The Error in Moral Discourse and
What to do about it

A thesis submitted to the University of Manchester for the degree of
Doctor of Philosophy in the Faculty of Humanities.

2011

Philip Brown
School of Social Sciences
Contents
Abstract .................................................................................................................................................. 4
Declaration ............................................................................................................................................... 5
Copyright Statement ............................................................................................................................. 5
Acknowledgements ............................................................................................................................... 6
Introduction ............................................................................................................................................ 7
1: Error Theory ....................................................................................................................................... 10
1.1 What is error theory? ....................................................................................................................... 11
1.2 Six Metaethical Theses .................................................................................................................. 15
1.2.1 The Assertion Thesis ................................................................................................................ 15
1.2.2 The Alethic Thesis ..................................................................................................................... 22
1.2.3 The Objectivity Thesis ............................................................................................................... 23
1.2.4 Moral Rationalism ...................................................................................................................... 34
1.2.5 Motivational Internalism ............................................................................................................ 41
1.2.6 The Supervenience Thesis ........................................................................................................ 46
1.3 Arguments for error theory ............................................................................................................ 47
1.3.1 The argument from queerness .................................................................................................. 49
1.3.2 The argument from relativity ..................................................................................................... 59
1.3.3 An Evolutionary Debunking Argument ..................................................................................... 63
1.3.4 Richard Joyce’s argument for error theory ............................................................................... 74
1.4 Conclusion ....................................................................................................................................... 81
2: Objections ......................................................................................................................................... 83
2.1 Error theory is self-undermining .................................................................................................... 86
2.2 Error theory defies commonsense .................................................................................................. 89
2.3 Moral facts are possible, thus moral facts are actual ...................................................................... 92
2.4 Partners in guilt arguments (1): The epistemic argument ............................................................ 99
2.4.1 The epistemic argument ........................................................................................................... 99
2.4.2 Some attempts at motivating the claim that epistemic reasons are categorical ......................... 106
2.4.3 Epistemic Virtues ...................................................................................................................... 115
2.4.4 Cuneo’s Arguments against Reductionism ............................................................................... 120
2.4.6 Summary of section 2.4 .......................................................................................................... 126
2.5 Partners in guilt arguments (2): The argument from the categorical requirements of practical reasoning .............................................................................................................................. 127
2.6 Conclusion ....................................................................................................................................... 135
3: Eliminativism and Relativism ........................................................................................................ 137
3.1 Some Options .................................................................................................................................. 140
3.2 Eliminativism ................................................................................................................................... 143
3.2.1 Is morality harmful? .................................................................................................................. 146
3.2.2 Should we eliminate moral discourse just because moral judgements are untrue? ............... 154
3.3 Reformist Realism: Moral Relativism for Moral Error Theorists ............................................. 160
3.3.1 Is moral relativism even coherent? ...................................................................................... 162
3.3.2 So should we turn relativist? .............................................................................................. 165
3.4 Conclusion .............................................................................................................................. 167
4: Fictionalism and Conservationism ............................................................................................ 168
4.1 What is fictionalism? ................................................................................................................. 169
4.2 Nolan, Restall and West ............................................................................................................. 183
4.2.1 What might fictionalist moral utterances help us to express? ................................................. 188
4.3 Pretence or non-pretence? ......................................................................................................... 195
4.4 The self-conscious fictionalist .................................................................................................. 201
4.5 Richard Joyce’s moral fictionalism .......................................................................................... 205
4.5.1 Joyce’s motivation for fictionalism ...................................................................................... 206
4.5.2 Why, according to Joyce, we needn’t believe the fiction ...................................................... 209
4.5.3 Why can’t we self-consciously employ the fiction? .............................................................. 213
4.6 Conservationism ......................................................................................................................... 216
4.6.1 Objections to conservationism ............................................................................................. 219
5: Summary and Conclusion ............................................................................................................ 226
Appendix ....................................................................................................................................... 231
References ........................................................................................................................................ 241

Word count: 76,333
Abstract

Moral error theory is the thesis that statements attributing moral properties to objects are always untrue. In my thesis I have two aims: to argue for error theory and defend it from a number of objections (chapters 1 and 2); and to consider whether and how we should go on with moral discourse, if we come to accept error theory (chapters 3 and 4).

In the first chapter I argue for error theory by defending a number of metaethical theses which taken together reveal that sincere utterances of indicative moral sentences commit us to ‘objectively prescriptive values’. I then go on to defend the arguments of J.L. Mackie and Richard Joyce that such values do not exist, and thus indicative moral sentences are systematically untrue.

In the second chapter I deal with five objections to error theory: (i) error theory is self-undermining; (ii) error theory defies commonsense; (iii) error theory is defeated by a modal counterargument; (iv) moral error theory entails an absurd epistemic error theory; (v) the error theorist’s denial that there are any categorical imperatives is untenable, as it is constitutive of being a rational agent that one is guided by certain categorical norms. I show how each of these objections can be dealt with.

In the third chapter I begin to look at what we should do with moral discourse once we have accepted error theory. The main foci of this chapter are eliminativism (the thesis that we ought to stop engaging in moral discourse) and reformist realism (the thesis that we should modify the semantics of moral discourse so that our moral terms can successfully refer). I argue that the rationales that have been provided for eliminativism, such as that moral societies are harmful to most of their members, and that it always harmful to make untrue judgements, are unpersuasive. I consider the most plausible way of being a reformist realist is to argue that we should become moral relativists, but I argue that such a move would be unstable and we would revert to non-relativist type.

In the fourth chapter I focus on moral fictionalism (the thesis that we should continue to use indicative moral sentences, while adopting some attitude to them other than belief). One of the more interesting motivations that has been offered for moral fictionalism is that moral discourse either facilitates or is essential for some non-moral description. I find no evidence for the stronger claim, but argue that there is some plausibility in the weaker claim. Another interesting suggestion (made by Joyce) is that pretending that certain actions are morally required or forbidden will help motivate prudent behaviour. I argue that although it is very plausible that thinking in moral terms can motivate prudent behaviour, Joyce fails to provide a convincing argument that we can retain these positive motivational effects if we abandon moral belief. In light of this, I argue that the only way for error theorists to retain the positive motivational aspect of moral discourse is by becoming conservationists. The conservationist argues that we can and should continue to form and be guided by genuine moral beliefs, even if we have become convinced of the error theory. Naturally, conservationism is open to a variety of objections, and I deal with the ones that seem most pressing.
Declaration

I declare that no portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time.

The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the “intellectual property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in the thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://www.campus.manchester.ac.uk/medialibrary/policies/intellectual-property.pdf), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.campus.manchester.ac.uk/library/aboutus/regulations) and in The University’s policy on presentation of Theses.
Acknowledgements

I would like to thank my thesis supervisors, Michael Scott and David Liggins, for the many valuable discussions I have had with them throughout my graduate study, and for their much appreciated encouragement.

The Philosophy Discipline Area at Manchester has been a wonderful place to study. The atmosphere has always been friendly and the staff and graduate students have provided me with a great deal of support and helpful criticism of my work. I would particularly like to thank Chris Daly, Philip Letts, Hichem Naar, Paula Satne, Joel Smith, Tom Smith and Graham Stevens.
Introduction

This thesis combines a defence of moral error theory with an exploration of the prospects for moral discourse if error theory is true. The first task occupies chapters 1 and 2, while the second task occupies chapters 3 and 4.

Moral error theory is the thesis that sentences predicating moral properties to objects are systematically untrue because there are no moral properties. In chapter 1 I discuss and develop some of the leading arguments for error theory. I present a number of theses that figure prominently in the metaethical literature, and show how the error theorist can exploit the tensions that exist between these theses to cast doubt on the idea that moral discourse can perform its distinctive function. In this way I aim to show that error theory is true.

In chapter 2 I respond to some of the most pressing objections to error theory. It has been claimed that error theory is self-defeating, and that it defies commonsense. I show how, ironically, the claims that motivate the self-defeating charge are themselves self-defeating, and how limited the resources of ‘commonsense’ arguments are when used against error theory.

It has been argued that the mere possibility of moral facts, together with the almost universally accepted thesis of the supervenience of moral properties on non-moral properties, is sufficient to demonstrate that there are moral facts at this world (which is denied by error theorists). I show how several arguments for error theory can easily be expanded so as to rule out not only the actuality, but the possibility of moral facts.

A number of authors have recently argued that the reasons that error theorists have given for denying the existence of moral facts should be equally effective against the existence of epistemic facts, but that the idea of an epistemic
error theory is absurd. Thus it is claimed that arguments for error theory have absurd implications beyond the moral sphere. I argue that epistemic discourse is not analogous with moral discourse to the extent necessary for this objection to succeed. Similarly, it has been argued that our reasons to do what is (non-morally) rational stand or fall with our moral reasons. I respond that the fact that an action would be rational is never the ultimate reason to do it.

