are for revolutionary fictionalism. On balance thus far, this indicates that revolutionary relativism is the better choice.

[171] I will discuss this issue in more detail in §7.4. It is worth noting that the other WNP responses would also require significant educational resources for the same reasons. Even conservationists, despite their advice to retain existing beliefs, would still have to explain error theory, doxastic voluntarism, and so on. However, since I reject the other WNP responses for independent reasons, further consideration of the various comparisons in terms of educational resource costs is superfluous here.

On the 'which morals?' issue, I argued in §4.4.4 that the most pressing problem for fictionalists is that, even if fictionalists can successfully separate moral and non-moral rules, it is unclear which moral rules we should adopt as fictions, and why.[172] Alternatively put, even if we were to adopt fictionalism and hence equip ourselves with a means to resist temptation, it would remain a pressing question which temptations we should wish to strengthen ourselves against giving in to. Joyce advocates adopting 'the most useful' (2017,p. 83) moral rules as fictions. But I argued that we will be unable to settle on the desired outcomes of adopting fictionalism – i.e. what the moral fictions we might adopt should be useful *for.* For example, recalling §6.2, should we prefer a world in which people do not steal, or one in which people have a radically different view of property and do not recognise stealing as such? If the fictionalist stipulates, as they perhaps might, that the moral fiction(s) we immerse ourselves in should be useful *for achieving the best prudential outcomes*, then which of these worlds is prudentially best? Joyce gives us no way of answering these questions (see §4.4.4 for further discussion). And even if we can agree which outcomes to pursue, I objected that we will then find it impossible to determine which specific moral fictions we would need to adopt in order to promote the agreed-upon outcomes.

Revolutionary relativism is preferable to fictionalism here because it takes the 'which morals?' question out of the metaethicist's hands entirely, and places it in the hands of each moral* community. Thus my proposal removes the burden of answering the question at the theoretical level. Simply put, Joyce needs to answer the 'which morals?' question, but I do not. The reader may raise an eyebrow at this, but the reason why this is the case becomes

[172] In §4.4.4 I also raised a concern about how fictionalists could specify what counts as a moral vs. non-moral fiction. I noted in §4.4.4 that one potential way around this for fictionalists may be to say that moral fictions are whichever relevant fictions we feel it is appropriate to hold relevant reactive attitudes about, and so on – roughly, they are whichever fictions we respond to *as if they were moral*. I still believe the point I raised in §4.4.4 demands a response, and that fictionalists cannot rely on the kind of response I mentioned unless they provide an argument to that effect. But here I will grant for the sake of argument that something like this move is available to fictionalists and set this matter aside, simply for the sake of streamlining the discussion and concentrating on the most pressing problems at hand.

clear if we consider what to do immediately after deciding to adopt fictionalism versus revolutionary relativism. If we follow Joyce's advice, then after we see that error theory is true, we need to establish the prudentially optimal moral fiction to entertain. For the reasons I just mentioned (and for further reasons which I will discuss in the next section), this would likely be extremely difficult, if not impossible. Whereas if we follow the revolutionary relativist's advice, deciding which moral* rules to accept is a matter for the collective judgement of each moral* community. On revolutionary relativism, there do not need to be any specific moral* rules picked out at the theoretical level because each moral* community is free to arrive at what it considers the most prudentially beneficial rules to accept for itself. Admittedly this merely passes the buck from metaethicists to moral* communities. But if in doing so it avoids a theoretical burden upon revolutionary relativism which fictionalism cannot avoid bearing, then this surely counts in favour of revolutionary relativism at the metaethical level.

Taking the above arguments together, I believe that revolutionary relativism is a better response to the WNP than revolutionary fictionalism is in moralgood. I will now turn to showing why the same is true if we are currently in moralbad as well.

6.3.2. Revolutionary relativism vs. revolutionary fictionalism in moralbad

We saw in §6.2.2 that conservationism was ill-equipped to deal with the possibility that we are currently in moralbad. If traditional morality is harmful, then the conservationist's advice that we should retain our pre-error-theory moral beliefs even after accepting a moral error theory necessarily involves perpetuating that harm. But fictionalists may be able to do better here, because fictionalism may allow a degree of choice about which of our previous moral beliefs we adopt as fictions. Despite implying in earlier work that we should 'entertain'

roughly our pre-error theory moral beliefs (2001 pp. 229-230), Joyce has also argued that we should entertain whichever moral beliefs are 'the most useful' (2001 p. 185, 2017 p. 83).

While he does not discuss the possibility which I have called moralbad at any length, we may reasonably presume that at least part of what Joyce has in mind when he uses the term useful is that the beliefs in question should not be prudentially harmful. Thus it is open to fictionalists to argue that, even if we are in moralbad, only a subset of our previous moral beliefs are actually harmful, and that there are other moral beliefs which are not. Then, fictionalists might suggest that we could at least in principle determine which moral beliefs lead to problematic outcomes and which do not, and then adopt only the latter - the *right* moral beliefs, as Joyce puts it (2001 p. 185) - as fictionalist moral make-beliefs.[173]

Unfortunately, I do not believe that fictionalists can make this argument successfully. The flaw in the argument is that it is implausible that fictionalists could ever have sufficient information to correctly determine 'the right moral beliefs' *in advance* of adopting them as make-beliefs. And they would need to do so in advance, as I will explain. A majority of those who have published on the matter disagree that moralbad obtains as things currently stand – the endeavour of calculating now which specific moral beliefs might result in demonstrable harm or benefit in the future is highly problematic. Remember that although certain moral beliefs may intuitively seem uncontroversially beneficial under normal circumstances, we are for the time being setting aside those intuitions and granting that we may be in moralbad. Therefore, if we are to adopt as make-beliefs whichever moral beliefs can deliver prudential

[173] In case it should be thought that I am putting a poor argument into fictionalists' mouths, given that I immediately go on to disagree with it, I must point out two things. One, the strategy I suggest here is consistent with what Joyce says in response to abolitionists about selecting the 'right' beliefs. Two, I will go on to make a similar (albeit in my view more successful) argument on behalf of revolutionary relativism below, and again at somewhat greater length in §6.4.2.

benefits while avoiding harmful outcomes, surely this must be a cautious process in which make-beliefs should be subject to revision.

Yet the flexibility required for this revisionary process is incompatible with fictionalism. To see why, recall section 4.4 and my discussion of immersion. A central plank of the fictionalist argument is that only by turning a habit of making judgements and speaking as if our moral beliefs were (capable of being) true into a 'life strategy' can we secure the behaviour-regulating benefits of traditional morality. Without essentially forgetting that we are fictionalists in all but our most critical contexts, we will fail to reap the claimed rewards in terms of beneficial regulation of our behaviour. It seems clear to me that any process of revising our decisions about which moral beliefs we entertain, according to the positive or negative outcomes they turn out to promote, must break our immersion in the fictional morality.

This means that fictionalism presents us with a window in time between accepting a moral error theory and immersing ourselves in a particular (set of) moral fiction(s). Once this window has passed, revising the moral fiction threatens the mechanism via which fictionalists claim their proposal can deliver prudential benefits. Repeated revisions, which the careful process of tailoring our moral make-beliefs to avoid harm would surely demand, must render the required immersion untenable.

By comparison, fictionalism's failure to cope with moralbad demonstrates why revolutionary relativism can succeed here: revolutionary relativism allows a degree of flexibility and offers scope for the revision of moral* rules according to the benefits or harms which eventuate from acting in accordance with them. On my proposal, the moral* status of actions (i.e. whether actions are right or wrong) is not permanently or objectively fixed. This is what

makes my proposal relativistic – the moral* status of actions is relative to the attitudes of the agents in the moral* community in question.

What this means here is that, if there is an appropriate shift in the moral* rules which participants in a moral* community accept, then actions which were previously deemed morally* permissible will on my proposal become morally* impermissible, or vice versa. Thus revolutionary relativism can offer flexibility where fictionalism cannot. This flexibility does not by itself mean that revolutionary relativism can cope adequately with moralbad, however. What is required is that moral* rules can be flexible *in response to the prudential issues raised by moralbad*. And this is precisely what revolutionary relativism delivers, as I explained in §6.2.2.

Therefore I conclude that, if we are in moralbad, fictionalism's inflexibility is its downfall, whereas revolutionary relativism's flexibility is its strength – revolutionary relativism is a better choice for error theorists if we are in moralbad. Taking this together with §6.3.1, I submit that revolutionary relativism is a better response to the 'what now?' problem than fictionalism is, regardless of whether moralgood or moralbad currently obtains.

6.4. Revolutionary relativism versus abolitionism

In §4.3 I discussed and rejected abolitionism, a response to the WNP defended most prominently by Garner today (e.g. 2007, 2019), and previously by Hinckfuss (1987). Abolitionists argue that current morality (i.e. the view of morality most people have before potentially becoming moral error theorists) is prudentially harmful. I will recap some of their reasons for claiming this below, but the upshot is that abolitionists advocate jettisoning moral thought and discourse, and that accepting an error theory gives us the ideal opportunity to do so.

While I discussed problems with more recent defences of abolitionism, my principal reason for rejecting abolitionism was that it rests on a claim which cannot plausibly be proven – we cannot be sure that current morality is not of net benefit, and save for an impractically large scale experiment, we never will be sure.[174] However as I argued in §6.1, I believe that abolitionism must be taken seriously as a challenge to all WNP responses. It is insufficient to dismiss the possibility that we are currently in moralbad. I will therefore explain why I believe that revolutionary relativism is a better choice of response to the WNP than abolitionism in moralbad (§6.4.2) as well as in moralgood (§6.4.1).

6.4.1. Revolutionary relativism vs. abolitionism in moralgood

My argument that revolutionary relativism is preferable to abolitionism here could not be simpler. Abolitionists advocate abolishing all moral thought and discourse. If moral thought and discourse were prudentially beneficial, at least on balance, then doing as abolitionists suggest would necessarily mean we lost out on any and all of the benefits involved. This is something of a mirror image of §6.2.2 above – where conservationism depended entirely on the claim that traditional morality is of net benefit, abolitionism depends on traditional morality's being harmful. Thus if my arguments in the last chapter that revolutionary relativism can deliver prudential benefits were correct to any extent at all, then revolutionary relativism is preferable to abolitionism in moralgood to the same extent.

However, the other WNP responses assume that moralgood obtains, which means that the meat of their arguments, and therefore of my responses to those arguments, share that assumption. Abolitionism is different, since it proceeds from the claim that we are currently

[174] The other problems with abolitionism I discussed (concerning the impossibility of implementing Marks' variant, and the confused nature of Blackford's) I will set aside here. I have already discussed implementation in this chapter, and will return to this theme more fully in chapter 7. And my proposal is (hopefully!) not fraught with confusion in the way Blackford's is. Thus it is more fruitful to concentrate here on the provability point.

in moralbad. So in the case of abolitionism, the majority of the philosophically interesting material comes into play when we consider how revolutionary relativism and abolitionism compare in moralbad, as I do in §6.4.2.

6.4.2. Revolutionary relativism vs. abolitionism in moralbad

According to abolitionists, unlike other WNP respondents, we are currently in moralbad. Abolitionists hold that current moral beliefs actually cause us to harm one another in ways which we could avoid if we jettisoned moral thought and discourse. As I discussed in §4.3, the *locus classicus* of this view is Hinckfuss' *The Moral Society* (1987). Hinckfuss argues that morality systematically fosters and entrenches elitism, for example, because moral psychology and education include a kind of brainwashing of the less powerful majority in society to make them believe that society should be structured in a way which actually harms their interests and benefits only the powerful minority (1987, sections 2 and 3).

Another of Hinckfuss' accusations is that because there is a plurality of apparently credible, mutually exclusive moral views, and because moral judgements purport to capture objective moral facts, morality impedes conflict resolution in the direst situations, and even paves the road to war (1987 section 4). Further examples include those discussed by Garner (2007), and referred to in §6.2.2 above. Abolitionism is therefore claimed to be prudentially beneficial because it removes the mechanism by which these harms come about.

Perhaps somewhat surprisingly, I do not contest this. If we are currently in moralbad, then I agree that it is likely that we would be better off than we are now if we abolished traditional moral beliefs, very possibly along with their related thought and discourse. Rather, my argument here is that even though becoming abolitionists would be beneficial in moralbad, adopting revolutionary relativism would be *even more* beneficial. Revolutionary relativism

allows us to both avoid the harms of traditional morality, and, I will argue, to gain coordinating benefits which are unavailable to us on abolitionism.

The reason I make this claim is grounded in the same flexibility I discussed in §6.2.2. My argument begins by observing that it is not plausible that every individual moral rule is unavoidably harmful, and I do not take abolitionists to claim this. For example, it is highly plausible that a policy of acting in accordance with the judgement that *it is wrong to allow one's own children to starve to death when one does not have to do so* is prudentially beneficial under virtually all circumstances.[175] Rather, I take abolitionists to argue that current morality is harmful as a whole because certain aspects of morality lead to (often unintended or unforeseeable) harm.[176] This does not mean that there are no practically normative judgements which may be of prudential benefit. Rather abolitionists suggest that we would be better off getting rid of the moral part of our discourse and deliberations, and thinking instead of what we have the best prudential reasons to do in ways which do not invoke moral terms (Garner 2007 pp. 511-512).

However this is perfectly compatible, I suggest, with the claim that traditional morality is nonetheless potentially of *some* benefit. It is simply that those benefits are, for abolitionists, outweighed by the tendency to result in harm which is an inalienable part of traditional moral thought and discourse. Thus to help us formulate the abolitionist proposal without glossing over these potential benefits – insufficient though they may be to offset the harms of morality, according to abolitionists – we might observe something like the following: at the heart of

[175] It may, with sufficient effort, be possible to come up with circumstances in which it might, according to some moral realists, be right to allow one's children to starve – perhaps a powerful alien has ensured that unless they starve, they will die in an even more awful way. But any such circumstances will be so exceptional that I set aside this possibility here.

[176] I say unintended and unforeseeable because I consider it compatible with moralbad that e.g. agents think they're acting in a prudentially beneficial way when they oppose theft, even if in fact their attitudes to property entrench prudentially undesirable elitism.

the abolitionists' proposal is an implicit claim that non-moral prudential reasons and motivations can provide a sufficient proportion of the benefits of traditional morality, without incurring the associated harms, that our lives will go best on balance if we abolish morality, and rely entirely on non-moral prudential reasons and motivations.

What I want to argue is that revolutionary relativism can similarly avoid the harms of traditional morality, but can also offer some of the benefits which even abolitionists must acknowledge are plausible. We saw in §6.3 and in previous chapters that Joyce highlights the ability of moral beliefs to help us overcome akrasia. To this benefit I would add that traditional morality almost certainly *can* help to facilitate coordination between members of a society (even if, as abolitionists claim, it often fails to *actually* do so).[177] Simply put, a society seems likely to function more smoothly if its members share a similar framework of actions to be avoided, responsibilities towards others, expectations of justified censure for transgression and so on. Traditional morality plays a role in this, as do legal systems or professional codes of practise, for example.

I am not contesting the abolitionist's obvious rejoinder that traditional morality may indeed foster coordination *in service of a prudentially harmful belief system*. What I am claiming is that revolutionary relativism not only offers a way to avoid the harms of traditional morality, it also offers similar coordination benefits, benefits which are not available if we jettison all vestiges of morality and rely only on non-moral prudential reasons as the abolitionists would have us do. If, like abolitionism, revolutionary relativism can avoid the harms of morality, then the contest between the two proposals in moralbad results in a draw. But if revolutionary relativism can also offer benefits which abolitionism cannot (or at least cannot to the same extent), then revolutionary relativism becomes the better choice for error theorists.

[177] See e.g. Joyce 2006, DeScioli & Kurzban 2009 & 2013.

On the first point, the avoidance of the harms of traditional morality, I appeal to the arguments I gave in §6.2.2. There, I showed that revolutionary relativism allows for flexibility in the face of discoveries about the harmfulness of accepted moral* standards and moral* beliefs. This in turn allows post-error-theory moral* thought and discourse to evolve towards a harm*less*, purely beneficial situation.

I will concede that this means that any harms caused or inherited by adopting revolutionary relativist beliefs will not be removed overnight. The evolution of revolutionary relativism towards a system of beliefs which causes no discernible harm is a process, and depends on at least two things. One, any harms must be discovered before they can be eliminated. This cannot be guaranteed, at least in the short term – if it could, there would have been no reason for Hinckfuss to write his 1987 book, and it seems likely that it would be self-evident that non-abolitionist philosophers would be forced to take abolitionism much more seriously, rather than this being a matter for arguments such as mine in §6.1. Two, benign beliefs would have to 'win out' over harmful beliefs, and be accepted by moral* communities. It is conceivable that this may not occur – perhaps the gifted orators in a community could unanimously and unknowingly defend harmful norms, and thus persuade the other members to accept norms which are actually harmful.

I do not think that this is problematic for revolutionary relativism, however. For one, neither can abolitionists claim that the adoption of their proposal would bring an immediate end to morality's harms. For all that Garner claims that it would be easy to adopt abolitionism (2007 pp. 511-512), moral prejudices are deeply ingrained in society.[178] Even if nobody ever judged anything morally wrong again, it is unlikely that our non-moral attitudes would change quickly or without resistance. For example, even if Hinckfuss is right that traditional moral beliefs

[178] Let us not forget that Garner's fellow abolitionist Joel Marks admits that he has found it nearly impossible to suppress his 'moralist reactions' despite years of trying to do so (2019 p. 101).

about property are harmful, I do not expect that abolitionists would change their non-moral attitudes towards property ownership at a stroke. A sense that others should not take our possessions without our permission is simply too embedded in most cultures to be so swiftly overturned. Thus the fact that it might take some time for revolutionary relativist communities to evolve their beliefs away from those which might cause harm does not count against my proposal in comparison with abolitionism in any significant way.

Turning to the requirements for the discovery of harm and the acceptance of the 'right' norms, my response is that I believe we have cause to be optimistic that revolutionary relativist communities could meet these requirements. This is because my proposal makes it explicit that the moral* norms a community accepts should always be plausibly beneficial rather than harmful. This in turn requires that communities collectively think very carefully about the norms they accept. It therefore seems natural that the nature of my proposal would lead to an increased awareness that moral* beliefs could potentially be harmful, and so metaethically aware members of moral* communities would be more open to and more vigilant for unintended harms caused by moral* beliefs. This means that moral* communities would be much more likely to identify any harmful beliefs than was the case with traditional moral beliefs. Having done so, and having made appropriate revisions to the accepted norms in light of any harms discovered, it seems reasonable to think that the resultant beliefs would be harmless – at the very least equally as harmless as the non-moral beliefs we would expect abolitionists to hold.

On the second point, the coordinating benefits of morality, a way of drawing out the issue is to consider the so-called free rider problem.[179] To describe this idea simply, consider a

[179] For an overview of the free rider problem (also known as the *n*-prisoner's dilemma), see e.g. Shafer-Landau 2010, pp. 190-194 & Russell 2013.

communal good such as a rail service.[180] The rail service provides us with a good which we desire - transport facilities. The continued provision of the service depends on most travellers paying for a ticket and thereby contributing towards fuels costs, maintenance etc.. Yet it is often possible to get away with not purchasing a ticket, and 'free riding'. While, collectively, we have an interest in people not free riding in this way (and thus endangering the continued availability of transport facilities), individually we each have a prudential reason to exploit the service whilst trying to get away with not buying a ticket. This is because the prudentially optimal outcome for each individual is to be able to travel whilst keeping the money they could have spent on a ticket. Yet if everyone acted according to their individual prudential reasons, the good would cease to be available.

So, how do we balance the desirable continued availability of the communal good with the apparent prudential reasons for each individual traveller? One way to do so is by using the law to impose punishments on those who try to ride for free. This goes at least some way towards bringing the individual's prudential reasons into alignment with the collective's reasons, by giving individuals a prudential reason to avoid punishment, and therefore pay for their ticket. And abolitionists can consistently go along with this. But it may still not be enough to dissuade free riders in sufficient numbers to safeguard the rail service's future. After all, policing such laws can be expensive, and lots of people could still successfully ride without buying a ticket.