In chapter 3 I begin to examine the options for what to do with moral discourse if error theory is accepted. I begin by briefly setting out what I take the various options to be. Then I begin a more detailed examination of the more interesting ones. In chapter 3 I look at eliminativism – the view that we should cease using moral discourse if error theory is true – and reformist realism – the view that we should change the meaning of moral terms in such a way that they will successfully refer.

Eliminativism has been defended on two grounds. Firstly, on the grounds that morality is harmful to most people other than the elite of society, and secondly on the grounds that it is unwise to carry on having false beliefs (whatever the nature of the belief) if this can be avoided. Ian Hinckfuss has provided the most forceful attack on morality as a source of harm. I argue that his depiction of an elitist and socially damaging ‘moral society’ is overblown and unpersuasive.

As to the second defence of eliminativism, it has been claimed that swallowing the myths necessary to support belief in moral facts will leave us as ‘epistemological wrecks’ (Garner 1993: 96), and that it is always to our advantage to reason from true beliefs. I show why both of these claims are too strong.

Lastly in chapter 3, I discuss the possibility of ‘reformist moral realism’, i.e. the attempt to modify central moral concepts in such a way that they do refer to
entities we can believe in. The most likely model for this I take to be moral relativism. But I argue that, even in its most plausible form, a relativised moral discourse is likely to be unstable, as we will always have reason to treat moral judgements as universal.

In chapter 4 I discuss moral fictionalism and conservationism. Both of these options allow us to retain moral discourse, without altering its semantics. After a preliminary discussion of the various forms that fictionalism can take, I consider its potential as a response to error theory. The important issues are the motivation for retaining moral discourse as a fiction, and whether morality can perform the function we want it to perform if it is considered to be a fiction. I argue that the only credible motivations for moral fictionalism are (i) that it allows for a more economical expression of some non-moral information; and (ii) that moral thought has some positive motivational benefits. I think (i) is plausible, although the benefits may be quite marginal. As for (ii), I agree that moral thought has positive motivational effects, but I argue that it cannot have such effects if it is regarded as a fiction. Richard Joyce (2001) has attempted to deal with this problem by arguing that there need be no phenomenological difference between a moral fictionalist and a genuine moral believer in every day contexts, but I resist his argument by showing how it leaves us with an implausible criterion for belief.

Next I discuss conservationism. This is the position that we can continue to engage in moral discourse, and to form and be guided by genuine moral beliefs in our day-to-day lives, even if we accept the error theory when we are thinking about metaethics. I defend this thesis, which has been independently developed by Jonas Olson (2011b), against a number of objections.
1: Error Theory

In this chapter I discuss the metaethical position known as *moral error theory* (henceforward just ‘error theory’, unless I need to distinguish moral error theory from other forms of error theory). I start with a preliminary characterization of error theory, accompanied by a sketch of what metaethics is, and what its methods are (section 1.1). I then go on to set out a number of metaethical theses that are central to the realism versus antirealism debate in metaethics (1.2). These are the *Assertion Thesis* (the claim that in uttering an indicative moral sentence, the speaker is attempting to assert a fact rather than, say, express a feeling); the *Alethic Thesis* (the claim that some moral statements are true); the *Objectivity Thesis* (the claim that the truth value of a moral statement varies independently of anyone’s actual feelings, beliefs, or decisions); *Moral Rationalism* (the claim that if there are moral facts, they give us irreducibly moral reasons for action); *Motivational Internalism* (the claim that a moral judgement is sincere (or properly formed) only when the person making the judgement is motivated *ceteris paribus* to comply with it); and the *Supervenience Thesis* (the claim that moral properties supervene on non-moral properties). I think four of these theses (*Assertion, Objectivity, Moral Rationalism* and *Motivational Internalism*) are simply true. The other two (*Alethic* and *Supervenience*) are true if moral discourse is in good standing.

Having done that, I go on to show how these theses figure in the arguments of the two most prominent error theorists, J.L. Mackie and Richard Joyce (1.3). I will deal with some relevant objections along the way, but I deal with objections to error theory at greater length in the following chapter.
1.1 What is error theory?

Error theory is a form of moral scepticism. Formally, it is akin to atheism. The theist holds beliefs about such things as God and the afterlife. The atheist thinks that all of these beliefs are untrue\(^1\) because the truth conditions of the sentences that express them do not obtain: there is no God, there is no afterlife. Similarly, the error theorist thinks that moral beliefs are systematically untrue: they have truth conditions that do not obtain, and perhaps could not obtain.\(^2\)

Introducing talk of moral ‘belief’ and ‘truth’ at the beginning of a metaethical discussion might be thought to beg important questions, such as whether our moral attitudes really are belief-like, and whether indicative moral sentences are capable of bearing truth-values. An important strand in metaethics – noncognitivism – has traditionally denied both these claims. It should be noted that over the past 30 years or so noncognitivists have attempted to develop theories that allow us to keep talking about moral belief and moral truth, without giving up the central noncognitivist tenet that the function of moral language is not to describe how the world is, was, or might be, but rather to express noncognitive attitudes (such as, but not limited to, attitudes of approval or disapproval) at how the world is, was, or might be. However, it may still be best to have a more neutral term – moral judgement – for the focus of metaethical debate.

A moral judgement is what we have or form when events we see or hear about strike us as wrong, or unjust, or unfair. It is what we arrive at when, in thinking about what to do, we conclude that some course of action would be morally right or wrong in the circumstances. And it is what motivates much of the advice,

---

1 I say ‘untrue’ rather than ‘false’ to respect the view that statements that quantify over non-existents are neither true nor false (see Strawson 1950). The outcome of that debate is immaterial here, however.
2 I’ll discuss the modal status of error theory in chapter 2.
criticism or praise we give concerning the actions and attitudes of others (or would give, if we had the nerve). Metaethicists concern themselves with the status of these judgements, in a way that is best explained by distinguishing metaethics from other areas of moral philosophy.

Moral philosophy can be split into three areas: applied ethics, normative ethics, and metaethics. Whilst the boundaries between these areas are blurry, the following general characterizations hold. Applied ethics shares the same aims as much non-philosophical discussion of moral issues, which is primarily focussed on sorting out the right from the wrong on particular issues, such as euthanasia or animal experimentation. Normative ethics is concerned with developing general theories of when actions are good or bad, right or wrong, of which the most widely discussed are consequentialist, deontological and virtue theories. Metaethics is concerned with a broad range of questions about such things as the meanings of moral terms, what speech acts are performed in uttering moral sentences, whether moral sentences have truth values, what the metaphysical implications of our moral judgements are, and how – if at all – we can have moral knowledge. It is also concerned with moral psychology and the nature of reasons. These are closely related issues – what we say about any one of them will affect what we say about the others – and the various metaethical theories usually have something to say about each of them.

Metaethical theorizing has its starting point in our pre-theoretical judgements about these issues, which also act as constraints on, or tests of, our theories. One of the key difficulties in metaethics – a difficulty that arises in many areas of philosophy – is that of how to accommodate all of our pre-theoretical judgements within a single theory. Michael Smith calls this ‘the moral problem’ (1994). He
identifies three intuitively appealing metaethical claims that appear to be inconsistent:

(1) Moral judgements of the form ‘It is right that I φ’ express a subject’s beliefs about an objective matter of fact, a fact about what it is right for her to do.

(2) If someone judges that it is right that she φs then, ceteris paribus, she is motivated to φ.

(3) An agent is motivated to act in a certain way just in case she has an appropriate desire and a means-end belief, where belief and desire are, in Hume’s terms, distinct existences. (1994:12)

Smith comments:

The apparent inconsistency can be brought out as follows: from (1), the state expressed by a moral judgement is a belief, which from (2), is necessarily connected in some way with motivation; that is, from (3), with having a desire. So (1), (2) and (3) together entail that there is some sort of necessary connection between distinct existences: moral belief and desire. But (3) tells us that there is no such connection. Believing some state of the world obtains is one thing, what I desire to do given that belief is quite another. (12)

A consistent theory could be developed by rejecting any one of these claims. Noncognitivists will either reject (1), or, more typically nowadays, offer a nuanced reading of it such that it isn’t inconsistent with (2) and (3). However, many moral realists (henceforward just ‘realists’) have rejected either (2) or (3).

I should say briefly what I mean by the term ‘moral realism’, which is used somewhat differently by different authors. I will use the term to cover a diverse range of theories that are united in being neither versions of error theory nor
versions of noncognitivism. Realists are distinguished from noncognitivists in that they take the function of indicative moral sentences to be the representation of moral facts (whereas noncognitivists take the function of such sentences to be the expression of noncognitive attitudes). They are distinguished from error theorists in that they consider at least some indicative moral sentences to be true.

So, in common with realists, error theorists take moral discourse to be descriptive, and thus take it that moral discourse is only performing its function successfully if it succeeds in describing how things are, morally speaking. In common with noncognitivists, however, error theorists deny that there is a moral reality to be described. The appeal of error theory lies in the fact that it can avoid the serious problems that are thought to beset both realism and noncognitivism. Some forms of realism have highly implausible metaphysical and epistemological implications, while others have serious difficulty accommodating some of our most fundamental intuitions about the nature of morality. Noncognitivists, on the other hand, have difficulty in giving an account of why moral sentences appear to behave logically in a way that is indistinguishable from how ordinary descriptive sentences behave. I will discuss these matters in due course.