Traditional morality is arguably beneficial here. If individuals judge that free riding is morally wrong, then by providing a mechanism for the censure of wrongdoers, some degree of motivation to do the right thing, and so on, it is easy to believe that morality helps individuals to overcome the temptation to free ride, and facilitates a prudentially beneficial coordination

[180] Rail services are just one example one might use. Others could include clean air, water supplies, the welfare state, and so on.

in ticket buying behaviour. In the WNP context, of course, moral reasons for action are off the table. But as I argued in the previous chapter, revolutionary relativism can go some way to reclaiming these coordinating benefits by explicitly linking moral* judgements to motivation, censure and so on. I do not necessarily wish to claim that revolutionary relativist moral* judgements are motivating in exactly the same way as traditional moral judgements.[181] But to any extent that they are motivationally efficacious at all, then to that same extent, revolutionary relativist moral* judgements offer coordinating benefits which are unavailable to abolitionists.[182]

Lastly in this section, I will draw attention to a point which I did not raise in chapter 4 as a problem for abolitionists, but which nonetheless counts in favour of revolutionary relativism over abolitionism. Garner is confident that abolitionism would be quite simple to adopt, claiming that,

> For *amoralists*, that is, error theorists who have already come to believe in the falsity of all moral judgments, cutting back on moral pronouncements will be no more difficult than cutting back on swearing, and not nearly as difficult as getting rid of an accent. (2007 p. 511-512, emphasis original)

But this is far from an uncontested claim. Nolan *et al.* believe that abolishing moral thought and discourse is likely to be much harder than Garner claims (2005 p. 307), and Olson agrees

[181] I will go into somewhat more detail about motivation shortly in §6.5.1.

[182] Hinckfuss or Garner could perhaps respond to me that one of the key harms of traditional morality is that it leads to the escalation of conflicts (see e.g. Garner 207 p. 502). And on my proposal, people could still have moral* disagreements and therefore conflicts could escalate. I do not think this is problematic for me, however – I take it that the reason why traditional morality leads exacerbates conflicts (according to some abolitionists) is that both sides in conflicts believe they have access to some kind of objective truth. The relativistic nature of revolutionary relativism should help avoid this (since both sides would have to agree that there is no objective truth to be had). And even if it does not, escalating conflicts are clearly not prudentially beneficial, and so the mechanisms I have already described would lead to alterations in the moral* norms people accept such that the conflicts would not escalate (potentially including resolving not to have any moral* rules about the relevant issues and instead treat them as abolitionists would).

(2014 p. 181). As I mentioned previously, one of Garner's fellow abolitionists admits that he has found it is exceedingly difficult to suppress his 'moralist reactions' (Marks 2019 p. 101). And Strawson (1962) and Streumer (2013a) both go so far as to doubt that eliminating moral thought and discourse is even psychologically possible. While I do not wish to digress into exploring the limits of moral psychology, suffice it to say that for error theorists who sympathise to any extent with Garner's opponents here, revolutionary relativism should be more attractive than abolitionism. This is because, while I do not propose that we retain moral thought and discourse unaltered, revolutionary relativism does preserve much of the vocabulary and deliberative methods of traditional morality. Therefore it would plausibly be much more psychologically straightforward to adopt revolutionary relativism than abolitionism, and therefore the preferable choice, all other things being equal.

Taking sections 6.4.1 and 6.4.2 together, I conclude that revolutionary relativism is a better response to the WNP than abolitionism, regardless of whether moralgood or moralbad currently obtains. This is because, as I have shown, abolitionism fails to be of benefit if we are in moralgood, and if we are in moralbad, revolutionary relativism can deliver benefits which abolitionism cannot. Thus revolutionary relativism is a better choice for error theorists going forwards than abolitionism is. In the next section, the last of the sections in which I compare revolutionary relativism with its competitors, I will discuss revolutionary expressivism.

6.5. Revolutionary relativism versus revolutionary expressivism

In §4.5 I discussed and rejected revolutionary expressivism (hereafter frequently abbreviated to RE), principally as proposed by Svoboda (2015).[183] Svoboda proposes that, if we accept

[183] Note that Svoboda prefers the name *revisionary* expressivism – see footnote 109 above. Svoboda's is not the only revolutionary expressivist view I touched upon – a related proposal by Köhler and Ridge was set aside because a) it seems to be at odds with moral error theory and WNP responses actually defended to date, and b) because despite the authors' arguments to the contrary, the proposal was

error theory, we nonetheless have good prudential reasons to retain apparently moral discourse and some kind of moral thought. For example he claims that moral discourse aids in the resolution of disagreements, and that this benefit can be preserved, whilst avoiding error, if we adopt revolutionary expressivism after we become error theorists. Similarly, Svoboda claims that traditional morality 'bolsters one's commitment to act for certain ends, increases one's self-control, and helps overcome weakness of will' (2015 p. 20), and that his proposal preserves this benefit.

Specifically, RE consists in using moral language in the post-error theory context to express desire-like attitudes, as opposed to beliefs. Thus, for example, where we previously used a sincere utterance of 'stealing is morally wrong' to express a belief that there is a moral fact about stealing, Svoboda recommends that as error theorists we should henceforth use the same utterance to instead express a desire (or a similarly conative attitude such as a hope, expectation, demand, preference or plan) that people do not steal.

I will compare revolutionary relativism and revolutionary expressivism in moralgood and moralbad, and show that my proposal wins out in each situation. In each case I will summarise the problems I raised with RE in §4.5, and show why relativism either does not face the problem in question, or fares better than expressivism does in the face of the problem.[184] This will be the last of the 'head-to-head' comparisons I will make in this chapter, and I will then

found to be plausibly self-defeating (albeit not necessarily conclusively so). See §4.5, especially footnote 108.

[184] I will assume here that revolutionary expressivists can either solve or avoid the Frege-Geach problem. As I touched upon in §4.5, aspects of revolutionary expressivism may be somewhat provisional until a solution is found. But since this is a very thorny philosophical matter which does not confront my own proposal, and since I take the arguments I gave in §4.5 to undermine revolutionary expressivism without grasping any of the relevant thorns, I set the matter aside. Note that revolutionary fictionalism may also face something like the Frege-Geach problem, too - Oddie & Demetriou argue (2007) that fictionalists of various kinds face a relative of the Frege-Geach problem, which they call the acceptance-transfer problem. For the same reasons, however, I omit further discussion of this issue as well.

go on to draw together the conclusions reached in this chapter and in the broader scheme of my thesis as a whole.

6.5.1. Revolutionary relativism vs. revolutionary expressivism in moralgood

It is not necessarily obvious that Svoboda assumes that we are currently in moralgood. This is because he sidesteps the question by characterising abolitionism as a view primarily motivated by epistemic worries about having false beliefs (2015 p. 9). Since RE does not rely on agents having moral beliefs, the suggested epistemological abolitionist worry is supposedly defused without appearing to take a position on whether abolitionists are right that traditional morality is prudentially harmful. I do not believe this characterisation of abolitionism is adequate (see footnote 151, above). But setting that aside, I believe that we may reasonably infer that Svoboda assumes we are currently in moralgood since he recommends RE on the basis of preserving benefits of traditional morality rather than avoiding its harms (2015 §5).

In §4.5 I argued that revolutionary expressivism is incapable of delivering the benefits Svoboda claims in the intrapersonal (i.e. motivational) and interpersonal (i.e. coordinative) contexts he identifies. Therefore my argument here will consist in taking these contexts in turn, and in each case providing a recap of my previous argument, followed by an argument that revolutionary relativism is a better response to the WNP in that context.

i) Intrapersonal context

In the intrapersonal context, Svoboda argues that traditional morality serves as a bulwark against giving in to temptation (2015 §5.2). This is plausible – according to the view of traditional morality typically taken by error theorists, moral norms essentially 'tell us' not to do immoral things, even if we feel tempted to do them. Thus if we judge that an action is

wrong, this judgement produces (or is otherwise intimately associated with – Svoboda uses the term 'provide' (2015 p. 20)) a motivation to resist giving in to any temptations we may feel to perform that action. Svoboda claims that RE can preserve this benefit because on RE moral judgements, as desire-like attitudes, are inherently motivational. I objected that revolutionary expressivism cannot deliver this benefit, at least not to the extent that Svoboda claims.

To summarise my argument, consider an agent, *A*, who makes a moral judgement, *M*. On revolutionary expressivism, we infer from this that A is motivated to act in accordance with *M*, since moral judgements, as desire-like attitudes, are inherently motivational. Now consider what happens when *A* experiences a temptation, *T*, to act in contravention of *M*. How does *A*'s judgement that *M* affect their response to *T*? I argued that a revolutionary expressivist understanding of moral judgement can explain what *A*'s motivations are at the point at which they judge that *M*. But this says little of use about *A*'s motivations when they then experience *T*.

This is because a reasonable way of interpreting *T* is as some kind of motivation to act. So an expressivist reading of *A* experiencing *T* merely posits two contradictory motivations (*M* and *T*). It offers no reason to think that the motivation implied by *M* remains compelling for *A* indefinitely, or that a prior motivation predominates a later one, or that moral judgements have any greater degree of motivational efficacy than any other kind of motivation. Bluntly, desire-like attitudes may be motivation*al*, i.e. they have to do with motivation. But there is no clear picture of whether or how they are motivat*ing* over time or in competition with conflicting motivations.[185]

[185] My argument here contrasts with a related problem raised by others against hermeneutic expressivists, and which may apply to Svoboda's proposal as well – that expressivism is flawed because it cannot allow for weakness of will at all. See e.g. Smith 1998. I will not pursue this here, however,

I do not wish to claim that adopting RE cannot furnish us with any way to resist weakness of will at all. If I disapprove of some action in a 'relatively fixed' (Blackburn 1998a, p.68) way, but find myself tempted to perform that action, then the motivational component of my existing disapproval of the action in question may make me less likely to succumb to temptation. This will be especially likely in a deliberative process where I consider my motivations, recall my disapproval, and so on. I take it that something roughly like this is how Svoboda expects instances of weakness of will to work out on RE. As I wrote in §4.5.2, I believe that this falls short of the account of weakness of will available on traditional morality. For example, one could argue that if I am a moral realist and believe that an action remains wrong entirely independently of anything to do with my approvals, disapprovals, temptations or any other conative attitudes I have, this will have a greater impact on my ability to resist weakness of will than the account available on RE allows. But even granting for the time being that adopting RE helps us resist weakness of will equally as well as we did as traditional moralists before we accepted error theory, I still believe that revolutionary relativism can do better yet.[186]

This is because nothing in Svoboda's proposal involves agents coming to have any new motivations (I use 'motivations' here as a shorthand for 'desire-like-attitudes which have an inherent motivational component or effect'). Svoboda proposes that 'moral judgments should be transformed into desire-like attitudes, and moral utterances into expressions of such attitudes' (2015 p. 15). But he does not recommend that we make different first-order moral judgements, or that we should come to have a whole raft of new conative attitudes.

since I prefer to focus on showing how my own objections to Svoboda's proposal are sufficient to show that revolutionary relativism is preferable to revolutionary expressivism.

[186] I leave it to conservationists to take up the argument whether or not morality and RE are equal here. This is because conservationists would presumably want to argue that retaining our realist moral beliefs as they propose would mean we also retain exactly the same account and degree of anti-akratic benefit as is available on traditional morality. They would therefore potentially have a greater interest than I do in arguing that the preserved benefits were greater than the similar yet diminished benefits available on RE.

He proposes only that we should replace our traditional moral judgements (which must, if they can help overcome weakness of will, be linked with motivations) with RE-style moral judgements which result in having the same practical motivations which we had as traditional moralists. So it is not part of Svoboda's theory that our motivations should change, just that we should use moral language to express them.

Revolutionary relativism, on the other hand, via the attitudes I labelled BELIEF and ACCEPTANCE, brings in extra factors which bear on agents' motivations, yet which are not necessarily part of moral deliberation on traditional morality as conceived of by error theorists, or on RE. That is to say, adopting my proposal and coming to have moral* beliefs makes us think about things and bear things in mind which are not part of the deliberative picture on competing WNP responses, or even, necessarily, on traditional morality. I will offer a couple of specific examples to highlight this.

First, the BELIEF component of my proposal includes as an integral part of moral* judgement that agents are participants in communities, participation in which is conditional upon accepting and acting in accordance with certain norms. This makes explicit the potential consequences of moral* transgression in terms of censure and ostracism in a way which is not necessarily the case for other proposals, or for traditional morality. These considerations must feed into almost any agent's deliberative process when facing temptation. Granted, it may be open to revolutionary expressivists to claim that community-focused DLAs *may* be involved in moral deliberation if we adopt their proposal, but they do not play any necessary role.

Likewise, moral* judgements necessarily involve the expectation of prudential benefits resulting from acting in accordance with the relevant norms. I argued in §4.5.2 that we cannot simply come to have DLAs as we choose, even if those DLAs are in favour of beneficial actions

or policies such as eating healthily. Yet moral* judgements necessarily involve the belief that acting in accordance with the relevant norms is prudentially beneficial. If doing so is not beneficial, my proposal includes provision for abandoning prudentially suboptimal moral* norms, regardless of the occurrent DLAs agents might have. Thus it is explicit at the outset that adopting my proposal may have consequences as regards our first-order moral/moral* beliefs, and that agents may have to reappraise their motivations.

This means that the explicit link between moral* beliefs and prudentially beneficial consequences of acting in accordance with those beliefs may well help moral* judgements function as a better bulwark against weakness of will than the equivalent judgements on RE. This is because when tempted to act contrary to a moral* judgement, revolutionary relativists will be aware that this will most likely be prudentially harmful in comparison with acting otherwise. On the other hand, as I argued in §4.5.2, it is unclear whether our DLAs naturally track prudential benefit. As such it is plausible that revolutionary relativist agents who deliberate in the face of a temptation to act immorally* will be motivated by prudential considerations which it is not clear will enter into the equivalent deliberative process on RE.

In these ways, adopting my proposal involves alterations in the factors which bear on our motivations. And those alterations arguably put agents who adopt my proposal in a better position to resist akrasia than agents who adopt RE. Even if all of this is wide of the mark, my proposal offers an account of akrasia which is very close to that available on traditional morality – beliefs about certain actions being right or wrong in a substantive sense, reactive attitudes towards agents who act contrary to those beliefs, links between judgements and practical motivations and so on. As such, my proposal is on at least an equal footing with RE as regards weakness of will even if we grant – as I do not believe we necessarily must – that RE is on a par with traditional morality on this issue. But as I have argued, I believe

revolutionary relativism can go further than this, and is actually of greater benefit in this respect than Svoboda's proposal.

ii) Interpersonal context

In the interpersonal context, Svoboda claims that RE 'preserves [traditional morality's] useful feature of accounting for normative and evaluative disagreement, because it can track these as attitudinal divergences' (2015 p. 22).[187] This feature of traditional morality is useful because it facilitates prudentially beneficial coordination among agents. When we have moral disagreements, if we have an account of what we disagree about, and what the disagreement consists in, then we are more likely to be able to solve our disagreement than we would be if we had no such account. Again, I objected that RE cannot deliver the benefits Svoboda claims. Or more accurately, I argued in §4.5.2 that RE can offer *an* account of moral disagreement, but it can do so only in a way which falls short of delivering the same benefits as traditional morality, and the benefits available on my proposal.

To summarise, I argued that the traditional model of moral disagreement is useful because it analyses moral disagreements as disagreements in beliefs about matters of fact. This is useful at least in part because in disagreements in beliefs about matters of fact, if the facts of the matter can be established, then in principle the disagreement can be resolved – any party whose beliefs do not match the facts must, according to standard accounts of beliefs, either change their beliefs or be irrational. And when disagreements are resolved, it seems likely

[187] Svoboda also discusses RE's ability to allow 'for a kind of moral reasoning among various parties' (2015 p. 21). I set this feature of his proposal aside since his argument in defence of this claim leads directly into the Frege-Geach problem. As I mentioned in §4.5, I take my own case against Svoboda's proposal to be sufficiently strong to support my claim that revolutionary relativism is the preferable response to the WNP without sailing into such stormy philosophical waters.

that the parties will be better able and more likely to coordinate their behaviour (or at least cease to argue and hinder one another).

This remains so even if it is metaethically controversial whether we can actually know what the moral facts are. Error theorists are perhaps likely to have some sympathy with a view of traditional morality according to which we cannot come to know moral facts – see e.g. Mackie 1977, p. 38-39 & Olson 2014 §5.2. But in typical cases of moral disagreement, so long as the speakers can satisfy *themselves* that they have agreed at least approximately on what the moral facts are and thereby resolve their dispute, then morality has played a useful coordinative role, regardless of whether or not any genuine moral facts have been established. The vast majority of moral discourse is between people who have little metaethical training, yet still competently use moral discourse in a way which is consistent with there being knowable moral facts, and which therefore seems to aid coordination.

My objection to Svoboda's proposal was that it accounts for moral disagreement as disagreement in attitude rather than as disagreement about matters of fact, and therefore cannot offer the same coordinative benefit as traditional morality. The upshot of this is that rather than retaining the relevant feature of traditional morality, revolutionary expressivism involves replacing the conventional model of moral disagreement with a different model, which it is less certain can yield the instrumental benefit of disagreement resolution which Svoboda claims it can. I did not therefore conclude that revolutionary expressivism yields *no* relevant kind of benefit, but that it is plausibly of less benefit than Svoboda advertises.

In contrast, on revolutionary relativism disagreement must be approached in two distinct ways. This is because we must distinguish disagreements between parties who are members of the same moral* community from disagreements between parties who are members of different moral* communities. When moral* disagreement occurs between members of the

same moral* community, revolutionary relativism offers an account of moral* disagreement which is substantially similar to traditional morality. This is because revolutionary relativists would view moral* judgements as beliefs about the normative standards the community accepts. And on revolutionary relativism, there would be an empirical, descriptive fact of the matter about which standards are accepted by the community as a whole.

While in real-world situations it may not be instantly and uncontroversially clear to all parties what the relevant facts are, it would in principle be possible to determine them.[188] This is much the same as the way in which traditional moral realists can disagree about the moral facts, but there is always in principle a matter of moral fact which can be determined. Indeed given the explicit manner in which revolutionary relativist moral* facts would be established, we might expect the moral* facts to be even clearer to disputants on revolutionary relativism than traditional moral facts are. If so, revolutionary relativism could actually improve upon the usefulness of traditional morality's ability to account for moral disagreement.

This picture is complicated somewhat by the availability on revolutionary relativism of a new locus of disagreement in comparison with traditional morality, namely disagreement about which moral* standards the community should accept. For (most) moral realists, there is no question about what the moral facts *should be*, merely about what the moral facts *are*. On my proposal, sincere moral* judgements implicate attitudes concerning which norms are plausibly most prudentially beneficial to accept, and therefore new questions are raised (see §5.3.4). I do not believe that this is a serious issue for revolutionary relativism, because this

[188] This may not be as simple as having a straw poll, but it can in principle be done. An analogy might be drawn to determining national attitudes towards contentious political issues – here, too, it would be simplistic to suggest that direct plebiscitary democracy is the only way to decide what is in the nation's interests, yet government is nonetheless possible. It should also be noted that outside of certain religious communities, even the staunchest of traditional moral realists can seldom point to a single, conclusively authoritative source for moral facts, and so revolutionary relativism is not necessarily any more vague on this point than the traditional morality we are trying to retain beneficial features of.

is still an empirical matter which is in principle soluble, even if this might be tricky in practise. As I discussed in §6.2 & §6.3, revolutionary relativism offers sufficient flexibility that finding the optimal moral* standards can be a gradual, iterative empirical process.

Perhaps more significantly in the present context, however, the same issue confronts revolutionary expressivists, but may be harder for them to solve. This is because Svoboda offers no guidance as to which desire-like attitudes revolutionary expressivists *should* have. It may seem easy for expressivists to respond to this by suggesting that what agents desire and care most about (i.e. their DLAs) will naturally track prudentially beneficial outcomes. But this cannot be relied upon – there are myriad examples of people who are passionately in favour of measures and practices which are demonstrably against their prudential interests.[189] It seems to me that one of two things must follow from this. Either 1) RE is neutral on the DLAs agents should have, in which case revolutionary relativism is preferable to RE on prudential grounds because it pushes agents towards prudentially beneficial outcomes while RE does not. Or 2) the desire-like attitudes of revolutionary expressivists should somehow track prudentially beneficial outcomes, in which case RE is no better off than revolutionary relativism on this matter, yet owes us an account of desire-voluntarism (i.e. how we can choose to have certain desires and not others) which Svoboda does not provide.[190]

When moral* disagreements arise between members of different moral* communities, things become somewhat more complicated. But, again, I do not think this is problematic for

[189] Comparatively uncontroversial examples (i.e. without touching upon matters such as Brexit voting patterns in heavily EU-subsidised areas of the UK!) include anyone in favour of the availability of junk food, pro-smoking groups, 'anti-vaxxers' and those who engage in various forms of self-destructive behaviour involving substance abuse or self-harm.