One interesting feature of error theory, which distinguishes it from all other metaethical theories, is that it is, in itself, neither revisionary nor non-revisionary. Most theories aim to be non-revisionary. They think that ordinary users of moral language use it successfully – realists think we succeed (at least some of us, some of the time) in describing how things morally are, and noncognitivists think we succeed in expressing our noncognitive attitudes; and as they think moral discourse performs a useful (perhaps vital) function, they don’t think there is any reason to change the way we use it. It is true that some virtue ethicists have argued that there
are serious problems with the moral concepts employed by ordinary users of moral language (e.g. Anscombe 1958); and the most prominent defender of moral relativism, Gilbert Harman, admits, at least in one place, that the moral concepts employed by most ordinary users are non-relativistic (in his contribution to Harman and Thompson 1996). In view of this, these theorists have argued that we ought to revise our moral concepts, and, it seems, consistency would demand that they urge revision (as they think there are better moral concepts available than the ones that most people have). However, whilst error theorists agree with these revisionists that the moral concepts of the typical user of moral discourse are defective, the question of whether (and how) to revise either the form of the discourse, or our attitude to it, remains on the table, even after we have accepted error theory. I discuss these matters further in chapters 3 and 4.

In *The Moral Problem*, Smith seeks to defend a version of moral realism that can accommodate the intuitions expressed by (1), (2), and (3). In this chapter I defend a range of metaethical theses (some of them related to (1), (2), and (3)), with the aim of showing that the only metaethical theory that can accommodate them is error theory. So let us now turn to examine the theses.

### 1.2 Six Metaethical Theses

#### 1.2.1 The Assertion Thesis

This is the thesis that indicative moral sentences attribute moral properties to objects. This is likely to sound like a truism, but we cannot treat it as one because some metaethicists deny it.
To describe something is to attribute properties to that thing, so to describe something morally is to attribute moral properties to that thing. The first thesis is simply that moral statements are descriptive or ‘assertoric’, in the sense that to utter a sentence of the form “X is F”, where “F” is a moral predicate, is to assert that X has the property of being F. Another way to put this is to say that moral discourse is ‘cognitive’ (implying that moral judgements are beliefs).

Assertion Thesis  In uttering an indicative moral sentence of the form “X is F”, where “F” is a moral predicate, a speaker asserts that X has the moral property of being F.

I am going to call this thesis ‘Assertion’ for short. Assertion is controversial because although everyone accepts that thick moral terms have descriptive content, it is not universally accepted that thin moral terms do. This is because not everyone believes that moral properties exist. If this is the case, and if moral discourse is in good standing, then sentences that appear to attribute moral properties to objects cannot, in fact, do so. This is the position taken by noncognitivists. Noncognitivists accept that thick moral terms have descriptive content, for they at least distinguish the type of behaviour that is being evaluated, but they deny that there are evaluative properties – goodness or badness, for instance – that are described. Instead, they hold that, in offering a moral evaluation of a character or action, we are not describing that object, but expressing some non-

3 A thick moral term is one that in some way combines description and evaluation. To call someone cruel is plausibly to criticize them morally, and to criticize them for a particular type of action (the intentional causing of unnecessary suffering). So ‘cruel’ is a thick moral term. To call someone bad or good, by contrast is to morally criticize or praise them, but not in a way that specifies why the person is being criticized or praised. ‘Bad’ and ‘good’ are thin moral terms. The terminology originates with Bernard Williams (1985). Not everyone accepts that supposedly thick terms have a fixed evaluative content (Blackburn 1992; Gibbard 1992).
belief (noncognitive) state, such as an emotion, a desire, or an intention. Thus noncognitivists defend varieties of the following thesis:

Expressivist Thesis In uttering an indicative moral sentence of the form “X is F”, where “F” is a moral predicate, a speaker does not assert that X has any moral property, but expresses a noncognitive attitude to X.

This thesis is offered as an elucidation of what we are actually doing when we utter indicative moral sentences, rather than as a suggestion for reforming how we use moral language. Noncognitivism is thus a non-revisionary theory.\(^4\)

One of the earliest defences of noncognitivism, that of A.J Ayer, was motivated by Ayer’s need to say something about the status of moral judgements, given his defence of the verificationist theory of meaning (Ayer 1946). On that theory, a statement has meaning only if it is either (i) verifiable (in principle) by empirical observation, or (ii) it expressed a tautology. Ayer accepted the arguments of G.E. Moore that moral properties – if they existed – could not be natural properties. These are properties which are either directly observable via the ordinary senses, or inferable from observation (so they are properties that are amenable to the natural sciences).\(^5\) As non-natural properties are, by definition, non-observable, Ayer claimed that no sentence could meaningfully refer to them. A general statement such as “Stealing is wrong” has no ‘factual meaning’ according

---

\(^4\) It might be better to say that it is standardly offered as a non-revisionary theory. I mention revolutionary expressivism as a possible response to error theory, in section 3.1.

\(^5\) This argument proceeds via Moore’s famous ‘open question argument’. The gist of this argument is that although moral properties have often been identified with some natural property – e.g. good with pleasant or desired – these identities cannot hold, because it is always an open question, of anything said to be pleasant or desired, whether it is also good, which, Moore claims, would not be the case if the properties were in fact identical (Moore 1903: 5-21).
to Ayer, but only serves to express moral disapproval. Although a statement about a particular action, such as “X acted wrongly in stealing that money”, conveys the factual content that X stole some money, the ‘wrongness’ that appears to be attributed to the act is not, in fact, some further factual information. Instead, the adverb ‘wrongly’ merely serves, once again, to express the speaker’s disapproval (Ayer 1946:110).

The verificationist theory has not survived. Contemporary noncognitivists motivate their position in a number of different ways, the most important of which is by stressing the practical function of moral deliberation and debate. In engaging in these activities, we seem to be trying to find out what to do or avoid doing, and what to approve of or disapprove of. But if moral facts are just facts of the same sort as facts about, say, the location or velocity of physical objects, then it can seem mysterious how in discovering the truth about them, we are finding out anything about what we ought to do or approve of. This is how Simon Blackburn puts it:

The reason expressivism [a form of noncognitivism] in ethics has to be correct is that if we supposed that, belief, denial, and so on were simply discussions of a way the world is, we would still face the open question. Even if that belief were settled, there would still be issues of what importance to give it, what to do, and all the rest. For we have no conception of a ‘truth condition’ or fact of which mere apprehension by itself determines practical issues. (Blackburn 1998: 70)

I think Blackburn is mistaken here. Error theorists agree with Blackburn that there are no such facts (and perhaps that there could be none), but it is wrong to say that we have no conception of them. My aim in this chapter is to show that we do indeed think of moral facts as things that determine practical issues.
One obvious problem for Ayer’s account (besides its verificationist motivation) is that it implies that we are radically mistaken when we think of our moral judgements as being *truth-apt*, i.e. as being the sort of judgements that are capable of truth or falsity. The error theorist, of course, holds that none of our judgements attributing moral properties to objects are true, which is radical enough, and attributes widespread error to users of moral discourse; but it does not attribute to speakers any misunderstanding of what they are doing when they engage in that discourse. In general, it is easier to believe that ordinary speakers are radically mistaken about the existence of a certain sort of entity than that they don’t even know what they are doing when they make statements that seem to refer to those entities.

A further problem is that, if moral statements are not truth-apt, what can be said about their meaning when they are embedded in more complex sentences that clearly don’t express a noncognitive attitude? And, relatedly, how can they take part in inferential – i.e. truth-preserving – relations? The problem is often illustrated with a modus ponens argument:

(i) If it is wrong to tell lies, it is wrong to get your little brother to tell lies.

(ii) It is wrong to tell lies.

thus

(iii) It is wrong to get your little brother to tell lies.6

It is clear that (i) does not express disapproval of lying (or the acceptance of norms that forbid lying, or the intention not to lie when that option is available, or

---

6 This is taken from Blackburn (1984: 190).
any other noncognitive state). But if all that can be said about the meaning of (ii) is that it expresses disapproval of lying, then it looks as if anyone offering this argument would be guilty of the fallacy of equivocation. This problem does not arise if we interpret the premises and conclusion as expressing beliefs. One might claim that, on this standard interpretation, while someone who utters premise (ii) asserts that lying is wrong, no such assertion is made by the antecedent of (i), so there is still a problem of equivocation. But on the standard interpretation, we can draw a distinction between propositional content and assertoric force. Both (i) and (ii) contain the propositional content that lying is wrong, but only utterances of (ii) apply assertoric force to this content (i.e. only (i) is used to assert this content). The assertoric force of (i), on the other hand, applies to the whole conditional. On the noncognitivist interpretation of (ii), however, there is no distinction to be made between the content of the utterance and the speech act that the utterance is used to perform. So noncognitivists are unable to avail themselves of the distinction drawn in the standard interpretations to escape from the charge of equivocation (Eklund 2009).

This problem – often referred to as the ‘Frege-Geach problem’7 – has received extensive discussion, with noncognitivists offering a number of ingenious replies, but it remains controversial whether any of these can succeed. I am going to proceed on the assumption that defenders of Expressivist cannot solve the Frege-Geach problem. (I have included an appendix in which I discuss a number of the attempts.)

However, even if noncognitivists can (or have) solved the embedding problem, it would still need to be shown that noncognitivism is the correct account

---

7 Peter Geach was the first to raise the problem, and he attributed the basic insight to Frege in a later paper (Geach 1960, 1965).
of how moral discourse functions. By contrast, it may seem that if Expressivist fails, Assertion would succeed by default. This is not the case, however, as there are further possibilities available for elucidating indicative moral utterances.

‘Ecumenical expressivism’, defended by Michael Ridge, has it that utterances of such sentences express both cognitive and noncognitive mental states. On one version of this theory, a normative utterance expresses both (i) “A suitable state of approval to actions insofar as they have a certain property”; and (ii) “A belief that makes suitable anaphoric reference back to that property” (Ridge 2007: 56). So, in stating that lying is wrong, for instance, I am expressing my disapproval of actions insofar as they have some property – say, failing to maximise aggregate happiness – and asserting that acts of lying have that property.

Another option is to distinguish between semantic and pragmatic content. Note that Assertion is about utterances not sentences. It is possible that moral sentences of the form “X is F” have moral properties as their truth conditions, but that typical utterances of those sentences do not. In fact, Mark Kalderon has claimed just this. In Moral Fictionalism (2005), Kalderon claims that whilst indicative moral sentences are truth-apt, standard utterances of moral sentences serve only to express noncognitive attitudes. Although Kalderon is not an error theorist, his thesis would not be incompatible with error theory, if the error theory was confined to a view about the status of moral sentences. In addition, Michael Scott and I develop a theory we call pragmatic antirealism (Scott and Brown 2011). According to this theory, although indicative moral sentences have truth-apt content sufficient for moral realism, these realist commitments are not expressed by moral utterances in most contexts. Pragmatic antirealism is also compatible with an error theory that is confined to a view about the status of moral sentences.
I am doubtful about the prospects for Kalderon’s fictionalism, but I think both ecumenical expressivism and pragmatic antirealism deserve a lot of discussion, which, unfortunately, I cannot give them here. If ecumenical expressivism turns out to be correct, then an error theory about both moral sentences and utterances will be unmotivated; and if pragmatic antirealism turns out to be correct, an error theory about moral utterances (but not moral sentences) will be unmotivated. So the conclusion I draw in this chapter is conditional on neither of these theses succeeding.

1.2.2 The Alethic Thesis

It is one thing for moral utterances to be truth-apt, another for any of them to be actually true. However, most metaethicists accept the following thesis:

\textit{Alethic Thesis} \quad \text{Some indicative moral sentences are true.}

As I said in section 1.1, most contemporary noncognitivists think that we can legitimately talk of moral truth and falsity. Error theorists, of course, deny \textit{Alethic}, but we will only be able to see why they deny it after we have gone through the other theses.

No doubt \textit{Alethic} will seem to many, if not self-evidently true, then at least unworthy of serious doubt. For the denial of \textit{Alethic} entails, amongst other things, that it is not true that either sadistic torture or paedophilia are morally wrong. That might seem like an alarming consequence, but it is worth saying at this early stage.

---

8 As Matti Eklund has shown, while Kalderon has no difficulty with the Frege-Geach problem when we consider an argument semantically, the problem reoccurs when we consider the argument pragmatically, i.e. when we consider what, by Kalderon’s lights, someone putting across the argument actually says (Eklund 2009).
that it is not a consequence of error theory that activities we consider heinous shouldn’t be prevented, nor that we ought not to be appalled by them. And nor does error theory entail that moral beliefs are not desirable things to have, even if error theory is true. In fact, I will defend the view that moral belief is desirable, even if error theory is true (section 4.6).

1.2.3 The Objectivity Thesis

Many philosophers have held that moral judgements are objective, in the sense that their truth value is independent of anyone’s feelings, beliefs or decisions. It is better if we narrow this to our actual feelings, beliefs, or decisions, as some philosophers hold that the truth value of our moral judgements is not independent of, say, how we would feel under idealized circumstances.

Objectivity Thesis: The truth value of any moral statement varies independently of anyone’s actual feelings, beliefs, or decisions.

It is worth noting why some realists prefer a stronger formulation. Russ Shafer-Landau writes that:

Moral realism is the theory that moral judgements enjoy a special sort of objectivity: such judgements, when true, are so independently of what any human being, anywhere, in any circumstances whatever, thinks of them. (Shafer-Landau 2003: 2; my emphasis).

The italicised words indicate that Shafer-Landau rejects the kind of objectivity that is compatible with moral facts being determined by the attitudes of suitably situated agents (e.g. an ideal observer (Firth 1952), or agents situated behind a veil of
ignorance (Rawls 1971)). This is largely because there is, in his view, no guarantee that any of the theories that depend on such devices will “yield prescriptions that match up with our views about what constitutes paradigmatically moral and immoral behaviour” (Shafer-Landau 2003: 41). It is presumably partly because of his rejection of such constructivist theories (but also partly because of what contemporary noncognitivists have to say about objectivity, more on which below) that he offers his own, stronger claim about the objectivity of morality:

… moral truths … obtain independently of any preferred perspective, in the sense that the moral standards that fix the moral facts are not made true by virtue of their ratification from within any given actual or hypothetical perspective. (2003: 15; emphasis original)

However, whether or not his complaints against constructivism are correct, it seems incredibly unlikely that this is part of what any ordinary speaker means to convey by their moral utterances. It is unlikely that most competent users of moral language would even understand it without first taking a course in metaethics (and perhaps not even then) (Scott and Brown 2011). For this reason, I think we ought to prefer my weak formulation of Objectivity.

Objectivity will be false if the truth conditions of moral statements are subjective. On one version of subjectivism – call it simple subjectivism – a statement such as “stealing is wrong” will be true just in case the person making the statement disapproves of stealing. So, in considering whether a moral statement is true, only the feelings of the person making the statement need be considered. That seems a plausible way to think about some evaluative sentences, relating to matters that most of us are inclined to regard as subjective anyway. For instance we are inclined to treat a statement such as “this cake is delicious” as true just in case the person making it is very much enjoying the cake, and false if the person making it
is just being polite, and doesn’t really like it at all. If two people tasting the same
cake give different verdicts about its deliciousness – Jill says it is delicious, Bill
says it is not – then most of us will be inclined to say either that they are not really
disagreeing, or that they are not deeply disagreeing. There is a sense in which we
can say they are disagreeing – they have a disagreement in attitude (Stevenson
1937). But the disagreement is not deep because Jill cannot express her
disagreement with Bill by saying that what he said was false (Schroeder 2008: 17).
This seems to be the right result when it comes to statements about deliciousness,
as there is no way to judge the truth of such statements other than by reference to
subjective experiences, and there is no sense to be made of saying that one person’s
experience of a cake as delicious is somehow the right (or wrong) one to have with
regard to that cake. “That cake is delicious” seems to be contradicted by “It is not
the case that that cake is delicious” (assuming they’re both about the same cake);
but we do not normally treat such pairs of statements as contradictories, unless they
are made by the same person.

By contrast, we do normally treat pairs of statements such as “x is wrong” and
“it’s not the case that x is wrong” – spoken by two different individuals, about the
same x – as contradictories. If simple subjectivism is true, however, we should no
more treat these statements as contradictories than we should the statements about
the cake. All our first-order moral disputes would turn out to have been entirely
pointless – because they didn’t turn on deep disagreements – just because of a
largely unnoticed fact about how moral language functions. And that seems highly
implausible.

It may be thought a bit rich for an error theorist to object to simple
subjectivism on the grounds that it would entail that all first-order moral disputes
are entirely pointless: isn’t that a consequence of error theory as well? I have two answers to give to that. My first answer is that I disagree that error theory entails that all first-order moral disputes are entirely pointless. It does entail that there is no moral fact of the matter that could settle such disputes. However, as I have already said, error theory in itself is neither revisionary nor non-revisionary. Moral discourse – including moral disputes – may be worth preserving even if error theory is true. My second answer is that, even if you think that error theory does entail that moral disputes are pointless, the error theorist at least thinks that such disputes would have a point if error theory were untrue, i.e. if moral discourse was in good standing. Simple subjectivism, on the other hand, entails that these disputes are pointless, even though, according to that theory, moral discourse is in good standing. If that were true, it would mean that most people are radically mistaken in thinking that the rules governing moral language allow a moral proposition voiced by one speaker to contradict another moral proposition made by another; and what seems implausible is that so many people could be mistaken on this fundamental point. The error theorist, on the other hand, does not deny that ordinary users of moral discourse implicitly understand its basic rules.