[190] What is required here is some kind of account of why agents' desire-like attitudes *actually would* track prudential benefit. It may well be that agents' lives will plausibly turn out better if those agents orient their desires in one way rather than another, and thus agents have a prudential reason to adopt desire-like attitudes which track prudential benefits. But this does not mean that agents *will* desire those things. Revolutionary expressivists may or may not be able to provide such an account. But either way, such a potentially complicated argument would be better coming from revolutionary expressivists themselves than being assumed, possibly incorrectly, by others.

revolutionary relativism. As we saw in chapter 5, I propose that we interpret moral* judgements primarily as beliefs about the moral* norms accepted by our moral* community, but that we also interpret sincere moral* judgements as implicating other attitudes as well. This means that when we encounter others who are not members of the same moral* community as us, beliefs about what our moral* community accepts may be considered irrelevant. But we may still disagree with members of other moral* communities in virtue of holding conflicting attitudes among the secondary attitudes implicated by our moral* judgements. This means that even if the factual model of disagreement available on revolutionary relativism fails when speakers are from radically different moral* communities, the model of disagreement in attitude which is available on RE is also available on my proposal. Plus it may be possible over the course of the disagreement for the speakers to become part of the same moral* community, and so return to the factual model after all.[191]

I conclude that revolutionary relativism is preferable to revolutionary expressivism in moralgood both in the intrapersonal context and, where disagreements arise between members of the same moral* community, in the interpersonal context. In the interpersonal context where disagreements arise between members of different moral* communities, revolutionary relativism can draw on the same resources as RE, and may be able to improve upon them. Therefore, revolutionary relativism is the preferable WNP response overall in moralgood. I will now turn to the comparison between RE and revolutionary relativism in moralbad.

[191] I will have much more to say about disagreement in the next chapter, since it is a classic source of problems for hermeneutic forms of relativism, and so I will need to defend my proposal against such problems. Therefore, while I take the argument here to be clear, if the reader requires more detail on this matter, for the sake of avoiding repetition I beg their indulgence until §7.1.

6.5.2. Revolutionary relativism vs. revolutionary expressivism in moralbad

If moralbad obtains, then the moral beliefs most people currently have actually led them to act in prudentially harmful ways (see §6.1, above). Yet it seems reasonable to assume that if we became revolutionary expressivists, our first-order moral judgements would align with our prior moral beliefs, at least initially. The term 'revolutionary' as I am using it here relates to a *meta*ethical revolution, and Svoboda's proposal is purely about how we interpret moral thought and discourse - he makes no mention of a wholesale revolution or even any significant change in our first-order moral judgements.[192] Therefore if we are to avoid harm in moralbad, the question expressivists must answer is this: what scope does RE allow for our moral judgements to change such that they no longer lead to harm?

As was the case with revolutionary fictionalism in moralbad, my argument here turns on flexibility, though it is a somewhat different flexibility argument to the one which I deployed in §6.3.2. My argument here comes in the form of a dilemma related to what is sometimes called the moral attitude problem.[193] In the present context, the dilemma concerns the nature of the desire-like attitudes which constitute moral judgements on RE. Recalling Blackburn (1998b p. 196), are these desire-like attitudes akin to 'mere' preferences, and subject to change as we see fit? Or is there nothing 'mere' about the desire-like attitudes in question, are they somehow resistant to change on a whim – and if so, why and how? This creates a dilemma for revolutionary expressivists: the latter possibility, which seems preferable for expressivists in general, makes RE less capable of avoiding harm in moralbad than

[192] Svoboda does suggest (2015 p. 16-17) that revolutionary expressivists may wish or need to cut back on making apparently truth-predicating moral pronouncements, particularly to avoid misleading others. But at the moment, we are interested only in how to best understand what revolutionary expressivists actually *would* say, not what they may not wish to say.

[193] For an excellent explanation and investigation of the moral attitude problem, see Miller 2013. Miller discusses the problem at various points, beginning with §3.6, p. 39.

revolutionary relativism. And the former possibility makes RE a poor response to the WNP overall. Either way, revolutionary relativism is preferable to RE in moralbad.

I will formulate my argument as simply as possible, while noting that the philosophical terrain here is significantly complex, and that I will be glossing over important issues in order to put my case in the kind of terms that the present context requires. It is a commitment of typical error theorists that traditional moral thought and discourse (unsuccessfully) aim at capturing objective facts. As such, moral judgements, when we think them true, have a compelling degree of apparent authority. By contrast, one might worry that non-cognitive attitudes such as desires or preferences are somehow weaker. If the wrongness of torture is somehow part of the fabric of objective reality, then it seems to follow that there is a stringent requirement that agents do not torture. Whereas if an agent who says 'torture is wrong' is simply expressing a desire, then the implied requirement to refrain from torture may seem no more stringent than the requirement to refrain from serving sprouts to a dinner guest who dislikes the taste of sprouts.

Now, if the desire-like attitudes which constitute moral judgements on RE were on a par with our gustatory attitudes concerning sprouts, this would bode well for RE in terms of flexibility in moralbad. In order to avoid harm in moralbad, we could expect that agents would change their attitudes if it brought about an overall benefit in much the same way that the nutritional content of certain foods can motivate us to eat them even if we are not fond of the taste. Indeed it may be that once we became habituated to the change in attitudes, the new attitudes would become genuine preferences independent of any benefits they brought. So long as we care about our wellbeing in a sufficiently stable and powerful way, we could acquire the moral taste, as it were.

This theoretical benefit would, however, come at the cost of making RE a deeply flawed response to the WNP. For in order to be of any benefit in the post-error-theory context, it seems that moral judgements must imply a level of commitment greater than half-hearted gustatory preferences. Otherwise moral judgements would fail to respect the Seriousness Constraint I discussed in §5.1, and thus fail to offer the benefits Svoboda claims – even aside from any arguments I have offered on the matter, they would surely lack sufficient motivational import to bolster us against giving in to weakness of will. And while such judgements could be the basis of genuine disagreement in attitude, it seems unlikely that anyone would care very much whether such disagreements were ever solved.

This is why, despite little on this matter appearing in Svoboda's paper, expressivists more generally have been careful to explain that the non-cognitive attitudes which constitute moral judgements according to their theories are much more significant than 'simple or 'mere' desires or preferences.[194] Blackburn vividly describes non-cognitive attitudes as sometimes immensely powerful aspects of our characters and motivations, asking 'If your attitude to me is one of contempt and disdain, or if you desire me dead, am I supposed to console myself by reflecting that these are merely attitudes and desires?' (1998b p. 197). Interpreting moral judgements as attitudes with the same level of impact and urgency as the desire to stay alive or the hope to remain free rather than be sold into slavery puts those judgements once again at the centre of the normative stage. Understood in this way, moral judgements are not the same as beliefs about objective moral facts, but they are arguably no less compelling. As such,

[194] It must be observed that Svoboda does not invoke, as I shortly will, Tiberius' notion of resilience, Blackburn's arguments from 1998, or any other argument along similar lines. However, given Svoboda's claim that 'Different versions of [revolutionary] expressivism are available, insofar as there are different versions of [hermeneutic] expressivism' (2015 p. 15) it is reasonable in the present context to present a case which draws from well developed hermeneutic expressivist accounts on Svoboda's behalf, at least until such a time as Svoboda or another revolutionary expressivist comes forward with their own view.

moral judgements on this account do seem potentially capable of playing the roles Svoboda describes, even if I disagree that they can successfully do so.

However, the theoretical benefit of viewing the non-cognitive attitudes which comprise moral judgements on RE in such powerful terms threatens to come at the cost of flexibility. For few people would plausibly be willing or able to change such deeply committed attitudes in order to avoid anything less than obviously catastrophic harms. And the harms of moralbad, while potentially serious and certainly best avoided, likely fall short of this at least most of the time – even over centuries of traditional morality, history has not been exclusively and literally the story of *bellum omnium contra omnes*.

This is part of the reasoning behind a 'third way', Valerie Tiberius' notion of resilience. Tiberius argues (2012) that in order for our normative judgements, which she understands in expressivist terms, to be able to form a proper basis for our plans and for 'good decisions', those judgements must be 'stable in the light of criticism and new experiences'. But this does not mean that one's moral judgements should be so stable that they cannot be altered. Rather, Tiberius advocates a degree of open-mindedness towards normative judgements which 'is likely to occupy the mean between the vices of pig-headedness and dithering uncertainty' (2012 p. 19).

If we construct a form of RE which includes something like Tiberius' view, this may seem to allow the degree of flexibility RE needs to be able to avoid harm in moralbad, while offering sufficient stability in moral judgements that they are not 'mere' desires. Provided, that is, that we grant that most agents, most of the time would take the avoidance of moralbad's harms seriously enough to reconsider their moral attitudes. I think we should be willing to grant that they would, at least as a possibility. Yet I nonetheless believe that revolutionary relativism is

preferable to revolutionary expressivism in moralbad even if the latter is elaborated upon by including something like Tiberius' view.

The reason I make this claim is that even if we accept Tiberius' view, it seems that changes in moral judgements must be subject to considerable inertia. It is natural for an expressivist view to be this way, in order to avoid any suggestion of 'mereness', as noted above. As Tiberius herself puts it, '[o]pen-mindedness will, therefore, by constrained by other virtues such as proper conviction' (2012 p. 19). This means that moral judgements, even on Tiberius' view, will be resistant to change. Exactly how resistant to change normative judgements would be is subject to debate, but empirical examples of attitudes which persist among otherwise reasonable people even in the face of compelling contradictory evidence are not difficult to find.[195] It seems clear, therefore, that changes in agents' moral judgements would be slow, sometimes resistant to evidence, and require considerable persuasion. The expressivist might respond by saying that we do not persist in our desire to eat some sweetmeat if we find out that it is poisonous. But expressivists cannot have it both ways – if an analogy to gustatory desires gives insufficient weight to moral judgements as I discussed above, then moral judgements must be resistant to change commensurately with the considerable depth of commitment with which they are made.

In contrast, revolutionary relativism has a much more direct route to the re-evaluation of accepted norms, if it should be discovered that those norms cause harm, as I laid out in §6.2.2. As I conceded in that section, the members of a community of revolutionary relativists would likely not change the norms they accept overnight. There is a degree of what could be called inertia here as well. But in essence, revolutionary relativism tells us that if we find that the

[195] One obvious example would be climate change, where those who do not endorse the view that climate change is caused by human activity – and there are many who do not - are opposed virtually unanimously by climate change scientists (see e.g. Cook *et al.* 2016).

norms we accept cause harm, then we must reconsider, cease to accept those norms, and replace them with norms which do not cause harm. This is an imperative matter, built directly into the way my proposal works. Whereas the best revolutionary expressivists can tell us, even with significant charitable embroidery, is that some revolutionary expressivists might argue that there are good reasons why we should consider changing our attitudes, provided that we subjectively wish to avoid harm. Of the two proposals, therefore, both can respond to moralbad to some extent. But when it comes to responding to the realisation that moralbad obtains, and then avoiding the harms implied by such a situation, I believe it is quite clear that revolutionary relativism is by far the better placed.

Overall, then, revolutionary expressivism as presented to date has little to offer in moralbad, and requires significant supplementation from other (hermeneutic) expressivist views. And even when supplemented appropriately, RE still seems to fall short of being able to offer the same benefits as revolutionary relativism. Taken alongside the arguments in §6.5.1, I conclude that revolutionary relativism is preferable to revolutionary expressivism as a response to the 'what now?' problem. This concludes the last of my comparisons between revolutionary relativism and competing WNP responses, and I will now move on to the concluding section of this chapter.

6.6. Conclusion

In terms of the overall structure of my positive thesis, in chapter 4 I laid out the 'what now?' problem, and argued that it necessarily confronts all moral error theorists. I then went on to explain how philosophers have sought to respond to this problem to date, and in each case argued that the response given was inadequate. In chapter 5 I outlined my own new response to the 'what now?' problem, revolutionary relativism. In the current chapter, I presented a dominance argument to the effect that revolutionary relativism is preferable to all other

responses to the 'what now?' problem. To date, respondents to the WNP have given insufficient weight to the challenge represented by abolitionism: what should we do if traditional morality leads to harm? Any response to the WNP must include a consideration of how the view would cope with the realisation that traditional morality is indeed harmful. Taking this into account, I compared each competing proposal with revolutionary relativism. In each case, I showed why revolutionary relativism was the preferable response to the WNP. Taking all of the sections of this chapter together, I conclude that revolutionary relativism is the best response to the WNP offered to date because it offers error theorists the best chance to secure the best prudential outcomes in response to the challenges presented by the WNP. If we accept the truth of moral error theory, the way forward which is supported by the best prudential reasons is to adopt revolutionary relativism.

This concludes my positive case for revolutionary relativism. In the next chapter I will defend my proposal against two kinds of criticism. First, I will defend it against some of the most important objections which confront all relevant forms of moral relativism, regardless of whether they are hermeneutic or post-error-theory views. Second, I will defend my proposal against key criticisms which apply specifically to my own post-error-theory formulation of moral relativism. Following this, I will present the conclusion of my thesis as a whole.

Chapter 7. Problems and Counterarguments

In this chapter, I will discuss a number of potential problems my proposal faces, and argue that they cannot derail my response to the 'what now?' problem. In the first part of the chapter in sections 7.1 and 7.2, these will be problems which standardly confront most or all variants of metaethical relativism. Naturally, readers with any knowledge of metaethical relativism will be concerned to know whether my proposal can cope with these objections. And more general readers will likely also find these problems and my attempts to deal with them useful in understanding the material and, hopefully, interesting. In the second part of the chapter in sections 7.3 and 7.4, I will anticipate the most problematic objections I can foresee which opponents might raise against my proposal specifically, even if they may not typically be raised as objections to other forms of relativism.

My aim in this chapter is to defend my proposal against the principal and most likely criticisms which might be made. This task, of course, can never be entirely completed. There is not sufficient space in a thesis such as this to consider every possible extant objection, and further objections which have not yet been formulated may always crop up in the future. But I will defend my proposal against the most salient and strongest objections in the available literature, and against the strongest objections I anticipate that others could raise.

Of the near-universal objections to hermeneutic moral relativism, several stand out as particularly troublesome. The most important of these is disagreement – according to some opponents, moral relativism is deeply flawed because it makes genuine moral disagreement impossible. I discuss this objection in §7.1. Then in §7.2, I will discuss several more-or-less related issues around moral epistemology, involving moral infallibility, moral dissidence and arbitrariness.

In the second part of this chapter, I will turn to objections which confront revolutionary relativism specifically. In §7.3, I will focus on what I call the Rational Choice Challenge, which concerns which moral* code(s) it would be rational for revolutionary relativist communities to adopt. This objection recalls aspects of the 'which morals?' issue I raised in earlier chapters when discussing revolutionary fictionalism in §4.4.4 (and which I also touched upon in §4.5.3), but is formulated in a way which specifically challenges my proposal. And in §7.4 I will discuss a problem which I foreshadowed in §4.4.3 concerning implementation – even if revolutionary relativism is fine on a theoretical level, it could be argued that my proposal would be so difficult and costly to implement that it would be impractical ever to do so.

This will complete my thesis. This chapter will then be followed by a final concluding chapter, in which I will draw together the ground covered in this project as a whole.

7.1. Disagreement

Hermeneutic forms of metaethical relativism (e.g. Harman 1975, Wong 1984, Dreier 1990) have been thought for a century or more to face a potentially fatal objection (see e.g. Moore 1922 pp. 333-334, *cf.* Dreier 2009, Finlay 2017). This objection centres around disagreement. Consider the following conversation, which I will call DISAGREEMENT: [196]

> Vladimir: It is wrong to torture people.
>
> Estragon: No, it is not wrong to torture people.

A slightly stilted quality (for the sake of clarity) aside, we have the strong intuition that so long as they are being sincere, Estragon and Vladimir disagree. And given the presence of the word 'no', Estragon himself clearly *thinks* he is disagreeing. Yet it is objected that relativist views

[196] I will use 'wrong' here just as an example. The problem and the analysis which follows also apply, *mutatis mutandis*, to other moral terms such as good, permissible and so on.

seem unable to account for this disagreement. In order to show why, and how revolutionary relativism can avoid this objection, I will approach the issue in three stages. First, I will give a brief outline of what the objection consists in when aimed at hermeneutic forms of relativism. Second, I will describe the 'building blocks' of my argument. And finally I will draw these blocks together to show how my proposal can deal with disagreement.

7.1.1. What the disagreement problem is

Roughly speaking, traditional relativist views analyse the meaning of 'wrong' in moral claims as relativising the content of the claim to the moral code or system - i.e. the set of moral principles - accepted by either the speaker themselves (in the case of speaker relativism, a.k.a. indexical relativism or subjectivism) or by the speaker's moral community (in the case of 'speaker's group' relativism, hereafter simply *group relativism*).[197] Thus when Vladimir claims that torture is wrong, relativists argue that what he means is that torture is forbidden by the moral code he or his community accepts. Likewise, when Estragon replies that torture is not wrong, relativists would have it that what he means is that torture is not forbidden by the moral code *he* or *his* community accepts. In other words, according to relativists, despite their apparent disagreement, Vladimir and Estragon's apparently contradictory claims could both be true at the same time.

The problem is especially acute for forms of speaker relativism. On speaker relativism, speakers asserting apparently contradictory moral claims are each merely reporting something about themselves: Vladimir accepts one moral principle regarding torture, and Estragon accepts another. Unless there is some extent to which their claims are mutually

[197] I borrow the term 'speaker's group relativism' from Dreier (1990 p. 21), as it captures the required meaning without incurring complicating baggage which alternatives such as 'cultural relativism' might. Despite its appropriateness, however, the term is slightly awkward, and so I will abbreviate it as indicated in the text from here on.

contradictory, this is no more a case of disagreement than Vladimir claiming that his own car is red and Estragon replying that his own car is blue. Rather than disagreeing, the pair are simply talking past each other.

Group relativist views have an advantage over speaker relativism here, because if Vladimir and Estragon are members of the same moral community, then they are indeed disagreeing about a matter of natural fact – namely, whether their shared community's moral code forbids torture. So at least in these cases, where the speakers belong to the same moral community, group forms of relativism can accommodate disagreement in a way which speaker relativism cannot. This being the case, and because revolutionary relativism is a kind of group relativism (even though I am still discussing hermeneutic views at this stage), I will henceforth set speaker relativism aside.

This does not mean that group-focussed forms of relativism are off the hook, however. For if in the example Vladimir and Estragon are members of different moral communities, then their respective uses of the word wrong simply assert non-mutually-contradictory claims about their respective moral communities - once again they fail to disagree and instead talk past one another. One way of illustrating this is to consider the following, slightly altered version of the conversation:[198]

> Vladimir: It is wrong to torture people
>
> Estragon: When you say that, what you say is true. However it is not wrong to torture people.

The disagreement objection points out that on relativism, Estragon's response here is correct. Yet it is nonsensical.

[198] This is drawn from Björnsson & Finlay 2010, p. 19.

This matters for several reasons. Most conspicuously, we are left with a phenomenon, moral disagreement, which definitely does seem to be part of moral thought and discourse on the one hand, and on the other hand a hermeneutic theory of moral thought and discourse which cannot account for that phenomenon. More pressingly for my purposes, it is important to note that having some way to account for moral disagreement is almost certainly very useful. Where people disagree about matters as important as e.g. torture, it is highly plausible that having a way to understand what is happening when they disagree will substantially increase the chances of their reaching an understanding and possibly agreement. This means that accounting for moral disagreement will help WNP responses deliver coordination among agents – one of the key putative prudential benefits of traditional morality which I am seeking to retain with revolutionary relativism. Indeed, even if my proposal could not account for disagreement, but adopting the proposal nonetheless improved cooperation between agents anyway, this would be a benefit. Finally there is the fact that moral disagreement is very psychologically familiar. One of my aims is to make moral* thought and discourse similar in use (albeit not in theory) to their moral counterparts, and so make the adoption and use of my proposal less psychologically challenging. Thus I have a vested interest in showing that revolutionary relativism can account for disagreement in a satisfying and intuitive way.

7.1.2. Coping with disagreement: Groundwork

I will begin my argument that revolutionary relativism can account for disagreement by drawing attention to a number of familiar arguments and phenomena. In simple terms, I will present my building blocks, and then go on to use them to construct my argument. Before doing this I should point out that there are a number of ways in which others have sought to respond to the disagreement problem on behalf of relativism, especially in recent years.[199]

[199] See e.g. Dreier 2009, Björnsson & Finlay 2010, Finlay 2017, Khoo & Knobe 2018, Suikkanen 2019.

My own response is somewhat different because it makes use of features of the WNP context which may be unavailable to hermeneutic metaethicists. But this does not mean that I am taking a position for or against any of those other views.[200] Rather, I believe that the specific features of revolutionary relativism allow me to account for disagreement without for the most part getting involved in the debates active in the hermeneutic literature.

That being noted, I will begin to build my own argument. I will start with several observations and principles which may seem unrelated at first, but which will be brought together as we progress. The first of the 'ingredients' for my account of disagreement is the pair of Gricean maxims of quantity. In order to make what we say 'conversationally suitable', Grice urged that we should, and indeed do, respect two maxims of quantity (1975 pp. 45-46):

i) Make your contribution as informative as is required (for the purpose of the exchange).
ii) Do not make your contribution more informative than is required.

Grice's view is not necessarily universally accepted. But it is so hugely influential and near-enough ubiquitous in the relevant literature that I am content to treat these maxims as guiding principles here.

It is important to note that despite their prescriptive formulation, Grice's maxims are analytical, not prescriptive – they are not primarily instructions for how to make suitable contributions to conversations (though they may also be read as such). Rather, Grice analyses conversations as 'cooperative efforts' (1975 p. 45), and presents his maxims, parts of an

[200] There are three main varieties. The first two, 'disagreement in attitude', exemplified in Björnsson & Finlay 2010, and the 'metalinguistic' view, in e.g. Khoo & Knobe 2018, are well summarised and discussed in Finlay 2017. The third, the 'proposition cloud proposal' is very recent, and is set out in Suikkanen (2019).

overall 'cooperative principle', as norms which we have internalised and to which we all more-or-less adhere. This, Grice argues, facilitates the cooperation required for conversations to be intelligible and fruitful. That is to say, we *already do*, as a matter of course and even if we are not consciously aware of it, adhere to Grice's maxims both in formulating our contributions to conversations and in interpreting the contributions of others.

Accordingly, we can put together a story about what happens in conversations which include remarks which do not, at least initially, appear to satisfy Grice's norms. When an interlocutor utters a sentence, we will typically initially expect to take it at face value. But if that sentence appears to violate one or more Gricean maxims, and so seems an inappropriate contribution to the conversation, we will not immediately conclude that the speaker has violated the Gricean maxims. Rather, we will typically charitably assume that what our interlocutor said *would* be an appropriate contribution to the conversation if it were looked at differently, and so seek to reinterpret their remark appropriately. Thus if someone says something to us which appears to be insufficiently informative, given the context of the conversation, we will typically and automatically question whether the face value meaning of what the person said was really what they intended to convey, and seek further implied meaning in their remarks which would respect the Gricean maxim of quantity. For example (drawing on Grice 1975 p. 52), a job reference which mentions a candidate's punctuality but which provides no useful information about the candidate's suitability for the job in question will typically be taken to tacitly imply that the candidate is unsuitable for the role – hence the phrase 'damning with faint praise'. Only if no such intelligible conversational implicature is found do we typically conclude that the person's utterance really did violate Grice's maxim, i.e. that it was an inappropriate contribution to the conversation.

A final note on this ingredient of my argument is that this Gricean story does not have to be infallible. Perhaps sometimes we do simply jump to the conclusion that speakers are idiots or

nervous or otherwise failing to make suitable contributions to conversations by Grice's standards. We may not be as unfailingly charitable as the story I have described suggests. This is not a problem for me. First, if we leap to judgement like this, we stand to miss important parts of what speakers are trying to communicate, so there is pressure to resist the temptation to form snap judgements. Second, even if the course of events is not guaranteed to follow the above story every single time, the familiarity of the process described shows that we clearly do act this way often, smoothly, and even without noticing that we are doing so. And finally, remember that my argument in this section is in response to the objection that relativist theories are systematically *incapable* of accounting for disagreement at a conceptual level. So if I can successfully argue that my invocation of Grice helps revolutionary relativism account for moral* disagreement *at all* then I am already ahead of the objection. That Grice's arguments are in fact widely accepted and clearly concern near-ubiquitous features of discourse just puts me further ahead.

The second building block of my argument is the observation that the utterance of an apparently descriptive sentence can have both illocutionary and perlocutionary dimensions (i.e. such sentences can be used both to assert that something is the case *and* to seek to affect an audience).[201] To capture the sense of this which I will draw on here, I will use the term *expectation*. If a speaker utters an apparently descriptive sentence under the expectation that her audience will agree, this may admit of both a reading according to which she anticipates that some fact about her audience's cognitive attitudes obtains, and also of a reading according to which she encourages her audience to agree with her. Both aspects can be part of the communicative intention of the speaker. Far from being a theoretical *deus ex machina*, this is a commonplace and familiar phenomenon. For instance, consider a politician

[201] This point draws on Stevenson (1963 p. 23-24). The example Stevenson uses is that of a mother who, in saying to her children 'we all like to be neat', rather than (or as well as) stating a descriptive matter of fact, tries to encourage her children to be neat. My example here is somewhat different, but is nonetheless inspired by a generalised reading of Stevenson's point.

who says 'We are patriotic people'. When a politician utters this sentence under the expectation that her audience will agree, it does not seem controversial that she both asserts a proposition about herself and her audience *and* seeks to arouse patriotic sentiments in that audience.[202] More would need to be said about this notion of expectation to describe it fully (for example, whether the apparently exhortative/suggestive dimension is actually a prediction predicated on psychological suggestibility), but I think the meaning I have in mind is intuitively sufficiently graspable for present purposes.

The third building block of my argument is to highlight that individuals are seldom if ever members of only one community, and this includes moral communities.[203] This is significant for group forms of relativism because the truth values and/or contents of moral claims are fixed by what an agent's moral community accepts. Not all moral communities will accept the same moral rules, and it may even be that different moral communities in which a given agent participates accept mutually contradictory moral rules.[204]

It would be possible for me to simply stipulate that on revolutionary relativism, individuals may participate in only one moral* community. But this would be extremely *ad hoc*, and would not sit with the requirement that where possible, moral* thought and discourse on revolutionary relativism should map closely on to their traditional moral counterparts.

[202] At the time of writing, a particularly prominent example in recent political discourse in the UK is describing matters of policy (which by their nature are often highly contested and very much arguable, depending on one's political leanings) as 'the right thing to do', thereby seeking to convince the listener that the rightness of the action in question is beyond question and should be accepted as fact.

[203] This point, alongside the FGM example which follows, was first raised in this sort of context by Shafer Landau (2004 chapter 10).

[204] An example which has attracted some attention in the West in recent years is the phenomenon of FGM, in which victims' families or wider ethnic communities act in accordance with moral codes which permit FGM, and which are thus at odds with the broader society's moral beliefs. One example among many of both the extent of this practice and the opprobrium which is directed at its practitioners by many people outside of the groups among which it is practised can be found in Topping & Carson 2014.

Therefore I must allow that the same multiplicity of moral* communities may apply to agents on revolutionary relativism.

This multiplicity of communities cuts both ways – on one hand, it poses a problem for relativism because it may be unclear which community is salient when considering how we should interpret agents' moral claims.[205] On the other hand, the fluid multiplicity of communities in which a given individual may participate means that when individuals come together, while they may be members of very different communities in some respects, there is a high probability that there will also be some overlap in the communities they each participate in, and therefore at least one salient shared community.

It is also important to note how participation in a moral* community might be defined. Partly this must surely have to do with which communities agents themselves identify with. But this cannot be the only factor. Moral and therefore moral* communities are to some extent coercive. For example, if the censure or approval of the members of a moral community exerts a psychological influence on an agent, then even if that agent would prefer not to consider themselves a member of the community in question, they are nonetheless to some extent a participant in it. An agent cannot cease to be a participant in a community on a whim, and the exertion of psychological influence (in such a way that it aids cooperation and is thus of prudential benefit) on agents by communities is part of the aim of my proposal. At the same time, this should not be taken too far – if members of a community which an agent categorically rejects persistently belittle that agent, for example, then it seems wrong to consider the agent a participant in that community simply because this bullying behaviour has produced a psychological effect, for example by making the agent miserable. For present purposes, then, I will define participation in a community as follows: when an agent regularly

[205] This is discussed as a problem for hermeneutic relativism by Shafer Landau (2004 chapter 10) and Suikkanen (2019).

engages with a community, and when there is a relationship of mutual psychological influence between the agent and other members of the community, then that agent is a participant in that community. A sharper definition of participation may be preferable if there were more space available here to present it, but I believe this definition will suffice for present purposes.

The fourth and final building block of my argument is to recognise that we routinely use ellipsis – i.e. we use terms which have explicitly relativised meanings without issue, even when there is no mention of the relativising parameters. This is because we make an assumption that the relativity of the terms' meaning is understood by speaker and audience. For example, if someone asks you which side of the road it is legal to drive on, you will not typically preface your answer with information about the judicial system, what country you are in, which countries' laws prohibit driving on which side of the road, and so on. You will typically simply answer left or right, depending on where you are.[206] Partly this has to do with the Gricean maxims of quantity I mentioned a moment ago, and partly it is because the extra information about the context of relativisation is assumed. Competent users of legal terms are aware that such terms are frequently elliptical, and should an American and a Japanese person find themselves fruitlessly failing to disagree about which side of the road one must legally drive on, one of the ways of resolving the situation is to realise that part of the problem is the relative nature of the legal terms being used.

7.1.3. How revolutionary relativism mitigates the disagreement problem

I am now in a position to offer my solution to the disagreement problem, and to show how it solves the disagreement problem in a sensible way. As I observed above, for group forms of relativism, including revolutionary relativism, disagreement between members of the same

[206] This parallels what Harman calls 'differences in situation' (e.g. 1978 p. 143).

community is not problematic. Thus my argument here is primarily directed at two things. First, I aim to show how revolutionary relativism can accommodate disagreements between members of different communities. And second, I will show that we have grounds to question whether disagreements between members of different communities happen as often as it may appear.

My argument begins by acknowledging that when revolutionary relativists encounter others, and particularly when they begin a moral* discussion, they will make an assumption about whether the other speaker is a member of the same 'main' moral* community.[207] We typically make numerous assumptions about people we encounter, such as the assumption that people we encounter in our native country will probably speak the same language we do. Assumptions about the various other communities in which interlocutors may be participants are no different. We tend not to expect centenarians to know a lot about video games, and we will not typically assume that natives in traditional dress whom we encounter in unexplored rainforests share our views on online privacy.

My argument then proceeds from the following claim: when revolutionary relativist interlocutors begin a moral* discussion, it is plausible that they will be quite good at knowing whether the other person is a member of the same 'main' moral* community. Partly this will be based on geographical experience, just as assumptions about languages are. And partly this will be based on the things others say. If an agent's assumption about the community

[207] Recalling §5.3.1, by 'revolutionary relativist interlocutors', here I mean people who knowingly use apparently moral discourse consistently with my proposal (whether they are aware of my actual proposal or not) and have an intuitive or 'working' knowledge of what this involves. I do not mean only people who also understand the full metaethical details of error theory, hold PhDs in philosophy etc. (though such people are also included). This is similar to how typical error theorists might describe 'the folk' as moral realists (see chapter 2 of this thesis), despite moral realism being a potentially very sophisticated and nuanced view when considered from an advanced metaethical perspective. Likewise, we have decades' worth of exceedingly sophisticated philosophical literature devoted to analysing what typical language users mean when using all manner of terms – yet none of this proves that only very highly educated theorists are competent users of the relevant terms.

membership of their interlocutor is incorrect, then it is also plausible that this would quite quickly become evident in the course of a moral* discussion.

This leads to there being three situations in which apparent moral* disagreements such as to the example I labelled DISAGREEMENT above may take place:[208]

1. The interlocutors correctly assume, or it becomes apparent, that they are participants in the same 'main' moral* community.
2. The interlocutors correctly assume, or it becomes apparent, that they are from different 'main' communities, but there may be some overlap in the moral* norms their communities accept.
3. The interlocutors correctly assume, or it becomes apparent, they are from communities so radically different that there is no overlap in the moral* norms they accept.

In each of these three cases, a different chain of events will play out, marked by the building blocks I described in the previous subsection. Note that I am not suggesting that revolutionary relativists should try to force this by approaching moral* disagreements with a mental 'script'. Rather, I am arguing that, if we bear in mind and draw together the resources I described in chapter 5 and the building blocks I laid out in the previous subsection of this chapter, we will see that a certain chain of events will plausibly naturally play out. My strategy here is simply to point out the stages and allow us to realise this.[209]

[208] Remember that the example labelled DISAGREEMENT in §7.1 was as follows:

Vladimir: It is wrong to torture people.

Estragon: No, it is not wrong to torture people.

[209] Naturally sometimes people will be wrong about various stages of this whole story. In such cases, perhaps it will be the case that interlocutors think they disagree, but in fact they do not. The point here, however, is to show that that on my proposal, speakers *can* genuinely disagree, even if sometimes they fail to do so.

I will lay out the 'story' in each of the three situations, highlighting where the various building blocks each come into play. Beginning with situation 1, where the interlocutors are members of the same (denoted by the label S) 'main' community:

S1. The speakers correctly assume, or it becomes apparent, that they are participants in the same 'main' moral* community.
S2. As self-aware revolutionary relativists, they are aware of this.
S3. Revolutionary relativists can interpret the speakers' apparent disagreement as a genuine and successful disagreement about a matter of natural fact – namely, whether their shared community's moral* code forbids torture.
S4. Because they're self-aware revolutionary relativists, the speakers know that they successfully disagree.

As I noted in §7.1.1, step S3 is available to standard hermeneutic group relativist views, and revolutionary relativism is no different here in being able to use this step to account for disagreement among members of the same community. Moving to situation 2, where the interlocutors are members of different (denoted by the label D) 'main' communities, but there may be some overlap in the moral* norms they accept, the 'story' runs as follows:

D1. The speakers correctly assume, or it becomes apparent, that they are from different 'main' moral* communities.
D2. As self-aware revolutionary relativists, they are aware of this, i.e. that a standard reading of the moral* terms they employ would lead to them talking past one another.
D3. Gricean maxims come into play – it would not be appropriate to rely on the standard meaning of the moral* terms used if this would lead to the interlocutors knowingly

talking past one another, because this would make their utterances insufficiently informative in the context of the conversation.

D4. The speakers assume a non-standard interpretation of the moral* terms used: the speaker making a moral* claim that ϕing is wrong is claiming that ϕing is forbidden by a standard of some community both interlocutors belong to, albeit not the community each speaker would typically most readily identify as a participant in.

D5. Further debate in the conversation then shifts to seeking a shared community to aid cooperation.

D6. If such a community is identified, the disagreement proceeds as in the same community case above.

D7. If no such community can be identified, the apparent disagreement shifts to situation 3.

This case is slightly more nuanced than the S1-S4 story above. Again, nothing particularly unusual is going on here, but the reasoning involved can be usefully expanded. To understand D4, consider that in the example of DISAGREEMENT, despite knowing that a standard reading of the moral* terms involved would result in talking past one another, Vladimir and Estragon nonetheless used the term *wrong*. Therefore, following the discussion of Grice above, they should each interpret their respective uses of *wrong* in a non-standard way which allows their utterances to make sense as appropriate contributions to the conversation. But why should they conclude that the appropriate non-standard interpretation of 'torture is wrong' is that 'torture is forbidden by the norms of some community we both belong to'? Recall that part of the explicit purpose of moral* discourse on my proposal is to foster prudentially beneficial

cooperation.[210] I submit that given that this is the case, the interpretation suggested in D4 is the smoothest and most natural way to accommodate this Gricean aim.

If that is true, then it is natural that in attempting to aid cooperation by facilitating genuine disagreement, the disputants would seek to identify the implied shared moral* community. If they can do so, they genuinely disagree. On the other hand, if the speakers cannot identify a salient shared moral* community, the conversation shifts into the third category of disagreement, in which the interlocutors do not belong to any shared community at all.

So finally we move to situation 3, where the interlocutors know from the outset that they do not belong to any shared community (denoted by the label R for 'radically different'), or where step D5 above fails and no shared community can be established. Here the story, again based on something like the example of DISAGREEMENT, runs as follows:

R1. The speakers correctly assume, or it becomes apparent, that they are from different 'main' moral* communities, and that there are no salient 'non-main' moral* communities in which they both participate.

R2. As self-aware revolutionary relativists, they are aware of this, i.e. that a standard reading of the moral* terms they employ would lead to them talking past one another.

R3. Furthermore, they are aware that no non-standard reading of the moral terms used would rescue them from talking past one another.

[210] This point draws on chapter 5 of this thesis. The claim that there is an explicit reason for the existence of moral discourse is not available to hermeneutic metaethicists in the way the equivalent claim here is available to me. This is because it is at best arguable whether moral discourse exists to foster prudential benefits – after all, depending on what kind of hermeneutic metaethicist one is, it could well be that one remains a target of abolitionists, i.e. it may be the case that moral discourse is actively harmful. On my proposal, however, moral* discourse has an explicitly prudential purpose. In making this observation, I put myself at odds with Suikkanen's assumption that 'the *raison d'être* of moral discourse and thought is to help us to avoid […] escalating vicious conflicts' (2019 §4).

R4. The speakers grant that they are literally talking past one another, and there is no factual disagreement between them.

R5. Gricean maxims come into play – it would not be appropriately charitable to interpret the speakers' use of moral* terms in a way which would lead to the interlocutors knowingly talking past one another without there being a fruitful alternative reading, because this would make their utterances insufficiently informative in the context of the conversation.

R6. The speakers assume a non-standard interpretation of the moral* terms used, which does not rely on there being any relevant shared community: 'the other person is seeking to influence me by expressing disapproval of ϕing or trying to get me to become a member of a shared community'.

In this situation, the speakers are forced to grant that they are talking past one another on a factual level. But here Gricean maxims come into play once again – since the speakers know that there is no factual disagreement between them, yet nonetheless use moral* terms, they must seek to interpret the use of those moral* terms in a way which makes sense and allows the utterances in question to be suitable contributions to the conversation. Put simply, the speakers assume that even if there is no factual disagreement, they are still using moral* terms for a reason. And the question then becomes: what could that reason be?

This is where the final building block I referred to earlier enters the frame – the fact that apparently descriptive sentences can be used both as assertions and to seek to influence an audience. In this situation, interpreting the speakers' utterances as assertions of fact has failed to produce a fruitful conversation because the speakers are talking past one another on a factual level. Yet the exhortative aspect of the conversational meaning of a sentence such as 'torture is wrong' remains. Thus a speaker who sincerely utters such a sentence may thereby seek to express disapproval of torture, in spite of having been forced under the

circumstances to abandon any factual claims about torture.[211] If this is the case, then the use of moral* terms in this situation still serves a purpose. Making one's attitudes clear in the deeply serious manner facilitated by moral* discourse is surely of prudential benefit because it can aid coordination among agents, and may even lead to the establishment of new moral* communities (or the acceptance of new or updated norms among pre-existing communities).

Alternatively (or also), the speaker could be attempting to elicit agreement that torture is to be disapproved of, and thereby establish the wrongness of torture as part of the basis or context for the rest of the conversation – essentially seeking to make their interlocutor a member of the same community as themselves.[212] Again, this is surely of potential prudential benefit, since the establishment of moral* community with interlocutors allows genuine disagreement to take place once again, and thus contributes to the realisation of the benefits offered by traditional moral disagreement in the first place.

Note that while others may wish to analyse all moral disagreements in these ways (see Finlay 2017 p. 191 for lists of those who might do so), for me these exhortative interpretations are 'fall back' positions. I expect that the vast majority of moral* disagreements among revolutionary relativists would be explained by the steps I have labelled S1-4 or D1-7. Yet alongside these more common cases, revolutionary relativism also has the resources to

[211] Here I am drawing on what is sometimes called a 'quasi-expressivist' account (e.g. Finlay 2017 p. 191). This means an account of disagreement which is substantially similar to the 'disagreement in attitude' model Svoboda relies on (see the preceding chapter of this thesis), but which operates on a pragmatic rather than a semantic level (i.e. in terms of what the speaker is expressing, rather than what the words used mean).

[212] Here I draw on what is sometimes called a metalinguistic account of disagreement (see e.g. Plunkett & Sundell 2013, Khoo & Knobe 2018). Roughly, this means an account of disagreement whereby to make a normative assertion that *p* 'is to propose updating the context so that it is common ground that *p* is true in that context' (Khoo & Knobe 2018 p. 25). To interpret speakers as attempting to establish a moral* community with their audience is my gloss on updating the context in the relevant sense.

account fruitfully for moral* disagreement between speakers who are not members of any salient shared moral* community.

To sum up, revolutionary relativism can account for genuine disagreement between speakers who belong to the same moral* community in much the same way that group forms of hermeneutic relativism can. In cases of disagreement between members of different 'main' moral* communities, revolutionary relativism offers scope for reinterpreting speakers' utterances in a way which allows genuine disagreement to take place. And if no community can be established among the disputants, the revolutionary relativist use of moral* terms can still be of prudential benefit. This means that on my proposal, at no point does moral* discourse become nonsensical in the way the disagreement objection to relativism implies. Therefore revolutionary relativism can meet the objection from disagreement – in the majority of likely cases revolutionary relativists can interpret apparent moral* disagreements as genuine factual disagreements, and even where revolutionary relativists must accept that no factual disagreement is possible, there are still prudential reasons to use moral* terms in line with my proposal.