Another serious problem with simple subjectivism is that it makes my moral judgements objectionably dependent on my own attitudes. The subjectivist claims that moral properties exist, but the truth conditions of our moral judgements – i.e. judgements about the distribution of those properties – are our subjective reactions to events or states of affairs (whether real or imagined). That means that I will speak truly in saying “x is wrong”, just in case I really do react negatively to real or imagined cases of x. That would make us all the arbiters of what is wrong, not just in reference to our own behaviour, but in reference to the behaviour of others (as
our moral opinions are about both ourselves and others). But although I might be highly opinionated about the moral behaviour of others, I cannot be relied upon to give moral advice to others: their moral judgements, based on their subjective responses, are the ones that are appropriate for them, in the sense that they fix what is true for them, morally speaking. Surely the concept of wrongness does not include the idea that each of us is the arbiter of what is wrong, but that such judgements should never influence anyone else’s thoughts or behaviour.

It doesn’t seem plausible that we could be mistaken about this: that moral properties will turn out to be subjective after all. Jesse Prinz disagrees. He thinks that moral properties are response-dependent properties (Prinz 2007). Just as statements about the colours of surfaces have been said by some to reduce to statements about the powers of physical objects to produce experiences of colour, so Prinz has argued that the property of wrongness is just the power of an object to cause feelings of blame or guilt within me.

He doesn’t deny that many of us think of wrongness as objective, but he argues that the fact that our concepts contain mistakes shouldn’t be taken to mean that they are wholly defective. For instance, he thinks that the typical person aims to pick out objective properties when they use colour terms, but that on a more plausible metaphysics of colour, our colour judgements are response-dependent. This should not, in his view, lead us to adopt an error theory for colour terms, because we successfully use colour terms to refer, whether or not we think that colours are mind-independent properties:

Think about BLUE as a concept with a fallback plan. It would refer to a our blue experiences. If there is no such property, it refers to the property of causing blue in us [sic], regardless of whether that property has any greater integrity. (Prinz 2007: 150-1)
Prinz accounts for this by claiming that the fact that some of our concepts contain mistakes does not show that they fail to have instances. Borrowing an example from Putnam (1975), he points out that many people think that the concept GORILLA contains the feature FEROCIOUS, but the fact that these creatures are not typically ferocious does not mean that many people fail to refer to anything when they use the term ‘gorilla’. By the same token, Prinz thinks that our moral concepts can erroneously attribute objectivity to moral properties without thereby failing to refer. (What they refer to, in his view, are the causes of our moral sentiments.)

I completely agree with Prinz that the question of what is the correct metaphysics for colours is not likely to make us revise the way we use colour terms in everyday contexts. But the comparison of colour properties and moral properties is inapt here. Whether or not colours are response-dependent properties, judgements about colours are seldom thought to be subjective in the way Prinz thinks moral properties are. According to response-dependence theories about colour, an object is blue if it appears blue to a ‘normal observer’ under some specified circumstances, such as in sunlight on a clear day. Because of this, we can say that colour-blind people often make wrong – as opposed to idiosyncratic – colour judgements, and that people with good colour vision make mistakes when they make colour judgements in poor lighting conditions. This means that response-dependence theorist can account for deep disagreements about colour judgements just as successfully as realists about colour. A deep disagreement about the colour of an object will occur when at least one of two parties is mistaken about how the object would appear to a normal observer under ideal conditions. (It is really just a matter of the conventions surrounding the correct use of colour terms. We could have used colour terms in such a way that a speaker correctly calls an object blue just in case
it appears to her to be blue in her present circumstances. But that isn’t how we normally use colour terms.)

The important difference between the case of moral concepts and the case of colour concepts, is that we cannot tell whether ordinary people are committed to an objectivist or a response-dependence metaphysics of colour merely by looking at how they use colour terms, but we can tell this in the moral case. And if their moral judgements commit them to objective properties that do not exist, then an error theory is appropriate, whether or not they could start using moral terms merely to refer to the causes of their moral sentiments.

A less simple form of subjectivism is the relativist theory defended by Harman (1975). Unlike simple subjectivism, Harman’s relativist analysis of moral statements does not merely take the attitudes of the person making the statement into account, but also the attitudes of the person whose actions are being morally appraised, and the attitudes of the speaker’s audience. This allows his theory to accommodate more in the way of disagreement than simple subjectivism is able to, does not make each individual the (sole) arbiter of moral truth, and makes it possible for our moral judgements to have an influence on the thoughts and behaviour of some other people. But despite these improvements on simple subjectivism, Harman’s relativism does not allow us to accommodate disagreement of the right kind, still gives us too central a role in determining moral truth, and limits the sphere of the legitimate influence of our moral judgements far too narrowly. Let us take a brief look at Harman’s analysis of moral judgements to see what he can and what he cannot accommodate.

Firstly, Harman limits his analysis to what he calls ‘inner judgements’. These are judgements about what an agent ought to do. In calling them ‘inner judgements’
he seems to be alluding to his belief that no such judgements will be true unless the agent concerned has certain goals, desires or intentions that would be satisfied by doing the act in question (1975: 9). So, in Bernard Williams’ terms, he believes that all reasons are internal (Williams 1981). Inner judgements are contrasted with another type of moral judgement that, by Harman’s lights, imply that the person they relate to is beyond the reach of our judgements about what they ought to do, because they have such different goals, desires or intentions. The judgements that someone is evil or a savage are given as examples of this contrasting type of judgement (Harman 1975: 5).

In his analysis of inner moral judgements, Harman focuses on intentions. A statement expressing the judgement that an agent $S$ morally ought to $\varphi$ in circumstances $C$ will be true, on his account, only if $S$, the speaker, and the speaker’s audience, intend to $\varphi$ in $C$. That these three parties all intend to $\varphi$ in $C$ is explained by there being a prior agreement amongst them – and other members of their group or society – to $\varphi$ in $C$. This allows room for disagreement, as there may be disagreements about what agreements have been reached, and what such agreements actually commit the parties to. It also means that no individual is the sole arbiter of moral truth: there is moral truth only when there is agreement in intention between two or more people. And it follows that our moral judgements can have a legitimate influence of the thoughts and behaviour of other people: if I say that you ought to $\varphi$ in $C$, and we have each agreed to $\varphi$ in $C$, then this is something that you have a reason to take note of.

These are real improvements over simple subjectivism, but they are not enough to make the theory a plausible account of how moral language actually works. First of all, as Max Kölbl has pointed out, what disagreement is possible within
Harman’s account does not seem to be genuine moral disagreement (Kölbel 2004). If I utter a moral statement which you disagree with, then assuming we do in fact agree in intention, our disagreement is really about what our prior agreements were, or what our circumstances are, or something of that nature. These cannot plausibly be called moral matters. To have a genuine moral disagreement it would seem that we would have to disagree in intention. But on Harman’s account, if I attempt to address a moral proposition to someone who disagrees with me in intention, then I in fact express no proposition at all. As Kölbel notes, Harman bites the bullet on this: we cannot say, for instance, that Hitler morally ought not to have committed genocide, as he just had such different intentions than us. We can only say that he was evil. But clearly most of us judge that Hitler ought not to have committed genocide, whatever his intentions, which is good evidence that the rules of moral discourse allow such judgements to be made.

On the other matters I mentioned, although I am not the sole arbiter of moral truth, any two people can jointly perform this role. They will not decide the moral truth for everyone, but only for themselves. But that is counterintuitive enough. Moral truth just doesn’t seem to be the sort of thing we can voluntarily create, whether on our own or as a group. And although, according to Harman, my moral judgements, if true, have legitimate influence on others, they have this influence only insofar as they agree with me in intention. But it is counterintuitive to say that I can be right about what someone morally ought to do only if they really, deep down, intend to act in that way (or would do if they were not confused). Relativism does not, then, look very plausible as an account of how moral discourse actually functions. In the next chapter I consider whether we could (and should) modify our moral concepts along relativist lines, as a response to the error theory (section 3.3).
One further take on the objectivity versus subjectivity debate is worth mentioning here. Error theorists think that indicative moral sentences have objective truth values. As we shall see, in combination with a thesis to be discussed below (moral rationalism), they think that this entails the existence of dubious mind-independent moral properties. But what if we can capture the objectivity of moral discourse without bringing in mind-independent properties? This is what Blackburn has attempted to do in response to the charge that noncognitivism is objectionably subjective.

Noncognitivism can appear to be akin to subjectivism. While it gives a different account of moral semantics than the types of subjectivism we have just been considering, it would seem, according to noncognitivism, that an expression of moral attitude is sincere when and only when the speaker has that attitude, and that there is no objective standard by which we can determine which of two opposing attitudes is the correct one. So as with subjectivism, moral worth seems objectionably dependent on our own attitudes. But typically we don’t think that things are right or wrong in virtue of the way we happen to feel about them.