7.2. Moral epistemology & related concerns

I will now turn to look at several related issues which revolve around moral epistemology to varying extents, and which have been raised as objections against hermeneutic moral relativism. Shafer-Landau captures well what many people intuitively feel is wrong with hermeneutic forms of moral relativism:

> …societies are sometimes based on principles of slavery, of war-like aggression, or of sexual, religious or ethnic oppression. [Group] relativism would turn these founding ideals into iron-clad moral duties, making slavery,

> sexism, and racism the moral duty of all citizens of those societies. The iconoclast – the person deeply opposed to conventional wisdom – would, by definition, always be morally mistaken. This has struck many people as seriously implausible. (2010 p. 279) [213]

There are several worries intertwined here which I will attempt to untangle and discuss in turn, showing in each case that revolutionary relativism cannot be derailed by the worry in question.

7.2.1. Infallibility

The first issue here is infallibility. Many of us share an intuitive distrust of any claims to infallibility. We feel that people and communities must surely be capable of making moral mistakes, and we will be suspicious of any metaethical theory which results in agents or communities being morally infallible. Yet this is exactly what moral relativism seems to imply – if moral truth is defined by what a moral community accepts, then the community cannot be mistaken about what it is morally right and wrong to do. So relativism doesn't just leave open the possibility of moral infallibility (which would be enough by itself to make many people question relativist views), it actually implies the claim that the community is *necessarily* morally infallible. This fails to accord with how we seem to use moral thought and discourse, because to most people moral error appears to be a very real and ever-present possibility. And so this implicit infallibility undermines moral relativism as a hermeneutic view.

[213] Shafer-Landau uses the term *cultural* where I have used the term *group*. As I noted in footnote 197, I am concerned that the term cultural relativism may carry unwanted baggage for some readers, and so prefer the term group relativism. Therefore for consistency, in the quoted passage I have replaced Shafer-Landau's term with mine. Nothing turns on the distinction between the two terms here.

In fact this infallibility seems wrong simply in and of itself - to borrow Blackburn's phrase, it seems 'unpardonably smug' (1998a p. 318).[214]

In order to cope with this worry, then, I need to show either that on revolutionary relativism, the moral* community is fallible, or that smugness is actually pardonable after all. Fortunately, on my proposal, the community clearly can be mistaken. And in §6.2.2 I laid out a 'story' about how a revolutionary relativist moral* community can respond if mistakes are brought to light.

Now I should admit that there is a twist here – on my proposal, what the community may be mistaken about is whether or not adherence to a given norm is prudentially beneficial. This implies that the community cannot be mistaken about the moral* truth, since moral* truth is defined in terms of the norms the moral* community accepts, as I described in §5.3.2.

There are two ways to respond here. On one hand, I do not have to grant the implication that the moral* community cannot be mistaken. Even though I believe the majority of matters will be clear enough, the finer details of right and wrong can be complicated and tricky to establish. Determining right and wrong is often not about consulting a rulebook, but is more about having a degree of sensitivity and being disposed to act in certain ways.[215] This is certainly the case with traditional moral beliefs, and I expect revolutionary relativism to be no different. This means that agents could sometimes be mistaken about the moral* norms their community accepts, and so would not be morally* infallible.

[214] Blackburn is talking about quasi-realism, but the smugness sentiment extends to relativism, too. By way of highlighting the importance of this issue to quasi-realists as well as relativists, Andy Eagan actually goes so far as to label one of the premises of his argument about Blackburn's view 'NO SMUGNESS' (2007 p. 210).

[215] This point draws on Hooker (2000 pp. 88-92). For example, '…figuring out whether a rule applies can require not merely attention to detail, but also sensitivity, imagination, interpretation and judgement' (2000 p. 88). Hooker is discussing consequentialism, but the point applies here too.

But even if I forego this line of defence, remember that like other WNP responses, and unlike hermeneutic forms of relativism, the whole point of my proposal is to deliver prudential benefits. Notwithstanding any notions of truth, any prudentially suboptimal norms a moral* community accepts must be reassessed and rejected once they are found to be prudentially suboptimal. And when this happens, the moral* truth, defined by what norms the community accepts, changes. Thus, albeit by a slightly indirect route, moral* communities may be mistaken about which norms are prudentially optimal, and if they are, the moral* truth will come to reflect this. In the WNP context, where objective moral truth is ruled out and prudential benefit is all that is on offer, I believe this is an adequate response to the infallibility worry.

7.2.2. Dissidence

The second issue we can identify in the quote from Shafer-Landau is raised by the iconoclast, or as I prefer to call them, the dissident.[216] This worry is closely related to the infallibility worry, and is that according to group relativists, the moral truth is defined by the community, and so moral dissidents – i.e. those who disagree with the rest of their community - must necessarily always be wrong. This remains so even if those moral dissidents are motivated by what seem like good reasons rather than out of a desire to do morally bad things. Yet we would typically want to leave open the possibility that even if all around them disagree, individuals who e.g. oppose slavery can be right to do so. This is especially so in the case of revolutionary relativism, which depends for its vital flexibility on the possibility of individuals

[216] Again, a slight terminological change: Shafer-Landau uses 'iconoclast', but I will use 'dissident'. This is because I feel the latter term a) carries less religious baggage and b) better captures the motivation of an agent who sees things differently than her community does because she believes that there is a prudentially better position which could be accepted by the community. This contrasts with the term iconoclast, which to me suggests opposition to convention for the sake of opposition itself. Such an individual will frequently come out as morally wrong on most metaethical views in a philosophically uninteresting fashion which I do not believe is Shafer-Landau's intended target.

or groups within a moral* community being at least in principle able to influence which moral* rules their communities accept.

I would respond that for revolutionary relativists, moral* dissidents would indeed hold moral* beliefs which are false according to their moral* community, and which are thus false, full stop. But this does not prevent dissidents being able to advocate changes in the moral* rules which their community accepts, so long as those changes are plausibly prudentially beneficial. Accordingly, as noted above, I can grant that the community can be morally* infallible yet also mistaken about whether the practical policies which are accepted as moral* truths are actually prudentially optimal. And given the prudential grounding of moral* rules described in chapter 5, this suboptimality necessitates a change in the moral* norms the community accepts, and thus a change in which moral* beliefs are true. This means that moral* dissidents may in the first instance necessarily be morally* mistaken as Shafer-Landau suggests. But this need not be a problem for revolutionary relativism, since my proposal provides a mechanism whereby prudentially beneficial norms advocated by dissidents can be accepted by moral* communities, and prudentially beneficial dissident beliefs thereby rendered true. Thus the dissident is not 'just plain wrong'. Rather she is the linchpin of how moral* communities can evolve for the better.

7.2.3. 'That can't be right!'

The third issue raised in the quoted passage is arbitrariness - many people feel very strongly that certain actions simply must come out as morally wrong, no matter which metaethical theory we consider true. Therefore any metaethical theory which threatens to allow e.g. slavery or rape to be anything other than automatically wrong, no matter what any community might to say about it, will be viewed as suspiciously arbitrary or even dismissed

out of hand.[217] Yet, again, this is what relativism seems to imply, and so any relativist theory which does not automatically rule certain actions out as wrong must be a bad theory. Simply put, we are intuitively more sure that slavery and rape are morally wrong than we are that moral relativists can be right.

In the post-error-theory context of this thesis, this cannot stand; since we are granting the truth of error theory, there is nothing – there *can be* nothing - which is morally obligatory, permissible or forbidden. Thus there can be no moral grounds on which to object that a given revolutionary relativist community accepts the wrong moral* norms. Opponents of hermeneutic forms of relativism are free to attempt to make some other kind of normative case that moral codes should or must proscribe e.g. slavery or else be incorrect or intolerably arbitrary. But there seems little reason to expect them to succeed when categorically normative reasons are off the table as they are in the present post-error-theory context - the arbitrariness objection implied by Shafer-Landau's remarks simply cannot apply if a moral error theory is true. All that there is left to argue about in the present context is whether there might be prudential advantages in adopting certain practical policies – and that is precisely what revolutionary relativism deals in.

Additionally, while I will not commit to any constraints on the moral* norms revolutionary relativist communities *must* accept, there are grounds to expect that most communities probably *would* accept norms which forbid certain practices. For example, people who would accept norms which authorise letting babies starve to death would have to be very psychologically unusual in terms of whether they care about human suffering and so on.

[217] Compare this with Joyce's constraint that 'a theory of imperatives that managed to supply strong categorical imperatives – that located Foot's "fugitive thought" – but for things like "Kill anyone who annoys you," "Steal when you can," etc., simply would not be a *morality*.' (2001 p. 67, emphasis original). Or later in the same book, '…a theory of moral imperatives had better get them "in the ballpark" of the kinds of things we uncontroversially consider such.' (2001 p. 75).

Whether for evolutionary or any other kinds of reasons, our cares and concerns are generally rather more stable and robust than the objection implies. And so even for communities which are radically different from ours, it is unlikely that their circumstances would be *so* radically different that they would give rise to prudentially beneficial norms which egregiously violate the 'that can't be right!' sentiment.[218]

As I said at the beginning of this chapter, it would be impossible to deal with all of the potential objections to my view. But taking §7.1 and the subsections of §7.2 together, I believe I have shown that revolutionary relativism can cope with the most serious of the objections to moral relativism in general. Therefore I will now turn to objections which may not apply to all forms of moral relativism, but which apply specifically to my proposal as a WNP response.

7.3. The Rational Choice Challenge

As with any philosophical view, it is impossible to anticipate all of the potential objections to my proposal which could eventually be raised. But we can be sure that some objections will apply to my proposal which do not apply to relativist views more generally, either because of the specific WNP context in which my proposal is made, or because of specific features of my response to the WNP. Of those which I can foresee, two stand out as particularly problematic, and in §7.3 and §7.4 I will tackle them in turn. The first of these challenges I call the Rational Choice Challenge (hereafter RCC). Aspects of the RCC may apply not only to other WNP proposals, but also to issues in other debates which fall outside the scope of this thesis.[219] In

[218] I am aware of an existing isolated community which reportedly actually does sometimes leave mothers giving birth to babies to die, and even sanctions killing babies which appear healthy to outsiders (Everett 2008 chapter 6). Yet even these actions are arguably based on a sense that members of that community must be tough to survive their harsh environment, and that protecting weak members of the community would place a burden on the other members so great that the community itself would be threatened.

[219] In the WNP context in general, something very like the RCC presented here may confront other responses – as I note in the text, the RCC is similar to objections I made myself to revolutionary fictionalism and revolutionary expressivism. More widely, similar or related issues have concerned, for

the present context, the RCC may be understood as a challenge similar to the 'which morals?' challenge I levelled at other WNP responses (see §4.4.4, and the related argument in §4.5.2). But the RCC is a particular issue for revolutionary relativism for two reasons. First, while it may be possible to level related challenges at other WNP responses, the formulation of the RCC I will give here is specific to my proposal because it makes use of moral* codes and community acceptance. And second, the RCC can be thought to apply not only at the time of the WNP itself (as with would be the case with the variations on the theme of the RCC which might apply to other WNP responses), but also at key points in the process of flexibility, a process I discussed a lot in the previous chapter and on which the arguments I have made in favour of my proposal rely quite heavily. As far as I am aware, the RCC has not yet been discussed as a challenge to WNP responses. Yet since its targets include such a crucial aspect of my proposal, if it can be made into a substantive challenge to my proposal – as I believe it can – the RCC demands a response.

To properly understand and formulate the RCC, we must take stock of where we are on the journey from accepting the error theory to forming a community whose members live consistently with my proposal. One way to do this is to envisage three points in time. First, we have the point at which we accept the truth of error theory, which we can label T^1. In one sense, T^1 is the 'revolution', the point at which we break with traditional morality and move into new territory. I am assuming that we have taken this step, and are now at a subsequent point in time, T^2. At T^2, we realise that the revolution which occurred at T^1 means that we are faced with a problem about what to do next – the What Now? Problem. At our current point in time T^2, then, we are looking towards the future, towards various situations which could potentially obtain at T^3, and considering which potential future we have the best prudential

example, philosophers working on theories of rational choice and justice, and others outside philosophy such as economists. See e.g. Cudd 1996.

reasons to seek to bring about. Naturally I have argued that the best prudential reasons support adopting revolutionary relativism at T^3.[220] But this is where the RCC comes into the frame: even if we accept that at T^3 we should, prudentially speaking, live consistently with my proposal, we can still ask, 'which potential revolutionary relativist community would it be rational for us to be?'. Putting this more carefully, we can formulate the RCC thus:

> RCC: From the perspective of the WNP at T^2, and looking to the future at T^3, is there a specific moral* code, or are there specific features of multiple potential moral* codes, which it would be rational for us to adopt?[221]

A key feature of the RCC which is implicit here is that the RCC concerns the *content* of the moral* beliefs which make up the moral* code accepted by a community. This is distinct from the arguments for adopting revolutionary relativism which I have presented thus far, which largely concern the meaning of moral versus moral* terms, or the truth values of moral* beliefs, but say comparatively little about the content of moral* codes. So, for example, if we believed prior to T^1 that murder was morally wrong, then adopting my proposal as defended thus far involves a change in the meaning of the term 'wrong': from a meaning which presupposes the existence of a certain type of categorical normativity to a meaning which relates to communities and the norms they accept. But adopting my proposal as defended thus far does not require that there is any change in whether or not it is murder which we

[220] Recalling §5.3.1, by 'adopting revolutionary relativism', I mean roughly living in accordance with my proposal, rather than necessarily critically engaging with every last metaethical detail. While the more people who adopt my proposal understand of it the better, I am not saying that everyone should e.g. have a PhD in philosophy at T^3. The same goes, *mutatis mutandis*, for phrases such as 'revolutionary relativist community' or 'moral* community' in what follows.

[221] This is not the only such 'rationality' question we could ask at this juncture, but it is the most interesting one in the context of my thesis thus far. Other questions include 1) at T^2, is it rational to move to a T^3 in which we live consistently with revolutionary relativism? And 2) for those who already live consistently with revolutionary relativism, is it rational for them to accept their community's rules? I take it that 1) is sufficiently similar to whether we prudentially ought to adopt my proposal to be answered by my arguments in previous chapters. And 2) would, in the current context, be a huge digression into general metaethical arguments rather than challenges to my proposal, and is thus outside the scope of this chapter. Therefore I set these questions aside here.

judge to be wrong, whatever that judgement might consist in. My proposal is about whether and how the judgement that murder is wrong can be true at T^3. The RCC, on the other hand, directly concerns the content of moral* beliefs. It asks which moral* beliefs - i.e. beliefs with what content - it would be rational for us to have. The RCC is therefore an *extra* challenge over and above the WNP, and would seem to require that I argue for the adoption of revolutionary relativism *plus* certain changes to the *content* of the moral* code which is accepted by the community.

Perhaps the most natural suggestion for how to respond to the RCC is to argue that we should accept those moral* rules which maximise prudential wellbeing in an aggregate, overall sense, similar to the calculus underlying rule utilitarianism (see e.g. Hooker 2000, especially chapter 2).[222] A significant degree of the prudential benefits plausibly brought about by traditional morality, and which my proposal aims to deliver, have to do with coordination and cooperation among agents. And my proposal is a group form of relativism in which communities *en masse* accept moral* rules. This would seem to support an intuition that my proposal rules out accepting any moral* rules which might erode aggregate prudential benefit, regardless of the effects at the level of particular individuals. All-in-all, then, thinking about the prudential benefits offered by accepting one rule or another at the community-wide aggregate level seems at least a reasonable place to start.

Yet this suggestion may be more problematic than it appears. Many people have argued that following principles which seek to bring about the greatest benefits at the overall, aggregate

[222] This is, for example, roughly what I read Olson as advocating when he emphasises the societal and coordinative aspects of his proposed retention of morality: 'human beings need morality to coexist peacefully, to prevent conflicts, to regulate and coordinate behaviour' (2014 p. 197). Similarly, Joyce recommends fictionalism to groups because of its usefulness *to groups*, rather than to individuals (2017 pp. 82-83).

level can lead to vastly unequal distributions of those benefits, which in itself is problematic.[223] For example it may be that 90% of the community decide to accept moral* rules such that it is morally* right, or at least permissible, that they enslave the remaining 10%. The 90% then experience a considerable uplift in their wellbeing due to the reduced need for them to perform manual labour and so on, while the enslaved 10% of the community have a terrible time. Yet because a large proportion of the community have nice, easy lives, and only a small proportion are forced into drudgery, the aggregate level of prudential wellbeing in the community is increased.

This is a simplification, of course, but the point is clear – maximising the aggregate level of prudential benefit in a community could be prudentially bad for some individuals. And so it is not clear that those individuals would have prudential reasons to accept the proposed moral* rule. This gives rise to the issue which is at the heart of the RCC, and which is the essence of the challenge to my proposal – when considering whether there is a revolutionary relativist community such that it would be rational for *us* to establish or join it, are we talking about a proposal which must, or would, be rational for everyone to adopt jointly and severally, or only for most people to adopt?

In order to respond to the RCC, in §7.3.1 I will begin by considering some ways in which there could be a moral* code such that all agents as individuals would be rational to adopt it. I will consider four responses to the RCC along these lines, and reject them all. After this, I will argue that we must reject the assumption that a good response to the WNP requires that the moral* code of a community must be such that all members of the community as individuals would be rational to adopt it. I will then provide my own solution to the RCC.

[223] For a much more thorough discussion of this well-known problem than can be accommodated here, see Hooker 2000 §2.8.

7.3.1. Rational for all agents?

Let's recap where we are in the latter stages of the overall story of this thesis. As we confront the RCC, we are currently at the point in time I labelled T^2 – we are in the aftermath of 'the revolution', and considering whether there could be a moral* code such that all agents would be rational to accept it at T^3. I am suggesting that one way to answer this is to consider how we might come up with such a code. I can see four options here, each of which I will explain, and then briefly give reasons why they are problematic. The four options are: seeking pareto optimality, egalitarianism, sufficientarianism and participating in rational bargaining. Having outlined and rejected each of these, I will then explain why revolutionary relativists must approach the RCC differently, and show how this can be done.

i) Pareto optimality

To find a moral* code such that it would be rational for all agents to accept it, first, we might suggest that a satisfying WNP response must require that so-called pareto improvements to the moral* code are made wherever possible (most likely from a starting point of the moral* code being a moral* equivalent of putative traditional moral 'truths'). Named after economist Vilifredo Pareto, a pareto improvement is defined as an improvement which can be made to a situation whereby one or more agents benefit, and no agents lose out. A situation is pareto optimal when no pareto improvements can be made. This applies most obviously to the distribution of material wealth or goods, but it can also apply to moral* rules. For example, accepting a rule against stealing plausibly conduces to fewer people having their possessions stolen, but it does not obviously result in anyone losing any of their possessions.[224] It would

[224] Some people might disagree with this theft example, especially some kinds of communist and those who I have described as abolitionists. Or thieves, for that matter. But I take it that the theft example illustrates the point sufficiently intuitively to satisfy more neutral readers for the time being. Since I will shortly reject the pareto optimality option for other reasons anyway, an intuitive illustration such as stealing will suffice for present needs, even if it may not survive sophisticated scrutiny.

therefore be a pareto improvement to accept a rule against stealing if one was not already accepted.

Since pareto improvements by definition mean that some people gain while no-one loses anything, it seems that it would be (instrumentally) rational for all agents to adopt any measures (including e.g. accepting moral* norms or changes to the moral* code of their community) which conduce to pareto improvements and thus contribute to bringing about a pareto optimal situation. And since we are dealing with prudential normativity in the WNP context, it seems unarguable that it would be rational for all agents to agree to implement any available pareto improvements to the current situation. This manoeuvre would appear to prevent the introduction of slavery, for example, as slavery would represent a change to the moral* code which caused some agents to lose out significantly. Also, seeking pareto optimality would satisfy the intuitions about aggregate *dis*-benefits I mentioned above, since the pareto improvements made *en route* to pareto optimality necessarily involve no loss of wellbeing or other prudential benefit for any agent. In the search for a situation which all agents would be rational to act so as to bring about, then, this seems a promising start – we can establish that any such situation would have to be pareto optimal, otherwise it would not be rational for all agents to seek to bring it about.

But the pareto optimality response is nonetheless problematic. This is because all manner of situations could be pareto optimal, yet many of them would fail to respect my proposal's requirement to adopt whichever moral* code is prudentially optimal, rather than pareto optimal (see chapter 5). To take an extreme example for the sake of illustrating the point, imagine a situation in which a catastrophically narcissistic or psychopathic 'master' somehow enslaves and is served by the whole of the rest of their community. Despite obvious prudential shortcomings for the vast majority of agents, this would be a pareto optimal state, since no relevant changes could be made - for example, by introducing a moral* code which

forbids slavery – without the 'master' losing out, even if only to a very minor extent. And pareto improvements by definition cannot result in anyone losing out at all. Yet it is a situation which almost no agents would be rational to seek to bring about – the only agents who might be rational to do so are the prospective master, and perhaps some similarly psychologically unusual agents who crave servitude.

Setting aside such extreme examples, even if we start with the world as it currently is, there are still issues of what we might call 'wellbeing distribution' which could not be solved by making pareto improvements. This is part of the motivation for abolitionism, as we saw in chapter 4. The upshot is that there is no one unique pareto optimal situation. And even if there were, pareto optimality is no guarantee that it would necessarily be rational for all agents seeking prudential benefits to bring that situation about.[225] Ultimately, then, despite a promising start, the pareto improvement model is unsatisfying as a guide in finding a moral* code which it would be rational for all agents to accept. On reflection, it turns out that conduciveness toward pareto optimality may be a necessary feature of the moral* code we are looking for, but it is not a sufficient one.

ii) Egalitarianism

The second way we might come up with a moral* code such that all agents would be rational to accept it at T^3 is by adopting a norm of egalitarianism. Thus at T^2 we envisage various potential T^3 states, $T^{3\text{-}a}$, $T^{3\text{-}b}$…. $T^{3\text{-}n}$ and seek to realise one in which agents are equal in the senses we consider important. Egalitarianism can take many forms, but for example we might focus on potential T^3 states where the moral* code conduces to equal distribution of wealth and liberty, perhaps by including norms outlawing ownership of other agents, norms

[225] The 'maximum aggregate' solution I mentioned in §7.3 is also pareto optimific, yet was also rejected.

promoting individual freedoms and norms relating to redistribution of wealth.[226] This would avoid the apparent gross injustice of permitting slavery, and would comport well with many people's intuitive sense of moral fairness. That being so, perhaps egalitarianism can guide us towards a T^3 which all agents would be rational to bring about.

Unfortunately, egalitarianism would also involve 'levelling down' to the extent that some agents might end up only slightly better off or even in some cases worse off than they would without morality*, and so it would threaten their prudential motivation for adopting revolutionary relativism.[227] Certainly, there may be net benefit in sacrificing some personal advantage in order to live in a more generally equal community. Yet it is unlikely that *all* agents would experience sufficient net benefits of this kind that it would be rational for *all* agents to bring about an egalitarian T^3. For example, it seems unlikely that many individuals who are very wealthy at T^1 would consider it rational to promote the acceptance of norms at T^3 which would result in most of their wealth being redistributed equally to others.[228] Thus it is unlikely that egalitarianism could be the basis for a moral* code which it would be rational for *all* agents to adopt.

iii) Sufficientarianism

Third, we could insist on principles of sufficiency to guide moral* rule acceptance.[229] This approach targets a hybrid between individual and aggregate prudential benefits, in order to

[226] This is of course only a sketch of egalitarianism, which is a sophisticated and vigorously debated view in its own right. See Arneson 2013 for an overview.
[227] For a classic treatment of this kind of reasoning, see Parfit 1997.
[228] Though I do not necessarily aim to endorse it here, a further argument could be made that pareto improvements are still possible even in a state of equality, along the lines of Rawls' 'difference principle' - social and economic inequalities may actually be prudentially advantageous, so long as everyone ends up better off as a result of them (see e.g. Rawls 1999 chapter 2, §12). Thus, again, it would be prudentially irrational for all agents to promote equality.
[229] Again, this is merely a sketch of a much more sophisticated view. See e.g. Crisp 2003 for a fuller discussion of these and related issues.

offer rational motivation for all agents. The initial step would involve establishing a minimum threshold of wellbeing. The precise nature and extent of wellbeing would still be debatable, but in the sense relevant to this thesis, this threshold could perhaps include acceptance of moral* norms promoting at least limited autonomy for all agents and norms which forbid excessive predation upon other agents, for example. Only moral* codes with these features would be deemed sufficient. Having established a basic minimum standard of what must be included in a moral* code, we could rule out any potential T^3 states which did not include acceptance of a moral* code with these features. Then, among the remaining potential T^3 states, we could seek to bring about whichever T^{3-x} would plausibly yield the greatest total aggregate degree of prudential benefit. By guaranteeing that all agents experience an acceptable minimum degree of prudential benefits, and then maximising the total sum of benefits above this level, perhaps some form of 'sufficientarianism' can guide us to a moral* code such that it would be rational for all agents to accept it.

There are several problems with this kind of sufficientarianism, two of which I will highlight. First, there seem to be no convincing grounds *available in the WNP context* for settling the question of what is and is not sufficient in the relevant sense. What we are looking for here is a moral* code which conduces to a T^3 state such that it would be rational for *all* agents to accept that code. But a reason for *all* agents - even cruel and predatory agents who would be happy to own slaves - to ensure an absence of slavery at T^3 would have to be some kind of categorical, objective reason. Yet this is precisely the kind of categorical objective reason which is ruled out by error theory. It seems likely that for any suggested standard of sufficiency, the same would apply – setting standards of sufficiency is incompatible with the truth of the error theory. The second problem with sufficientarianism is that, like egalitarianism, it also involves a kind of levelling down – again, we are looking for rational motivation for *all* agents to accept a moral* code, yet the contented oligarch has little interest

in accepting sufficientarianist moral* norms because she may lose out as a result of e.g. the redistribution of wealth required to guarantee the sufficient standard for everyone else.

iv) Rational Bargaining

Finally, fourth, we could insist that the best way to find a T^3 which it would be rational for all agents to bring about would be to rely on a process of rational bargaining. The notion of moral bargaining has been discussed in a relativist context by Gilbert Harman (see e.g. Harman & Thomson 1996 chapter 2).[230] But what I have in mind here is more similar to the view advanced by David Gauthier in *Morals by Agreement*.[231] The idea is that agents can realise greater advantages by cooperating than by living in isolation. This is easy to illustrate with a simple example, though naturally the picture can be much more complicated in real life. Think of two farmers. If each exclusively farms their own field, then each can grow 50kg of produce per year, making a total between the two of them of 100kg. Yet if they cooperate, for example if they work at different times of day and therefore more efficiently scare off birds which would otherwise eat the seeds sown on the fields, then between them they can grow 120kg of produce. Thus cooperation yields a 20kg 'cooperation dividend'. This dividend can then be divided between the farmers as they see fit.

Morality* comes into the picture through the norms which encourage this beneficial cooperation, and those which govern the distribution of the cooperation dividend. That

[230] It may be expected that since Harman is a relativist, and I am advocating a form of relativism, I would discuss his model of moral bargaining in preference to what I go on to discuss instead - Gauthier's view. But I read Harman's discussion of bargaining as primarily hermeneutic, intended to explain how we might have arrived at traditional morality, warts and all. Naturally in the present context, we are considering responses to the WNP, so such explanations are somewhat redundant. Gauthier's view, on the other hand, promises to offer a more direct response to the issues under discussion here.

[231] Once again, the view mentioned here is necessarily only the roughest sketch of a much more developed view, most famously presented in *Morals by Agreement* (Gauthier 1987). But broadly speaking, I believe what I call a 'Gauthierian' response here is sufficiently in keeping with the spirit of what he says to merit the adjective. For a third party source which can provide more information and analysis than space here permits, again see e.g. Cudd 1996

morality* should promote cooperation is no surprise. But in this example, different moral* codes could also lead to different distributions of the dividend. For instance, imagine that the applicable moral* code placed more emphasis on rewarding time spent working, and less emphasis on people's needs. In that case (which is roughly similar to traditional morality in the modern West), it would be considered morally* right that the farmer who worked the longest hours would receive a larger share of the dividend, even if she had no family to feed while the other famer has hungry children. Conversely, if the applicable moral* code primarily emphasised helping those with the greatest need, then it might be thought right that the farmer with a family should receive a greater proportion of the dividend even if he worked fewer hours. More balanced moral* codes could have a more nuanced view, and more agents cooperating will open up more complex distributive considerations, and so on. The question we are seeking to answer in this subsection therefore becomes a matter of what moral* code would yield a division of the cooperation dividend which it would be rational for all agents to agree upon.

The Gauthierian response is that it is rational for all agents to ensure that their maximum concession – the amount of the dividend which they agree should be taken by others – is as small as possible.[232] Thus agents may bargain with one another in an attempt to arrive at an agreement over the norms of distribution which facilitates further cooperation and which also allows each agent to minimise the portion of the dividend which they agree to concede to others. So long as the bargaining agents are rational, sufficiently informed, and free of

[232] Gauthier has subsequently updated his view (2013). But I will set this aside here, as while his earlier work is a widely influential classic in the field, there are reasons to be much more skeptical about his later view in the current context. For example, the later view assumes that agents are predisposed to cooperate (2013 p. 610), whereas WNP respondents are typically trying to promote cooperation among agents who are *not* so predisposed. As such, it is better to focus on the 'classic' view here, and forego the considerable digression that further discussion of Gauthier's later view would require.

coercion, this process of rational bargaining, it could be argued, is the best way to come up with a moral* code at T^3 which it would be rational for all agents to accept.

There are numerous problems with the rational bargaining view. One of the objections I wish to raise is that it is built entirely from a mistaken, inadvertently biased first step. In the background of the rational bargaining approach is an assumed view which is actually tendentious. This view is that humans are in some important sense free of social ties and potentially asocial to begin with, and then as instrumentally rational agents they freely participate in society in order to gain advantages of some kind. This view is hardly unfamiliar, and is taken as a neutral, uncontentious starting point by many philosophers - not least, as Holly Smith points out (1991 p. 230), Rawls and Hobbes. Gauthier makes this explicit: 'A person is conceived as an independent center of activity, endeavoring to direct his capacities and resources to the fulfilment of his interests' (1987 p. 9). A radical kind of individualism and a transactional view of the rationality of human interaction are baked into Gauthier's fundamental conception of what it is to be a person.

Yet this conception has been rejected by others for more than a century. An example here is John Dewey, who argued that individual liberty could not be understood separately from agents' social context:

> Society in its unified and structural character is the fact of the case; the non-social individual is an abstraction arrived at by imagining what man would be if all his human qualities were taken away. Society, as a real whole, is the normal order, and the mass as an aggregate of isolated units is the fiction. (Dewey 2008 p. 232) [233]

[233] See Festenstein 2018 §2 for more on this aspect of Dewey's thought.

The rational bargaining model rests on views of personhood and society which are by no means the obvious, philosophically neutral positions which they may seem. In the present context of attempting to establish what it is rational for all agents to do at T^2, a relatively uncontentious basis for any response to the RCC is required if we are to avoid considerable digression and backtracking through the arguments discussed in sizeable sections of this thesis. That contradictory views exist about underlying assumptions as fundamental as personhood and society means that rational bargaining has no such uncontested basis.

Even if we are willing to overlook this arguably profound misconception, there are other more immediate problems. First, the Gauthierian response's insistence that it is rational for agents to ensure that their maximum concession is as small as possible entails that in the bargaining process, an agent's bargaining position makes a difference. For surely she who contributes most, and who may therefore withhold the greatest future contribution if they are dissatisfied, must have the greatest say in how the dividend is distributed. Perhaps we could charitably weaken this aspect of the response and say instead that agents are rational if they ensure that their maximum concession is as small as possible, unless they desire to receive a smaller concession for some reason contingent only upon their desires and not upon any external factors.[234]

With this amendment in place, rational bargaining does not necessarily lead to the strongest or best resourced agents always getting more of the dividend than others – the most powerful bargainer could also be the most selfless person in the community. But it does mean that

[234] This weaker version, including the slightly awkward 'contingent only upon…' aspect, may be required in the WNP context. This is because the requirement that agents pursue maximum reward regardless of whether they may have desires to the contrary (e.g. relevant altruistic desires) is likely to be incompatible with the truth of error theory, and therefore violate the ROBET constraint I discussed in chapter 5. Since I already have several independent reasons for rejecting the rational bargaining model, I do not treat this as a separate objection to the view. But in other contexts, proponents of anything like the view I have described here may well have to confront this problem, too.

those in inferior bargaining positions may be forced to accept a bad deal. For example, anyone who is very old, disabled, too young to work and so on sometimes cannot contribute as much as others to society's dividend. Yet if others do not contribute, then these groups stand to lose an awful lot. On the other hand, very powerful groups or individuals may not depend on those less able groups' contributions to the dividend at all, and so will be in a position to force a very bad deal on to those groups if they wish (and given human behaviour throughout history it seems inevitable that, at least some of the time, they will so wish).

Our current society with its traditional moral outlook is often considerably more concerned with fairness and helping the less well off than this. So, if we assume a rough moral* equivalent of currently widely held moral beliefs as a baseline, anyone who is old, disabled, too young to work and so on would not be rational to accept this response to the RCC. Therefore the rational bargaining model once again fails to offer a way to find a moral* code which it would be rational for all agents as individuals to accept.

Drawing the findings in this subsection together, I conclude that it is not easy to see how any moral* system could be rational for *all* agents to accept. As we look toward T^3, then, we will have to take a different approach. Therefore in the next section, I will lay out a different approach which is based around what moral* code it would be in a *community's* interests to accept in order to bring about *aggregate* prudential benefits. This, I will argue, is how we can best respond to the RCC and how we can most profitably understand the benefits of adopting revolutionary relativism.

7.3.2. The revolutionary relativist's response

To recap, the original question posed by the RCC is this: from the perspective of the WNP at T^2, and looking to the future at T^3, is there a specific moral* code, or are there specific features

of multiple potential moral* codes, which it would be rational for us to adopt? This led to a further question of interpretation – whether we are talking about what it would be rational for all agents as individuals to do, or what it would be rational for moral* communities as a whole to do. I examined the possibility of finding a moral* code such that it would be rational for all agents as individuals to accept, and found that no answer was satisfactory. While my suggested strategies in this section may not be exhaustive of all the possible strategies, the conclusion (at least unless someone comes along with a more defensible position) must be that no answer to the RCC based on the rationality of all agents as individuals is likely to be successful.

That being the case, it seems that the prudential benefits offered by adopting my proposal must be thought of in an aggregate sense. Recall, though, that there was a worry about this. The example I gave above was where 90% of a community enslave the remaining 10% - the community may thereby be better off in an aggregate sense, yet this comes at a high cost for the 10% who are enslaved.

To consider this clearly, let us imagine a group of people who are at T^2, i.e. who are grappling with the 'what now?' problem. They have resolved to adopt revolutionary relativism and thus constitute an emerging moral* community. These people have not yet embarked upon a post-error theory life, having been traditional moralists until their recent 'revolution'. So, as is consistent with traditional morality, the members of the group are variously well off in terms of material wealth, available opportunities, happiness, eudaimonic wellbeing and so on. Let us suppose for the sake of argument that this means that each person has a roughly quantifiable degree of prudential goods, and that we could therefore identify variations in their prudential good 'score' at T^2 versus at various potential T^3 situations.

Now, imagine that this group accepts a moral* code M at T^3 such that the aggregate degree of prudential goods enjoyed by the community is increased in comparison with T^1. Let's say the prudential good score of each member of the group increases by 50 points at $T^{3\text{-}M}$ compared with the alternative possible T^3s. All members of the group, that is, apart from one individual. This unfortunate individual, let's call her Suzy, actually experiences a 'prudential score' fall of 5 points at $T^{3\text{-}M}$ compared with T^1. The overall effect on the group of adopting M is an aggregate increase in prudential goods, so it seems rational for the community as a whole to adopt M. And the same goes for nearly all members of the group as individuals.

This means that we may wish to place a constraint on all moral* codes which could instantiate M: *the likely consequences at T^3 of accepting M must be better than the available alternatives for sufficiently many members of the group to accept M that the community as a whole can be said to accept M*. Otherwise it would be impossible to bring about the desired level of cooperation and coordination at $T^{3\text{-}M}$.[235]

But it does not appear to be rational for Suzy to accept M. So what about her? Recall from §5.3.4 that on my proposal, acting in contravention of the moral* norms the community accepts is taken to justify reactive attitudes such as guilt and blame, and may lead to more severe censure. Thus when the majority of our group of intrepid budding revolutionary relativists accept moral* code M because it is to their advantage as individuals, and to the group as a whole's advantage on aggregate, Suzy will be dragged along because she faces censure if she does not comply. This remains the case even if it is not rational for Suzy as an individual to accept M herself. Thus Suzy is effectively coerced into accepting M, or at least into acting in accordance with M.

[235] This constraint is quite similar to the claim that in considering which moral* code to accept, communities must be what Railton called 'socially rational'. See Railton 1986, especially the second half of the paper.

Once M has become the accepted moral* code of her community, however, Suzy's position changes. Since she will face censure if she does not continue to toe the line, once at $T^{3\text{-}M}$, it becomes rational for Suzy to accept M, or at least to act in accordance with it. We might say that following the adoption of M by her community, Suzy is placed in a new 'incentive structure' – because of the attitudes of those around her, it comes to be Suzy's in selfish interests to adhere to M. This does not mean that Suzy must be resigned to her fate, however. Recalling chapter 6 (especially §6.2.2), if there were an amendment to M which could benefit Suzy – be that a pareto improvement for her individually, or a benefit to the community as a whole, including her - revolutionary relativism requires that the amendment be given proper consideration, and accepted if prudentially beneficial.

Beyond this, I leave it to specific communities to decide what constitutes an acceptably beneficial moral* code. It may be that within the constraint I have described, a community may conclude that some variety of pareto-optimific moral* code is best, or they may follow a sufficientarianist path which, for example, outlaws slavery, or they may choose any other way of settling the matter – I leave this to moral* communities themselves.

Crucially, it should be noted that this is not an abdication of responsibility on my part. Rather, this lack of specification is *required* if my proposal is to respect the ROBET and RC constraints (see §5.2.1 & §5.2.2). I have defined prudential benefits in terms of the satisfaction of the practical desires which the fully informed, fully rational counterparts of current agents would have for their current selves. And it is part of the argument for the error theory that the desiderative sets of current agents are sufficiently diverse that there can be no universally held desire among these counterfactually idealised counterparts of current agents (see §3.3.7 & e.g. Joyce 2001 §3.8). That being the case, to claim that there must be some universal desideratum is to claim that there are grounds for categorically authoritative practical reasons, which is obviously at odds with the truth of the error theory. This means that there

can be no 'one norm to rule them all' at any point in my proposal – within the constraints I laid out in chapter 5 different communities *must* be free to decide first-order moral* matters as they see fit.

This means that I reject the implication I mentioned near the start of §7.3 that I must advocate adopting revolutionary relativism *plus* moral* judgements with specific first-order contents. Revolutionary relativism remains a strictly second-order, metaethical proposal. In response to the RCC, I do not offer a specific moral* code or set of moral* principles which must feature in the set of appropriate moral* codes. Rather, I offer a constraint on appropriate moral* codes, and can go no further than this. This is, however, a positive result – it gets us beyond the impasse of the negative conclusion reached at the end of §7.3.1. It also offers scope for improving the lot of any potential 'Suzies' by highlighting how their position may be ruled out by specific moral* communities, and how their position may be improved if it is not ruled out by the norms accepted by the specific moral* community in which any given Suzy finds herself.

My direct response to the RCC, then, is no, there is no specific moral* code which it would be rational for us to adopt.[236] Nonetheless, in discussing how we might respond to the RCC, I have shown two important things. The first is that it is not surprising that there is no such specific moral* code, since specifying a moral* code would be inconsistent with the error theory, and is therefore impermissible in the WNP context. And the second is that revolutionary relativism nonetheless has the resources to get us further towards a response to the RCC which allows individual communities to respond as they see fit, including in ways which may help diffuse the worries which, on reflection, lie behind the RCC.

[236] Once again, for ease of reference, the RCC was formulated as follows: from the perspective of the WNP at T^2, and looking to the future at T^3, is there a specific moral* code, or are there specific features of multiple potential moral* codes, which it would be rational for us to adopt?

7.4. Implementation

The final potential objection to my proposal which I will discuss concerns implementation. An opponent could argue that even if my proposal is entirely successful on a theoretical level, it would be so difficult and require such enormous resources or psychological stress to actually implement that it simply is not a practical response to the WNP. And despite my concentration throughout this thesis on the theoretical (it is, after all, a philosophy thesis!), this is a realistic worry, certainly in comparison with some other WNP responses. As an example, one variation of this might be to argue that conservationism may seem to fare well here because it requires that agents simply continue to have their familiar moral beliefs. Thus, intuitively at least, conservationism seems comparatively easy to implement. I will briefly run through the concerns this potential objection involves, and then show why revolutionary relativism either avoids the worry in question, or compares more favourably to competing WNP responses than my notional opponent may think.