However, Blackburn has long argued that the noncognitivist can accommodate the intuition that the assertability of moral judgements should not change as our attitudes change, by invoking higher-order attitudes (Blackburn 1984: 217-18). He asks us to focus on the sort of person who would endorse the view that the wrongness of, say, kicking dogs varies as our sentiments vary. The way that Blackburn thinks we should interpret this view is as endorsing a certain sort of sensibility: “…one which lets information about what people feel dictate its attitude to kicking dogs”; to which Blackburn comments: “But nice people do not endorse such a sensibility”. In this picture, the ‘nice person’ has a first-order
attitude of disapproval of kicking dogs, and a second-order attitude of disapproval of changing one’s attitude to kicking dogs if one’s feelings change.

We might object that although this response may allow the noncognitivist to say how a person’s first-order moral attitudes could be mistaken, the problem just gets transferred to the second-order attitudes that Blackburn appeals to. The second-order attitudes of disapproval of changing one’s attitude to kicking dogs if one’s feelings change is just an attitude that one might or might not happen to have. It will do no good to bring in a third-order attitude that disapproved of not having the second-order attitude, because then the problem would transfer to the third-order attitude (and so on). It is questionable how effective this response is, however. It might be responded that the noncognitivist has provided an explanation of why we treat the assertability of moral judgements as independent of our actual attitudes, and that to complain that this does not leave our moral judgements objective enough to tally with what the ‘ordinary user of moral language means to say’ is to credit the ordinary user of moral language with a level of metaethical sophistication that he or she is unlikely to have.

A more effective response is just to note that because the second-order judgement is just another noncognitive judgement, it will fall prey to the Frege-Geach problem just as surely as first-order moral judgements. If we are not persuaded by any of the solutions that noncognitivists have offered for the Frege-Geach problem, we will not be ready to accept their account of Objectivity.

Realists and error theorists agree that if moral discourse is in good standing, then Assertion, Alethic, and Objectivity are all true. In uttering a sentence of the form “X is F”, where “F” is a moral predicate, the speaker asserts that X has the moral property of being F; and its being F is an objective property of X. Realists
hold that some of these sentences are true, and so say Alethic is true. Error theorists take the contrary view. Why this disagreement? Well, that turns on the issues of what moral judgements commit us to – i.e. what the objective properties turn out to be like – and whether we can accept that commitment. The nature of moral properties begins to be revealed when we consider the next thesis on our list.

1.2.4 Moral Rationalism

Objective properties are not a problem per se. Most people will think that it is true to say that some objects have some properties, irrespective of our attitudes. Moral properties will not be a problem if they are identical with, reducible to, or in some way realized by properties that we already believe in, such as natural properties. Moore’s open question argument (see note 5) was designed to show us that no attempt to identify moral properties with natural properties could succeed, and that moral properties are thus non-natural. That conclusion was too swift, however, for there is the third possibility that they don’t exist at all.

Mackie did not exploit the open question argument to get to the conclusion that moral properties would be non-natural if they existed. I think we would be wise to follow him in this regard, because the argument is open to some serious objections. The following is perhaps the most compelling. The naturalist proposes that moral property \( M \) is identical with natural property \( N \). Moore says it is always an open question whether things that are \( M \) are also \( N \). The evidence for this claim seems to be that someone competent with the terms for \( M \) and \( N \) could ask this question without displaying conceptual confusion; but if the terms for \( M \) and \( N \) were merely synonyms, they would display conceptual confusion (they would be asking, in
effect, “Is M the same thing as M?”). However, as Frege showed, a pair of terms can have the same reference, and yet competent users of those terms can be unaware of this (Frege 1892). An ancient stargazer, in wondering whether Hesperus was Phosphorus, was obviously not wondering whether Hesperus was identical to itself, but whether the star that appeared in the evening was the same star as the star that appeared in the morning. And as it turned out, although ‘Hesperus’ and ‘Phosphorus’ have different senses, they do indeed refer to the same thing. So the naturalist can claim that, in just the same way, her proposed identification of M and N need not be obvious to competent speakers.

What other reason might we have for denying that moral properties will turn out to be natural properties? The best way to argue this is to show that they cannot be natural properties because, if we say that they are, we will have to deny some very plausible metaethical claims. The first of these is the claim that moral rationalism is true (which I am about to show); the second is that motivational internalism is true, which I deal with afterwards.

Moral rationalism is the thesis that moral properties in themselves give us reasons to undertake or avoid actions that would have those properties if they were undertaken. That an action would be wrong is itself a reason not do it, regardless of our goals or desires. The contrasting view is that moral properties give us reasons to undertake or avoid actions only when we have some goal or desire that is satisfied by doing right things or avoiding wrong things.

Does the reason we (seem) to have to act morally depend on what our goals or desires are? Suppose that I think some action, say drug dealing, would be wrong. I can certainly imagine that someone who proposes to sell illegal drugs has no desire

---

9 See the second chapter of Miller (2003) for this and other objections to the open question argument.
that is satisfied by her refraining from selling drugs, or at least that any such desire she does have is outweighed by her desire to make money, or support her own addiction, or whatever. We can also suppose that this individual is fully aware of the harm that is caused by the trade in illegal drugs, so it is not the case that she would desire to avoid drug dealing if she were fully aware of these facts. If our reasons to act morally depend on our desires, then while this drug dealer may be behaving immorally, she is not failing to act on a reason she has (see Foot 1972, Brink 1984, and Railton 1986 for views of this type).

The problem with this sort of view is that, if it is true, then there can only be such things as ‘moral reasons’ in a derivative sense. If I say that I did something because it was the right thing to do, or avoided doing it because it was the wrong thing to do, then I have not fully specified the reasons behind my actions. I would have to add “and I desire to do right things/avoid wrong things”. My reasons for action are always ultimately explained by my desires. Although it may be convenient to talk of ‘moral reasons’, such talk will be analogous to the way we talk about tax reasons or legal reasons. I might say that I moved my business to Switzerland for tax reasons, but my ultimate reason will probably be that by moving my business I get to keep more of my turnover for myself, which is something I desire. I get to do this by exploiting Switzerland’s lower taxes, so I talk about ‘tax reasons’, but the reason for which I act really has nothing to do with taxes. In the same way, the denier of moral rationalism can still talk about doing things for ‘moral reasons’, but this masks the fact that the reasons that she thinks justify her actions do not, in fact, have anything to do with morality.

Is this a problem? Yes, because if a person’s reason for ‘doing the right thing’ is always that it promotes something else she cares about, then it is never the case
that she does the right thing *because* it is the right thing to do. But it seems that her motivation to do an act must at least include this motivation, if she is truly to be described as acting rightly. This is how Kant illustrates the point:

For example, it certainly conforms with duty that a shopkeeper not overcharge an inexperienced customer, and where there is a good deal of trade a prudent merchant does not overcharge but keeps a fixed general price for everyone, so that even a child can buy from him as well as everyone else. People are thus served *honestly*; but this is not nearly enough for us to believe that the merchant acted in this way from duty and basic principles of honesty: his advantage required it; it cannot be assumed here that he had, besides, an immediate inclination towards his customers, so as from love, as it were, to give no one preference over another in the matter of price. Thus the action was done neither from duty nor from immediate inclination but merely for the purposes of self-interest. (Kant 1785: 397)

There is much that might be objected to in this passage. Kant’s moral philosophy is notoriously demanding in what it takes for our actions to have moral worth: they must not to any extent be motivated by self-interest, or by love or ‘immediate inclination’, but solely by duty to the moral law. His assessment of the merchant in this passage may strike us as rather uncharitable too. Can’t he be motivated *both* by self-interest and by friendly feelings? But one thing that does seem right is that if the merchant were *solely* motivated by his self-interest, and would overcharge inexperienced customers if that suited his self-interest, then we would not say that his actions were morally praiseworthy. He has ‘done the right thing’, but he hasn’t acted in a morally praiseworthy manner, but only in a self-interested manner. So acting morally is about acting for moral reasons, not self-interested reasons.

Another way to make the point is by pointing out that we think that we have reason to regret the wrong things that we do. If our reasons for being moral were only reasons to promote our own self-interest, then we should think that our reason to feel regretful in such situations was only that our self-interest had been
undermined. But it seems a conceptual error to think that the regretfullness of a wrong action depends on the harm that that action brings to the actor, rather than the other people who are actually or potentially harmed by the action (Joyce 2001: 33).

That last point might be taken to show that I am drawing a false dichotomy between self-interested reasons and irreducible moral reasons, when there is another type of reason to consider, namely *other-interested reasons*. While self-interested reasons are those that arise out of our concern for our own well-being, other-interested reasons arise out of our concern for the well-being of others. Whilst the person acting from pure self-interest certainly seems to deserve no moral praise, the person acting out of concern for others may seem a paragon of virtue. Indeed, the idea that our sole reason to do what is right is *that it is right* (or because it is our duty), has been thought by some to be highly objectionable (Smith 1994: 75). The person who volunteers to help the needy out of compassion seems more morally admirable than the person who volunteers only because she thinks it is her sacred duty, and is not actually moved by anyone’s plight.