To state the objection as clearly and forcefully as possible, I would divide it into four related worries:

i) Revolutionary relativism is very complicated, and it is implausible that sufficient numbers of people would be able to understand it.

ii) Revolutionary relativism involves huge changes to how we speak and think about what were traditionally moral matters, so even if people could in principle grasp it, the educational resources required to implement the proposal would outweigh any prudential benefits of doing so.

iii) Even if we thought we were intellectually capable of making the required changes to our ways of thinking and speaking, and we were willing to devote

the required effort to trying to do so, it may turn out that we are not actually psychologically capable of making those changes.

iv) Other WNP responses could be so much easier than revolutionary relativism to implement that it would be worth overlooking any theoretical shortcomings and favouring them as the best WNP response instead.

Since the objection can be understood in terms of these four related worries, I shall reply to each of these points in turn. On the first point, I would reply by pointing to companions in guilt, and then by observing that all such companions are perhaps rather less guilty than it may seem. The companions I have in mind are pretty much all metaethical theories, both hermeneutic views and WNP responses. Metaethics in general is widely acknowledged to be a tricky and sometimes hard to grasp discipline, yet as a discipline, it is entirely aimed at understanding phenomena – moral thought and discourse – which are in evidence throughout almost all agents' lives. Virtually all mature human beings engage in moral discourse and make moral judgements by some definition or other, and virtually the whole of humankind manages to do so without finding themselves incompetent or under-qualified. In the most basic terms possible, if views such as quasi-expressivism or non-naturalist realism are at all credible, then it cannot be the case that revolutionary relativism is ruled out because it is too sophisticated.

Several times in the course of this thesis (in §5.3.1, and in footnotes 207 and 220) I have touched upon what it means to adopt my proposal, i.e. whether or not living roughly in accordance with my proposal would count, or whether adopting my proposal would require everyone to study advanced theories of metaethics. Each time I have stressed that while a more sophisticated understanding would potentially be helpful, an intuitive awareness or even just living in accordance with the proposal whilst remaining apparently oblivious to the details would suffice. To supplement and back up this claim, I again draw attention to an

article I discussed in §6.3.1 by Goodwin & Darley (2008), in which the authors show that 'lay' people are intuitive objectivists about some moral matters, and relativists about others. Despite the complexity of the many and various metaethical views which would come under the headings of objectivism and relativism, clearly lay people are competent users of moral terms. And in many cases this is without being more than dimly aware of the distinction between objectivism and relativism at best, yet alone the full details of the various available metaethical analyses of their thought and speech. This demonstrates that most people are capable of living consistently with advanced theories of metaethical relativism, even without training in metaethics. On this level, my proposal may even appear quite modest – along with some theoretical tweaks which would not necessarily be particularly evident on an everyday level, I am simply proposing that we take roughly the way some of our moral beliefs already work and extend it to cover all of our relevant beliefs.

In the WNP context more specifically, I would also argue that revolutionary relativism is not conspicuously more complicated than its competitors. Naturally in this context we start from a position of accepting a moral error theory, which must involve a degree of metaethical awareness.[237] And even then, the WNP response which it would arguably be most straightforward to adopt – conservationism, since it involves simply retaining the same moral beliefs we had before we accepted an error theory – is actually quite demanding of its adherents. This is because even conservationist agents would need to understand some quite sophisticated details of conservationism, or else conservationism itself would threaten to disintegrate into propagandism – a view dismissed out of hand by virtually all who have expressed a view on it (including me, in §4.5). Also, we must remember that according to error theorists, the correct hermeneutic view, based on how people actually use moral

[237] Here I leave out the possibility of the intuitive error theorist. It is simply not believable that there are very many agents who are untrained in metaethics, yet when confronted with the full details of a moral error theory would reply 'oh that – yes, obviously that's the case, I thought everyone knew *that*!'.

language, is some kind of non-naturalist realism (see *inter alia* chapter 2 of this thesis for more specifics). Therefore for post-error-theory agents to become any variety of fictionalist or expressivist would also require some quite extensive conceptual reupholstery for the majority of agents. In fact, it is plausible that a form of relativism – even a novel and slightly unusual form – would actually be rather easier for most agents to grasp, if Goodwin and Darley are right that most agents start out as intuitive relativists about some matters. And while abolitionists may be tempted to claim that simply abolishing moral thought and discourse would be conceptually straight forward, we saw in §4.3 that this can be seriously doubted. Overall, I reject the idea that the worry I labelled i) is a threat to my proposal.

My response to point ii) is similar, in that experimental philosophy shows us that most people have an intuitive understanding of moral relativism in general terms, and so the educational resources required to implement my proposal should not be all that vast. This must surely be the case in comparison with competing WNP responses. Revolutionary fictionalism is quite unlike the view of moral realism on which the error theory is predicated. Thus even in spite of potential ways of making it easier to explain (by analogies to acting, for example), Joyce's proposal would likely require significantly greater educational resources to implement than mine. The same goes for revolutionary expressivism – as a hermeneutic theory this criticism may not apply, but in the WNP context it has already been established that most people do not use moral discourse consistently with expressivism – rather, most people are intuitively moral realists of some kind.[238] So implementing Svoboda's proposal would need to include teaching huge numbers of people about a conception of moral thought and language which they do not intuitively grasp. Despite Garner's claims that abolitionism would be easy to adopt, as we saw in §4.3.2 (and touched upon again in §6.1), we have strong reasons to doubt this – at least one philosopher who has made a committed, long-term attempt to adopt

[238] As well as chapter 2 of this thesis, also see e.g. Enoch 2011a chapter 3.

abolitionism in everyday life has found it near-impossible to do so consistently.[239] This implies that the education and training resources required to implement abolitionism on a significant scale would be prohibitive. Given the above, I do not see how my proposal could be derailed by an objection that it would require excessive educational resources to implement.

My responses to points iii) and iv) are blunt: not only are we psychologically capable of embracing moral relativism, we actually do so already, including without even realising it. Thus it is not credible that we would be unable cope psychologically with an amended version of moral relativism such as the one I propose. And my responses to points i) and ii) show that at minimum, revolutionary relativism would be just as easy to implement as other WNP responses, if not in fact easier. Taking the above worries and responses together, I reject the claim that foreseeable difficulties around implementation could derail my proposal, especially in comparison with competing WNP responses proposed to date.

I wrote at the beginning of this chapter that it would be impossible to defend my proposal against every single existing or potential objection. But I have offered defences against what I feel are the strongest and most pressing objections possible, both those drawn from existing literature and those anticipated but not yet raised elsewhere. This concludes my defence of my proposal. I will now go on to close out this thesis with a short concluding chapter.

[239] Indeed, it may even be impossible to do so – see e.g. Streumer 2017 chapter 9.

Chapter 8. Conclusion

In this thesis, I have argued that if we accept a moral error theory, we should adopt a new metaethical view, which I have called revolutionary relativism. I began in chapter 2 by laying out how error theorists typically view traditional morality, in order to show the conceptual underpinnings of the moral error theory which would form the backdrop for the discussions which followed. I explained that error theorists typically hold that moral judgements are beliefs, and that moral discourse is typically assertoric and expresses those beliefs. Being beliefs, moral judgements are capable of being true or false, and error theorists typically claim that moral judgements are true only if the actions or situations they are about have some kind of property such that all agents have an inescapable, authoritative reason to act in accordance with the judgement, regardless of their desires or ends or any institutions they may be participating in. I then explained that error theorists argue that no action or situation can have the kind of property required for moral beliefs to be true, and that as a result, all moral beliefs or utterances which ascribe moral properties are false.

In chapter 3, I discussed how error theorists argue for the claim that nothing can have the relevant kind of property. I very briefly discussed the arguments by probably the most well-known moral error theorist, J. L. Mackie, and then moved on to two more recent and sophisticated arguments for a moral error theory, those of Jonas Olson and Richard Joyce. Both arguments focus on moral normativity, and the view that no agent can have an authoritative practical reason which does not relate to their desires, ends or to an institution which they are participating in. I argued that Olson's argument was less convincing than Joyce's, and that Joyce's is the most successful and influential argument for a moral error theory defended to date. That being the case, I went on to treat the matters discussed in chapters 2 and 3 as the background assumptions throughout the rest of the thesis. That is, I

went on to discuss what we should do if we accept an error theory roughly the same as that laid out at the end of chapter 2, for something like the reasons Joyce gives.

In chapter 4 I argued that accepting a moral error theory means that we are confronted with an unavoidable and urgent problem, the 'what now? problem, or WNP. The WNP is unavoidable because whether we engage with it or not, we will do *something* after we accept an error theory, and we fail as philosophers if we do not try to understand what the options might be and how to judge which the best option is. Some philosophers have argued that morality in general is a pernicious institution, and that we should therefore treat the WNP as an opportunity to rid ourselves of morality all together. Others have argued that morality is so useful to us that we would be better off if we retain moral thought and discourse, or at least adopt something quite like them in an attempt to hold on to some or all of the prudential benefits which abolishing morality risks losing. I outlined and discussed each of the main positions which have cropped up in the literature on this topic so far – abolitionism, conservationism, revolutionary fictionalism and revolutionary expressivism. I found all of them to be problematic, and therefore concluded that we must seek a new answer to the WNP.

In chapter 5, I offered my new response to the WNP, a proposal I call revolutionary relativism. I argued that we can respond to the WNP in a way which respects the commitments of error theory and which also avoids the pitfalls of the other WNP responses by replacing our previous moral beliefs with beliefs about practical norms which our communities accept. These beliefs could not properly be called *moral*, since they do not entail or presuppose the existence of the kind of categorical normativity which error theorists argue does not exist. But since they would be very similar to moral beliefs in use (if not on a conceptual level), I used the term moral*. I unpacked this by defining prudential benefit in a way which is compatible with error theory, and by grounding the claim that we should adopt my proposal in its capacity to deliver

this kind of prudential benefit. I then laid out what revolutionary relativism consists in – the structure, truth conditions and other commitments of moral* beliefs. I explained that moral* beliefs are beliefs about norms which agents' moral* communities accept, and that acceptance of a norm depends on adherence to that norm plausibly being prudentially beneficial. I also built in to moral* belief certain further commitments intended to ensure that agents are motivated to act in accordance with their moral* beliefs, and that there is scope for the community to reappraise the norms it accepts in line with the prudential benefits or harms of adhering to those norms.

In chapter 6 I began the defence of revolutionary relativism by comparing it with existing WNP responses in terms of how well it could cope with the problems I raised in chapter 4. I argued that many of the philosophers active in this area have failed to take seriously enough the claim by abolitionists that traditional morality is harmful, and that we cannot be completely sure whether it is harmful or not. I therefore compared my proposal with existing responses in two distinct scenarios – one in which traditional morality is not harmful, which I called moralgood, and one in which traditional morality is harmful, which I called moralbad. I compared my proposal with each of the existing WNP responses in turn, in each case arguing that revolutionary relativism could better cope with the objections I had raised in chapter 4 in both moralgood and moralbad, and so was the preferable response to the WNP overall.

Finally in chapter 7 I defended revolutionary relativism against the most pressing objections which I could foresee. The first objection I discussed was one of the most intractable problems for hermeneutic relativists, that moral relativism renders moral disagreement unintelligible. I argued that revolutionary relativism could account for moral disagreement in a way which copes with the disagreement problem. I set out several 'building blocks' for my argument, drawing on Gricean maxims, features of moral language and views of community participation. I then offered a story about how my proposal could accommodate various

different levels of disagreement, which would include the vast majority of likely cases of disagreement. And even in those cases which are more tricky for my proposal, it nonetheless avoids unintelligibility and provides scope for prudential benefit. I also discussed traditional problems for relativism around infallibility, dissidence and arbitrariness, in each case showing that they could not derail my proposal. I then moved on to respond to objections which may not typically arise for hermeneutic views, but which could be directed at my proposal because of its post-error-theory context. I discussed how we might decide which set of moral* beliefs it would be rational to adopt. I showed how this may be a more complex question than it appears, and considered various ways in which we might arrive at an answer. I concluded that there is no one set of moral* beliefs which it would be rational to adopt, but that this is unsurprising because specifying that communities must accept one set of beliefs over another would be inconsistent with the truth of error theory. Nonetheless, I argued that revolutionary relativism allows each moral* community to find its own optimally rational answer. Finally I discussed the worry that it may be so difficult or costly to implement my proposal that any prudential benefits of doing so would be outweighed. I argued that my proposal would be no more difficult to implement than any competing proposal, and may actually be easier to put into practise.

Drawing together all of the above, I conclude that revolutionary relativism is the best available response to the 'what now?' problem, and that if we accept a moral error theory, we should become revolutionary relativists.

Bibliography

Arneson, R., 2013. 'Egalitarianism', *The Stanford Encyclopedia of Philosophy* (Summer 2013 Edition), Zalta, E. N. (ed.), available online at https://plato.stanford.edu/entries/egalitarianism/ Accessed September 2019.

Austin, J. L. 1962. *How To Do Things With Words*. Oxford: Oxford University Press.

Baghramian, M. & Carter, J. A., 2018. 'Relativism', *The Stanford Encyclopedia of* Philosophy (Winter 2018 edition), Zalta, E. N. (ed.), available online at https://plato.stanford.edu/entries/relativism/ Accessed June 2019.

Balaguer, M., 2015. 'Fictionalism in the Philosophy of Mathematics', *The Stanford Encyclopedia of Philosophy* (Fall 2018 edition), Zalta, E. N. (ed.), available online at http://stanford.library.usyd.edu.au/entries/fictionalism-mathematics/ Accessed November 2019.

Bedke, M., 2014. 'A menagerie of duties? Normative judgements are not beliefs about non-natural properties', *American Philosophical Quarterly*, vol. 51, no. 3., pp. 189-201.

Berkeley, G., 2002. *A Treatise Concerning The Principles of Human Knowledge*, Wilkins, D. (ed.)., available online at https://www.maths.tcd.ie/~dwilkins/Berkeley/HumanKnowledge/1734/HumKno.pdf Accessed June 2019.

Björnsson, G. & Finlay, S., 2010. 'Metaethical Contextualism Defended', *Ethics*, issue 121, pp. 7-36.

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

Blackburn, S., 1998b. 'Moral Relativism and Moral Objectivity', *Philosophy and Phenomenological Research*, vol. 58, no. 1, pp. 195-198.

Blackburn, S., 2010. 'Must We Weep for Sentimentalism?', in *Practical Tortoise Raising and Other Philosophical Essays*. Oxford: Oxford University Press, pp. 109-128. Page numbers in the text refer to the online version available at https://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199548057.001.0001/acprof-9780199548057-chapter-7 Accessed October 2019.

Blackford, R., 2019. 'After Such Knowledge – What?', in *The End of Morality: Taking Moral Abolitionism Seriously*, Garner, R. & Joyce, R. (eds). Oxford: Routledge, pp. 59-76.

Boghossian, P., 2006. 'What is Relativism?', in Greenough, P. & Lynch, M. P. (eds.), *Truth and Relativism*. Oxford: Clarendon Press, pp. 13-37.

Bukoski, M., 2016. 'A Critique of Smith's Constitutivism', *Ethics*, vol. 127, pp. 116-146.

Burgess, J. P., 2007. 'Against Ethics', *Ethical Theory and Moral Practise,* vol. 10, pp. 427-439.

Cohen, S., 1972. *Folk Devils and Moral Panics*, London: MacGibbon & Kee.

Cook, J. *et al.*, 2016. 'Consensus on consensus: a synthesis of consensus estimates on human-caused global warming', *Environmental Research Letters*, vol. 11, no. 4.

Copp, D., 2001. 'Realist-Expressivism: A Neglected Option for Moral Realism', *Social Philosophy and Policy*, vol. 18, pp. 1-43.

Copp, D., 2009. 'Realist-Expressivism and Conventional Implicature', *Oxford Studies in Metaethics, vol. 4*. Oxford: Oxford University Press, pp. 167-202.

Cornock, D., 2012. 'Cameron: Wales has a strong voice at government's heart', available online at https://www.bbc.co.uk/news/uk-wales-19861810 Accessed June 2019.

Cree, V. E. (ed.) 2016. *Revisiting Moral Panics*, Bristol: Policy Press.

Crisp, R., 2003. 'Equality, Priority, and Compassion', *Ethics*, vol. 113, no. 4, pp. 745-763.

Cudd, A. E., 1996. 'Is Pareto Optimality a Criterion of Justice?', *Social Theory and Practice*, vol.22, no. 1, pp. 1-34.

Cuneo, T. & Christy, S., 2011. 'The Myth of Moral Fictionalism', in *New Waves in Metaethics*, Brady, M. (ed.). London: Palgrave Macmillan, pp. 85-102.

Daly, C. & Liggins, D. 2010. 'In defence of error theory', *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, vol. 149, no. 2, pp. 209-230.

DeScioli, P. & Kurzban, R., 2009. 'Mysteries of Morality', *Cognition*, vol. 112, pp. 281-299.

DeScioli, P. & Kurzban, R., 2013. 'A Solution to the Mysteries of Morality', *Psychological Bulletin*, vol. 139 no. 2, pp. 477-496.

Dewey, J., 2008. 'The Ethics of Democracy', in *The Early Works of John Dewey, Volume 1, 1882–1898: Early Essays and Leibniz's New Essays, 1882–1888*, Boydston, J. A. & Axetell, G. E. (eds). Carbondale, Illinois: Southern Illinois University Press.

Dreier, J., 1990. 'Internalism and Speaker Relativism', *Ethics*, vol. 101, no. 1, pp. 6-26.

Dreier, J., 1992. 'The Supervenience Argument Against Moral Realism', *The Southern Journal of Philosophy*, vol. 30, no. 3.

Dreier, J., 2009. 'Relativism and Disagreement', *Philosophical Perspectives*, vol. 23, pp. 79-110.

Dreier, J., 2010. 'Mackie's Realism: Queer Pigs and the Web of Belief', in *A World Without Values*, Joyce, R. and Kirchin, S. (eds). New York: Springer.

Eagan, A., 2007. 'Quasi realism and fundamental moral error', *Australasian Journal of Philosophy*, vol. 85, no. 2, pp. 205-219.

Eklund, M., 2015. 'Fictionalism', *The Stanford Encyclopedia of Philosophy* (Winter 2015 edition), Zalta, E. N. (ed.), available online at http://plato.stanford.edu/archives/win2015/entries/fictionalism Accessed November 2019.

Enoch, D., 2006. 'Agency, Shmagency: Why Normativity Won't Come from What Is Constitutive of Action', *The Philosophical Review*, vol. 115, no. 2, pp. 169-198.

Enoch, D., 2011a. *Taking Morality Seriously*. Oxford: Oxford University Press.

Enoch, D., 2011b. 'On Mark Schroder's Hypotheticalism: A Critical Notice of "Slaves of the Passions"', *The Philosophical Review*, vol. 120, no. 3.

Eshleman, A., 2016. 'Moral Responsibility', *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), Edward N. Zalta (ed.), available online at https://plato.stanford.edu/archives/win2016/entries/moral-responsibility/ Accessed June 2019.

Everett, D., 2008. *Don't Sleep, There Are Snakes*. London: Profile Books.

Finlay, S., 2008. 'The error in the error theory', *Australasian Journal of Philosophy*, vol. 86, pp. 347-369.

Finlay, S., 2009. 'Oughts and Ends', *Philosophical Studies*, vol. 143, pp. 315-340.

Finlay, S., 2011. 'Errors upon errors: A reply to Joyce'. *Australasian Journal of Philosophy*, vol. 89, pp. 535-547.

Finlay, S. & Schroeder, M., 2017. 'Reasons for Action: Internal vs. External', *The Stanford Encyclopedia of Philosophy* (Winter 2012 edition), Zalta, E. N. (ed.), available online at http://plato.stanford.edu/entries/reasons-internal-external/ Accessed November 2019.

Finlay, S., 2017. 'Disagreement Lost and Found', in Shafer-Landau, R. (ed.), *Oxford Studies in Metaethics, vol. 12*. Oxford: Oxford University Press, pp. 187-205.

Festenstein, Matthew, 2018. 'Dewey's Political Philosophy', *The Stanford Encyclopedia of Philosophy* (Fall 2018 Edition), Zalta, E. N. (ed.), available online at https://plato.stanford.edu/archives/fall2018/entries/dewey-political Accessed November 2019.

Foot, P., 1978. *Virtues and Vices*. Berkeley: University of California Press.

Garner, R., 2007. 'Abolishing Morality', *Ethical Theory and Moral Practise*, vol. 10, pp. 499-513.

Garner, R., 2011. 'Morality: The Final Delusion?', *Philosophy Now*, issue 82, January/February 2011. Available online at https://philosophynow.org/issues/82/Morality_The_Final_Delusion Accessed October 2019.

Garner, R. & Joyce, R. (eds) 2019. *The End of Morality: Taking Moral Abolitionism Seriously*. Oxford: Routledge.

Garner, R., 2019. 'A Plea for Moral Abolitionism', in *The End of Morality: Taking Moral Abolitionism Seriously*, Garner, R. & Joyce, R. (eds). Oxford: Routledge, pp. 77-93.

Gauthier, D., 1987. *Morals by Agreement*. Oxford: Clarendon Press.

Gauthier, D., 2013. 'Twenty-five on', *Ethics*, vol. 123, pp. 601-624.

Gendler, T. (2013). 'Imagination', *The Stanford Encyclopedia of Philosophy* (Fall 2013 edition), Zalta, E. N. (ed.), available online at http://plato.stanford.edu/archives/fall2013/entries/imagination/ Accessed October 2019.

Gibbard, A., 1990. *Wise Choices, Apt Feelings: A Theory of Normative Judgement*. Oxford: Oxford University Press.

Gibbard, A., 2005. 'Truth and Correct Belief', *Philosophical Issues*, vol. 15, pp. 338-350.

Goodwin, G. P. & Darley, J. M., 2008. 'The psychology of meta-ethics: Exploring objectivism', *Cognition*, vol. 106, pp. 1339-1366.

Green, M., 2017. 'Speech Acts', *The Stanford Encyclopedia of Philosophy* (Winter 2017 Edition), Zalta, E. N. (ed.), available online at https://plato.stanford.edu/archives/win2017/entries/speech-acts/ Accessed November 2019.

Grice, H. P., 1957. 'Meaning', *The Philosophical Review*, vol. 66, no. 3, pp. 377-388.

Grice, H. P., 1975. 'Logic and Conversation', in *Syntax and Semantics, vol. 3*, Cole, P. & Morgan, J. L. (eds). New York: Academic Press, pp. 41-58.

Grice, H. P., 1989. *Studies in the Way of Words*. Cambridge, Massachusetts: Harvard University Press.

Hájek, A., 2018. 'Pascal's Wager' *The Stanford Encyclopedia of Philosophy* (Winter 2012 Edition), Zalta, E. N. (ed.), available online at http://plato.stanford.edu/entries/pascal-wager/ Accessed November 2019.

Hare, R. M., 1952. *The Language of Morals*. Oxford: Oxford University Press.

Harman, G., 1975. 'Moral Relativism Defended', *The Philosophical Review*, vol. 84, pp. 3-22.

Harman, G., 1978. 'What is Moral Relativism?', in *Values and Morals*, Goldman, A. I. & Kim, J. (eds.). Dordrecht: D. Reidel Publishing Company, pp. 143-161.

Hattiangadi, A., 2006. 'Is Meaning Normative?', *Mind & Language*, vol. 21, no. 2, pp. 220-240.

Hinckfuss, I. 1987. *The Moral Society: Its Structure and Effects*. Full text available online at http://www.bim-bad.ru/docs/hinckfuss_ian_moral_society.pdf Accessed November 2019.

Hinckfuss, I., 2019. 'To Hell With Morality', in *The End of Morality: Taking Moral Abolitionism Seriously*, Garner, R. & Joyce, R. (eds). Oxford: Routledge, pp. 21-38.

Hooker, B., 2000. *Ideal Code, Real World: A Rule-Consequentialist Theory of Morality*. Oxford: Clarendon Press.

Horwich, P., 1998. *Truth* (2nd edition). Oxford: Oxford University Press.

Hubin, D. C., 1980. 'Prudential Reasons', *Canadian Journal of Philosophy*, vol. 10, no. 1, pp. 63-81.

Humberstone, I. L., 1992. 'Direction of fit', *Mind*, vol. 101, issue 401, pp. 59-83.

Hume, D., 1739. *A Treatise of Human Nature*, available online at http://www.davidhume.org/texts/thn.html Accessed November 2019.

Husi, S. 2014. 'Against Moral Fictionalism', *Journal of Moral Philosophy*, vol. 11, no. 1, pp. 80-96.

Jackson, F., 2008. 'The Argument from the Persistence of Moral Disagreement', in *Oxford Studies in Metaethics, vol. 3*, Shafer-Landau, R. (ed.). Oxford: Oxford University Press, pp. 75-86.

Jaquet, F. & Naar, H., 2016. 'Moral Beliefs for the Error Theorist?', *Ethical Theory and Moral Practice*, vol. 19(1), pp. 193-207.

Joyce, R., 2000. 'The Fugitive Thought', *The Journal of Value Inquiry*, vol. 34, pp. 463-478.

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

Joyce, R., 2002. 'Expressivism and motivation internalism', *Analysis*, vol. 62.4, pp. 336-344.

Joyce, R., 2005. 'Moral Fictionalism', in *Fictionalism In Metaphysics*, Kalderon, M. E. (ed.). Oxford: Oxford University Press, pp. 287-313.

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

Joyce, R., 2011. 'The error in "The error in the error theory"', *Australasian Journal of Philosophy,* vol.89, pp. 519-534.

Joyce, R., 2012. *Enough With The Errors! A Final Reply To Finlay*, available online at http://personal.victoria.ac.nz/richard_joyce/acrobat/joyce_2012_enough.with.the.errors.pdf Accessed November 2019.

Joyce, R., 2017. 'Fictionalism in Metaethics' in *The Routledge Handbook of Metaethics*, Plunkett, D. & McPherson, T., (eds). London: Routledge.

Joyce, R., 2019. 'Moral Fictionalism: How to Have Your Cake and Eat It Too', in *The End of Morality: Taking Moral Abolitionism Seriously*, Garner, R. & Joyce, R. (eds). Oxford: Routledge, pp. 150-166.

Kalderon, M. E., 2005. *Moral Fictionalism*. Oxford: Oxford University Press.

Kalf, W., 2013. 'Moral Error Theory, Entailment and Presupposition' *Ethical Theory and Moral Practice,* vol. 16, issue 5, pp. 923-37.

Kalf, W., 2019. 'The belief problem for moral error theory', *Inquiry*, DOI: 10.1080/0020174X.2019.1612779.

Katsafanas, P., 2013. *Agency and the Foundations of Ethics: Nietzschean Constitutivism.* Oxford: Oxford University Press.

Katsafanas, P., 2018. 'Constitutivism about practical reasons', in *The Oxford Handbook of Reasons and Normativity*, Star, D. (ed.). Oxford: Oxford University Press.

Khoo, J. & Knobe, J., 2018. 'Moral Disagreement and Moral Semantics', *Noûs*, vol. 52, pp. 109-143.

Köhler, S. & Ridge, M., 2013. 'Revolutionary Expressivism', *Ratio*, vol. 26, issue 4, pp. 428-449.

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

Korsgaard, C., 1999. 'Self-constitution in the ethics of Plato and Kant', *The Journal of Ethics*, vol. 3, pp. 1-29.

Korsgaard, C., 2009. *Self-Constitution: Agency, Identity, and Integrity*. Oxford: Oxford University Press.

Kripke, S., 1982. *Wittgenstein on Rules and Private Language*. Oxford: Blackwell.

Lenman, J., 2013. 'Ethics Without Errors', *Ratio*, no. 36, pp. 391-409.

Lillehammer, H. & Möller, N. 2015. 'We Can Believe the Error Theory', *Ethical Theory and Moral Practice*, vol. 18, issue 3, pp. 453-459.

Lutz, M., 2014. 'The "Now What" problem for error theory', *Philosophical Studies*, vol. 171, issue 2, pp. 351-371.

Lynch, M., 2004. *True To Life. Why Truth Matters.* London: MIT Press.

Mackie, J., 1977. *Ethics*. London: Penguin Books.

Mahon, J. E., 2016. 'The Definition of Lying and Deception', *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), Zalta, E. N. (ed.), available online at https://plato.stanford.edu/archives/win2016/entries/lying-definition/ Accessed November 2019.

Malatesti, L. & McMillan, J. (eds), 2010. *Responsibility and Psychopathy: Interfacing Law, Psychiatry and Philosophy*. Oxford: Oxford University Press.

Marks, J., 2019. 'Beyond the Surf and Spray', in *The End of Morality: Taking Moral Abolitionism Seriously*, Garner, R. & Joyce, R. (eds). Oxford: Routledge, pp. 94-109.

Merli, D., 2008. 'Expressivism and the Limits of Moral Disagreement', *The Journal of Ethics*, vol. 12, pp. 25-55.

Mezzofiore, G., 2015. *Ali Mohammed al-Nimr crucifixion: Defiant Saudi Arabia rejects-international interference in protester's crucifixion*. News article, available online at http://www.ibtimes.co.uk/ali-mohammed-al-nimr-crucifixion-defiant-saudi-arabia-rejects-international-interference-1522958 Accessed November 2019.

Miller, A. 2013. *Contemporary Metaethics* (2nd edition). Cambridge: Polity Press.

Moore, G. E., 1922. *Philosophical Studies*. New York: Harcourt, Brace & Co.

Nietzsche, F., 1998. *Beyond Good and Evil*, Faber, M. (trans.). Oxford: Oxford University Press.

Nietzsche, F., 2001. *The Gay Science*, Williams, B. (ed.), Nauckhoff, J. & Del Caro, A. (trans.). Cambridge: Cambridge University Press.

Nietzsche, F., 2007. *On The Genealogy of Morality* (Revised Student Edition), Ansell-Pearson, K., (ed.), Diethe, C. (trans.). Cambridge: Cambridge University Press.

Nolan, D., Restall, G. & West, C. 2005. 'Moral fictionalism versus the rest', *Australasian Journal of Philosophy*, vol. 83 no. 3, pp. 307-330.

Oddie, G. & Demetriou, D., 2007. 'The Fictionalist's Attitude Problem', *Ethical Theory and Moral Practise*, vol. 10, pp. 485-498.

Olson, J., 2010. 'In Defence of Moral Error Theory', *New Waves in Metaethics*, Brady, M. (ed.), New York: Palgrave Macmillan.

Olson, J., 2014. *Moral Error Theory – History, Critique, Defence*. Oxford: Oxford University Press.

Olson, J., 2019. 'Nihilism and the Epistemic Profile of Moral Judgement', in *The Routledge Handbook of Moral Epistemology*, Jones, K., Timmons, M. and Zimmerman, A. (eds.). New York: Routledge, pp. 304-15.

Parfit, D., 1997. 'Equality and Priority', *Ratio,* vol. 10, issue 3, pp. 202-221.

Perl, C. & Schroeder, M., 2019. 'Attributing error without taking a stand', *Philosophical Studies*, February 2019, pp. 1-19.

Pigden, C., 2007. 'Nihilism, Nietzsche and the Doppelganger Problem', *Ethical Theory and Moral Practise*, vol. 10, pp. 441-456.

Plato, 2008. *Republic* (Benjamin Jowett trans.). Available online at http://classics.mit.edu/Plato/republic.html Accessed November 2019.

Plunkett, D. & Sundell, T., 2013. 'Disagreement and the Semantics of Normative and Evaluative Terms', *Philosophers' Imprint*, vol. 13, no. 23.

Price, H., 1998. 'Three Norms of Assertibility, or How the Moa became Extinct', *Noûs*, vol. 32, Supplement: Philosophical Perspectives, 12, Language, Mind, and Ontology, pp. 241-254.

Railton, P., 1986. 'Moral Realism', *The Philosophical Review*, vol. 95, no. 2, pp. 163-207.

Railton, P., 1994. 'Truth, Reason, and the Regulation of Belief', *Philosophical Issues*, vol. 5, pp. 71-93.

Rawls, J., 1999. *A Theory of Justice (Revised Edition)*. Cambridge, Massachusetts: The Belknap Press of Harvard University Press.

Ridge, M. 2013. 'Disagreement', *Philosophy and Phenomenological Research*, vol. 86, issue 1, pp. 41-63.

Ridge, M., 2014. 'How to insult a philosopher', in *Having It Both Ways: Hybrid Theories and Modern Metaethics*, Fletcher, G. & Ridge, M. (eds). Oxford: Oxford University Press.

Robertson, S., 2008. 'How to be an Error Theorist about Morality', *Polish Journal of Philosophy*, volume II, no. 2, pp. 107-125.

Russell, B., 1910–11. 'Knowledge by Acquaintance and Knowledge by Description', *Proceedings of the Aristotelian Society*, vol.11, pp. 108–128.

Russell, H., 2013. 'The Free Rider Problem', *The Stanford Encyclopedia of Philosophy*, (Spring 2013 Edition), E. N. Zalta (ed.), available online at https://plato.stanford.edu/entries/free-rider/ Accessed November 2019.

Sayre-McCord, G., 2008. 'Hume on Practical Morality and Inert Reason', in *Oxford Studies in Metaethics, Volume 3*, Shafer-Landau, R. (ed.). Oxford: Oxford University Press.

Sayre-McCord, G., 2013. A Moral Argument Against Moral Dilemma, draft paper, available at http://philosophy.unc.edu/files/2013/10/A-Moral-Argument-Against-Moral-Dilemmas.pdf Accessed November 2019.

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

Schroeder, M., 2008. *Being For: Evaluating the Semantic Program of Expressivism*. Oxford: Oxford University Press.

Schwitzgebel, E., 2019. 'Belief', *The Stanford Encyclopedia of Philosophy* (Fall 2019 Edition), Zalta, E. N. (ed.), available online at https://plato.stanford.edu/archives/sum2015/entries/belief/ Accessed November 2019.

Shafer-Landau, R., 2004. *Whatever Happened to Good and Evil?* Oxford: Oxford University Press.

Shafer-Landau, R., 2005. 'Error theory and the possibility of normative ethics', P*hilosophical Issues*, vol. 15, pp. 107-120.

Shafer-Landau, R., 2010. *The Fundamentals of Ethics*. New York: Oxford University Press.

Shepski, L., 2008. 'The vanishing argument from queerness', *Australasian Journal of Philosophy*, 86:3, pp. 371-387.

Shoemaker, D. & Tognazzini, N. (eds) 2015. *Oxford studies in agency and responsibility. Volume 2, Freedom and resentment at 50*. Oxford: Oxford University Press.

Sinnott-Armstrong, W., 2019. 'Consequentialism', *The Stanford Encyclopedia of Philosophy* (Summer 2019 Edition), Zalta, E. N. (ed.), available online at http://plato.stanford.edu/entries/consequentialism/ Accessed November 2019.

Smith, H., 1991. 'Deriving morality from rationality', in *Contractarianism and Rational Choice: Essays on David Gauthier's Morals by Agreement*, Vallentyne, P. (ed.). Cambridge: Cambridge University Press, pp. 229-253.

Smith, M., 1993. 'Objectivity and Moral Realism: On the Significance of the Phenomenology of Moral Experience', in *Reality, Representation and Projection*, Haldane, J. & Wright, C. (eds). Oxford: Oxford University Press.

Smith, M., 1994. *The Moral Problem*. Oxford: Blackwell.

Smith, M., 1995a. 'Internal Reasons', *Philosophy and Phenomenological Research*, vol. 55, pp. 109-131.

Smith, M., 1995b. 'Internalism's Wheel', *Ratio*, vol. 8, no. 3, pp. 277-302.

Smith, M., 1998. 'Ethics and the A Priori: A Modern Parable', *Philosophical Studies*, vol. 92, pp. 149-174.

Smith, M., 2006. 'Is That All There Is?', *The Journal of Ethics*, vol. 10, no. 1/2, pp. 75-106.

Smith, M., 2010. 'Beyond The Error Theory', in *A World Without Values: Essays on John Mackie's Moral Error Theory*, Joyce, R. & Kirchin, S. (eds). New York: Springer, pp. 119-139.

Smith 2015. 'The Magic of Constitutivism', *American Philosophical Quarterly*, vol. 52, no. 2, pp. 187-200.

Smith, M., Lewis, D. & Johnston, M., 1989. 'Dispositional Theories of Value', *Proceedings of The Aristotelian Society, Supplementary Volumes*, vol. 6, pp. 89-174.

Sobel, D., 1999. 'Do the desires of rational agents converge?', *Analysis*, vol. 59, issue 263, pp. 137-147.

Sobel, D., 2019. 'Good and gold', in *The End of Morality: Taking Moral Abolitionism Seriously*, Garner, R. & Joyce, R. (eds). Oxford: Routledge, pp. 3-20.

Stevenson, C., 1963. *Facts and Values*. Yale University Press: New Haven, Connecticut.

Strawson, P. F., 1962. 'Freedom and Resentment', *Proceedings of The British Academy*, vol. 48, pp. 1-25. Available online at http://www.ucl.ac.uk/~uctytho/dfwstrawson1.htm Accessed October 2019.

Street, S., 2006. 'A Darwinian Dilemma for Realist Theories of Value', *Philosophical Studies,* vol. 127:1, pp. 109-166.

Streumer, B., 2011. 'Are Normative Properties Descriptive Properties?', *Philosophical Studies,* vol. 154, pp. 325-348.

Streumer, B., 2013a. 'Can We Believe the Error Theory?', *Journal of Philosophy,* vol. 110, pp. 194-212.

Streumer, B., 2013b. 'Why There Really Are No Irreducibly Normative Properties', in *Thinking About Reasons: Themes from the Philosophy of Jonathan Dancy*. Bakhurst, D., Hooker, B. & Little, M. O. (eds). Oxford: Oxford University Press.

Streumer, B., 2017. *Unbelievable Errors: An Error Theory about All Normative Judgements*. Oxford: Oxford University Press.

Suikkanen, J., 2013. 'Moral Error Theory and the Belief Problem', in *Oxford Studies in Metaethics, Volume 8*, Shafer-Landau, R. (ed.). Oxford: Oxford University Press, pp. 168-194.

Suikkanen, J., 2019. 'Contextualism, Moral Disagreement, and Proposition Clouds', in *Oxford Studies in Metaethics, volume 14*, Shafer-Landau, R. (ed.). Oxford: Oxford University Press, pp. 47-69.

Svoboda, T., 2015. 'Why Moral Error Theorists Should Become Revisionary Moral Expressivists', *Journal of Moral Philosophy*, advance article. References in the text are to the advance version, now available here - https://www.academia.edu/18083907/Why_Moral_Error_Theorists_Should_Become_Revisionary_Moral_Expressivists Accessed November 2019. Subsequently published in 2017, *Journal of Moral Philosophy*, vol. 14, issue 1, pp. 48-72.

Swinburne, R., 2008. 'God and Morality', *Think*, Winter 2008, pp. 7-15.

Tiberius, V., 2002. 'Maintaining Conviction and the Humean Account of Normativity', *Topoi*, vol. 21, pp. 165-173.

Tiberius, V., 2012. 'Open Mindedness and Normative Contingency', in *Oxford Studies in Metaethics, Volume 7*, Schafer-Landau, R. (ed.). Oxford: Oxford University Press, pp. 182-204. Also published online at https://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199653492.001.0001/acprof-9780199653492-chapter-6 Accessed October 2019. Page numbers in the text refer to the .pdf document of the online version.

Toppinen, T., 2015. 'Pure Expressivism and Motivational Internalism', in *Motivational Internalism*, Björnsson, G., Strandberg, C., Olinder, R. F., Eriksson, J. & Björklund, F. (eds). New York: Oxford University Press, pp. 150-166.

Topping, A. & Carson, M., 2014. 'FGM is banned but very much alive in the UK', *The Guardian*, available online at https://www.theguardian.com/society/2014/feb/06/female-genital-mutilation-foreign-crime-common-uk Accessed November 2019.

Williams, B., 1981. 'Internal and External Reasons', in *Moral Luck*. Cambridge: Cambridge University Press.

Williams, B., 1985. *Ethics and the Limits of Philosophy*. London: Routledge.

Wood, A., 1993. 'Marx against morality', in *A Companion To Ethics*, Singer, P. (ed.). Oxford: Blackwell.

Wong, D., 1984. *Moral Relativity*. Berkeley: University of California Press.

Wrenn, C., 2010. 'True Belief is not Instrumentally Valuable', in *New Waves In Truth*, Wright, C. & Pedersen, N. (eds). New York: Palgrave MacMillan, pp. 174-188.

Wright, C., 1992. *Truth and Objectivity*. Cambridge, Massachusetts: Harvard University Press.

Wright, C., 1996. 'Truth in Ethics', in *Truth in Ethics*, Hooker, B. (ed.). Oxford: Blackwell.