I don’t doubt that we often act for other-interested (i.e. altruistic) reasons. However, acting on purely other-interested reasons cannot be sufficient for acting morally, for the same reason that acting on purely self-interested reasons cannot be. Someone can be criticised for not doing the right thing regardless of whether they are concerned for the people involved. Another reason that it is not sufficient is that someone can do the wrong thing for other-interested reasons (e.g. a parent who lies to the police to protect her criminal child). I don’t even think acting on other-interested reasons is necessary for acting morally. For we surely *can* morally admire someone who acts on principle even though, in so acting, she thwarts her
other-interested concerns (e.g. a parent who turns her child in to the police, not because she thinks it is the child’s best interest, but simply because it is the right thing to do).

It seems to me that if we praise someone more when they are motivated by concern for others, rather than acting solely from a sense of duty, that is because their motivations reveal something morally admirable about their character: they are the sort of person who is moved by the concerns of others. But it is not enough that they display this character trait for us to credit them as acting rightly (for the reasons just given). They must also act out of concern to do the right thing because it is the right thing.

There is one last thing to take into account before stating our thesis, which is that moral considerations have priority over non-moral considerations – at least when we are talking about what is morally right and wrong. The fact that it is morally wrong to undertake some course of action is supposed to outweigh any other consideration, such what would be enjoyable, or profitable, in the circumstances. And if a certain course of action is the morally right thing to do in a certain set of circumstances, then that too is supposed to outweigh any other considerations in favour of alternative courses of action. So considerations about what is right or wrong are meant to cut off debate about what to do, in a way that other considerations don’t.

Not all moral considerations have an automatic priority over competing non-moral considerations. That it would be morally good to undertake a certain course of action is not a decisive consideration in favour of doing it. This has to be the case, because there are innumerable things that you could be doing at any moment that seem morally praiseworthy, and it would be impossible to do all of them. And
it is normally thought morally acceptable to devote some of one’s time to one’s self-interested pursuits, which would be impossible if one was under an obligation to do good at every moment. We may be morally criticisable if we never deliberately do good, but we cannot be obliged to do every good action that is available to us to do.

So let us take all this into account when we state our thesis:

*Moral Rationalism: If* $\phi$-*ing* is the right (wrong) thing for an agent $S$ to do in a particular circumstance, then $S$ ought, for irreducibly moral reasons, to $\phi$ (not to $\phi$) in those circumstances; and there are no alternative considerations that can outweigh the moral ought.

I’ll call this thesis *Rationalism* for short. If *Rationalism* is true then the properties of ‘being right’ and ‘being wrong’ have very special implications for practical decision making. If what we morally ought to do always outweighs what we ought to do, given alternative considerations, then morality seems inescapable, especially when we consider that these properties are supposed to be objective: existing independently of our feelings, beliefs or decisions. Foot calls this morality’s ‘binding force’ (Foot 1972); Mackie spoke of objectively good actions as having ‘to-be-pursuedness’ built into them, and wrong actions as being imbued with ‘not-to-be-doneness’ (Mackie 1977: 40).

The combination of *Rationalism* and *Objectivity* makes moral properties difficult to believe in. Mackie also brings motivational internalism and the

---

10 Although she did not think that moral reasons were in fact inescapable. She was referring only to what she took to be an implication of Kant’s ethics.
supervenience thesis to bear in his queerness argument, so I shall discuss these before getting to Mackie’s and Joyce’s arguments.

1.2.5 Motivational Internalism

We all know that people can be motivated to do that which they judge to be wrong, and that people can fail to be motivated – or, at least, effectively motivated – to do what they judge it is right to do. (Someone is effectively motivated to φ when they are motivated to φ and they actually endeavour to φ.) However the motivational internalist holds that, despite these familiar phenomena, there is a necessary connection between moral judgement and motivation.

One way of putting this is to claim that it is impossible to make moral judgements that leave one indifferent (Lenman 1999). Suppose that I have a large number of moral opinions: there is a set of actions $R$ that I think are right to do in certain specified circumstances, and a different and non-overlapping set of actions $W$ that I judge wrong to do in specified circumstances. If I am perfectly morally conscientious then, by definition, I shall always be motivated to engage in the sort of actions contained in $R$ and avoid the actions contained in $W$. In fact, we can make an even stronger claim. If I am perfectly morally conscientious, then I shall always be effectively motivated to engage in the $R$ actions and avoid the $W$ actions, when it is appropriate to do so. That is to say, if I judge it right (wrong) to φ in circumstances $C$, I will always try to (try not to) φ in $C$. If I am not perfectly conscientious I might sometimes lack the courage of my convictions, and fail to perform an action from $R$ when the circumstances are right, or yield to the temptation of performing an action from $W$. However, this need not indicate that I am indifferent to my moral opinions. For it may be that I have a different and
conflicting non-moral motivation that gets the better of me. We might say that I am somewhat motivated to do the right thing, but not effectively motivated. Alternatively we can say that if I am in some way troubled by my not acting in accordance with my moral opinions – which may manifest itself as guilt or shame – this reveals that I am not indifferent.

Let us state this thesis as:

*Motivational Internalism* I judge that I morally ought (ought not) to φ in circumstances C only if I am, *ceteris paribus*, effectively motivated to φ (not to φ) in C.

I am going to refer to this thesis as *Internalism*, for short, and to its supporters as ‘internalists’ and its opponents as ‘externalists’. As stated, this thesis only concerns judgements about what I ought or ought not to do, and not judgements about what it would be merely good to do. It seems to me that there will only be a necessary connection between judgements of the first type and motivation. It is easy to imagine someone agreeing that it would be good if she did a whole host of things – e.g. devoting all her time and effort to charity a, devoting all her time and effort to charity b, devoting all her time and effort to charity c … and so on – without feeling in the slightest bit motivated to do any of these things. I think it’s true that there is a necessary connection between thinking that it is good to φ and having a *pro-attitude* toward φ-ing; but I don’t think that ‘having a pro-attitude toward φ-ing’ equates to ‘being motivated to φ in certain circumstances’.

Restricting *Internalism* to ought-judgements also allows me to say that the necessary link is between these judgements and *effective* motivation. We fall short of being perfectly conscientious not just when we fail to be somewhat motivated by
judgements about what we ought to do, but when we fail to be effectively motivated. The advantage of being able to state the thesis in this way is that we can tell what people are effectively motivated to do by seeing what they endeavour to do, whereas things get a lot more murky if we have to talk about degrees of motivation.

*Internalism* would be trivially true if it were a thesis about perfectly conscientious agents. By definition, they cannot fail to endeavour to do what they think is right. Very few of us are so conscientious. We fail to do what we consider right and engage in actions we consider wrong, at least on some occasions. The reasons we generally cite for these lapses are moral vices: we were being selfish, or cowardly, or greedy, or spiteful etc. People sometimes cite an episode of mental illness, such as depression, to explain their lack of motivation. It is these sorts of cases that we want to capture with the ceteris paribus clause. However, if *Internalism* is true, we cannot cite indifference as the reason that we fail to be effectively motivated.

Notice that *Internalism* is a claim about my judgements about what I morally ought to do, rather than a claim about what I actually ought to do. It should not be confused with a claim discussed by Derek Parfit, which he calls ‘moral internalism’:

*Moral Internalism* We cannot have a [moral] duty to act in some way unless (if we knew the relevant facts, and deliberated rationally, we would be motivated to do this thing). (Parfit 1997: 100-102)

*Moral Internalism* is a claim about what we actually ought morally to do. It is important not to confuse these two claims. They are not equivalent, and neither
entails the other, although they are compatible (Smith (1994) would seem to endorse both of these theses).

Internalism would be false if the following situation was possible: I am highly morally opinionated, yet indifferent to my moral judgements. Suppose that when I consider whether to do something, my decision is never affected in one way or another by whether I judge the action to be right or wrong. We can also suppose that I am a complete and utter egoist. When I consider what to do, my bottom line is always what’s in it for me, and if that conflicts with my moral assessment of the situation, my egoistic principle always wins out (and suppose, for good measure, that I never feel guilty or ashamed about this). I am what is known in the literature as an amoralist, where this is used as a technical term. The question then is whether amoralists are possible.

It is certainly true that we can imagine someone saying something like, “I know that drink driving is wrong, but I do it anyway and feel no shame about it”. Perhaps psychopaths provide real-world examples of people who hold such attitudes. A standard way that internalists have of dealing with cases like these is by saying that such people are using moral terms in an ‘inverted commas’ sense (Hare 1952: 124-6). If an agent seems to make a moral judgement to which she is indifferent, she is best understood to be alluding to the moral judgements of others. When she says “drink driving is wrong”, what she really means is something like “drink driving is conventionally held to be a wrong”. The inverted commas reply gains some plausibility from the fact that we sometimes do use evaluative terms within ‘inverted commas’ (for instance, we might talk about what good girls do or

---

11 I take it that an amoral person is standardly understood to be someone who just acts without regard to moral principles, without necessarily being morally opinionated. In the philosophical sense of the term, the amoralist also acts without regard to moral principles, but in addition is morally opinionated.
don’t do, in the context of a discussion of sexist attitudes, where it is clear that by ‘good’ we mean ‘good according to the outdated attitudes under discussion’). The suggestion is then that the ‘amoralist’ always uses evaluative terms with inverted commas.

Against this, externalists can say that it fails to take the amoralist seriously enough (Brink 1986). The amoralist uses (and succeeds in using) moral vocabulary to pick out moral facts in just the same way that everyone else does, she is just different from most of us in being indifferent to what they are when it comes to deciding how to act. The externalist can concede to the internalist that there is a connection – in most of us – between our moral opinions and what we are motivated to do. It is just that this connection is contingent rather than necessary. The externalist has to deny that moral judgements are essentially practical, but can go on to offer an explanation of why most people are motivated by them. For instance Brink has claimed that moral facts are facts about human flourishing and well being (1986: 115). We will be motivated to act in accordance with our moral opinions only in so far as we are motivated by thoughts of flourishing and well being, but – fortunately – most of us are motivated by such thoughts.

I think we can dismiss Brink’s claims merely on the grounds that they conflict with Rationalism. If people were only motivated to do that which they consider right by their self- and other-interested concerns, then they are not acting for a reason which deserve moral merit. In fact this should lead us to see how Internalism follows from Rationalism.

For this claim to succeed, we shall have to make the additional claim that competent users of moral discourse implicitly know that Rationalism is true. I am quite confident about making that claim, as the two main claims I put forward when
formulating *Rationalism* seem so obviously true: (1) that moral merit is only deserved for actions done for moral reasons; and (2) that moral oughts have priority over non-moral considerations. So, an agent who believes that she morally ought to φ believes that (a) there is a moral reason for her to φ; and (b) that that reason outweighs any competing considerations there may be for not φ-ing. Now, it is certainly possible that this agent could believe both (a) and (b) yet fail to be motivated to φ. But if that were to happen, there must be some explanation for why she is not motivated, for it is irrational to fail to be motivated to do that which you think you have most reason to do. Such irrational episodes are instances of ‘weakness of will’ or akrasia. Thus, in judging that I morally ought to φ in C, I am *inter alia* judging that I will be effectively motivated to φ in C, unless I succumb to some moral vice, or I am suffering from a mental illness (or there is something else preventing me from being fully rational). Thus *Internalism* follows from *Rationalism*. Anyone who appears to make moral ought judgements to which she is indifferent is either insincere, or they are trying but failing to make a moral judgement,\(^{12}\) or they are using moral terms in an inverted-commas sense.

### 1.2.6 The Supervenience Thesis

Last, but by no means least, we have the supervenience thesis. Almost everyone accepts some form of this thesis. The basic idea can be captured in this way. If we are morally comparing a pair of situations \(A\) and \(B\), any moral differences that we attribute to them must be grounded in some non-moral difference that exists between them. And simple numerical difference won’t do. The fact that you and I

---

12 This is what Smith argues is happening (Smith 1994: 70).
are numerically different will not be enough to ground a moral difference in our circumstances.

Supervenience Thesis: moral properties (if they exist) supervene on non-moral properties, such that the subvening non-moral properties wholly determine the supervening moral properties.

The nature of the supervenience relation between moral and non-moral properties is controversial, particularly with regard to what exactly is supposed to supervene on what. This is not surprising given the broad acceptance of Supervenience coupled with the diversity of accounts of what our moral judgements commit us to, ontologically speaking. Nicholas Sturgeon (2009) has argued that given the different ontological commitments of rival metaethical theses, there is in fact no one supervenience thesis that is generally agreed upon. I have stated Supervenience as blandly as possible to avoid such subtleties.

That completes my discussion of the six metaethical theses. In the next section, we shall see how these theses can be used to support error theory.

1.3 Arguments for error theory

If moral discourse is in good standing, then the following claims are true: (1) in uttering an indicative moral sentence of the form X is F (where ‘F’ is a moral predicate), the speaker asserts that X has the moral property of being F (Assertion); (2) some indicative moral sentences are true (Alethic); (3) such sentences have objective truth values (i.e. truth values that vary independently of anyone’s actual beliefs, feelings, or decisions) (Objectivity); (4) the moral properties of actions give us irreducibly moral reasons to act (or refrain from acting), and some of these
properties – namely the properties of being the right/wrong thing to do – provide reasons that outweigh any competing, non-moral considerations (*Rationalism*); expressions of moral ought judgements are sincere (or well-formed) only when the person making them is *ceteris paribus* motivated to act in accordance with them (*Internalism*); and moral properties supervene on non-moral properties, such that the latter wholly determine the former (*Supervenience*). In the rest of this chapter I will show how we can take what we have learned about the above theses to develop arguments for the error theory.

The best known arguments for error theory come from J.L. Mackie’s *Ethics: Inventing Right and Wrong*.\(^\text{13}\) They are the *argument from relativity* and the *argument from queerness*. In that book they are presented as independent arguments (with the relativity argument being discussed first). However, in my view, the relativity argument is best understood in the light of the queerness argument, and is not, in fact, independent of it. So I’ll start with a discussion of the queerness argument (section 1.3.1). The gist of that argument is that vindicating moral discourse requires us to posit non-natural properties, but that moral phenomena – such as our propensity to make and be affected by moral judgments – can be adequately explained without positing such properties. The second explanation is to be preferred, in part, because its metaphysics is more credible. Once I have laid out this argument in outline, I will turn to look in more detail at how error theorists have sought to explain why we don’t need to invoke moral properties to explain moral phenomena, and it is in this context that I will look at the relativity argument (1.3.2). I will show why that argument is inadequate, before going on to consider a more promising evolutionary debunking argument (1.3.3).

\(^\text{13}\) Although they are sketched out in a much earlier article (Mackie 1946).
After discussing Mackie’s arguments, I will go on to discuss Richard Joyce’s argument for error theory (1.3.4). Mackie’s arguments may not be effective against the sort of philosophers who holds that our moral obligations can be discovered rationally, but denies that this requires us to posit non-natural moral properties. I will show how Joyce’s argument is effective against this sort of opponent.

1.3.1 The argument from queerness

Mackie argues that moral discourse commits us to objectively prescriptive values (which he usually calls ‘objective values’ for short). Such entities would, according to Mackie, be ‘queer’ if they did exist, by which he just seems to mean that they would not fit into the naturalistic picture of the world that he and many other philosophers and scientists have hoped to adhere to:

If there were objective values, then they would be entities or qualities or relations of a very strange sort, utterly different from anything else in the universe. Correspondingly, if we were aware of them, it would have to be by some special faculty of moral perception or intuition, utterly different from our ordinary ways of knowing everything else. (Mackie 1977: 38)

So objective values raise both metaphysical and epistemological worries.

What does Mackie mean by ‘objectively prescriptive values’? Well, we can attribute to Mackie an acceptance of Objectivity, Rationalism, and Internalism. From Objectivity, he thinks that a moral statement can only be true if there is some property, independent of the mental states of person making the statement (or any other person), on which the truth of the statement depends. From Rationalism he thinks that this property in some way provides reasons for action (or refraining from action), but not in a way that depends on the mental states of any person (i.e.
their antecedent desires or beliefs). And from Internalism he thinks that anyone who accepts the truth of the statement will be motivated to act in accordance with it. So moral values are objective because Objectivity is true, and prescriptive because Rationalism and Internalism are true.

One might think that Rationalism alone is enough to cover the prescriptivity of objective values, and some later error theorists, having doubts about the truth of the thesis, have questioned the need to bring in Internalism in explaining the strong sense of prescriptivity that the error theory requires (Garner 1990; Joyce 2001: 17-29). However, it seems that Mackie is trying to give a sense of why moral values would be prescriptive in the strong sense that error theorists require. When we accept a moral statement, such as “It is wrong to steal”, it is as if we accept a directive not to steal. (Although I wouldn’t say that telling someone that it’s wrong to steal is necessarily to command them not to do so – when two people are having disagreement over a particular moral issue, it does not seem as if they are issuing each other different commands – although in certain contexts the command may be implied.) Given Objectivity, we take such directives to be in place independently of our actual desires or interests, but why, in that case, should we care about them? We might have an explanation if such commands not only direct us in how to act, but also motivate us to act in that way too, and that seems to be what Mackie has in mind when he writes:

Plato’s Forms give a dramatic picture of what objective values would have to be. The Form of the Good is such that knowledge of it provides the knower with both a direction and an overriding motive; something’s being good both tells the person who knows this to pursue it and makes him pursue it. (Mackie 1977: 40)
I think that this is an interesting attempt to elucidate the ‘binding force’ of morality. If objective prescriptions don’t have the power to move us to act, then why should anyone care about acting in accordance with them, unless they are motivated by some other consideration, such as self-interest? But, as I said above when discussing Rationalism, a person does not act in a morally commendable manner if they are solely motivated by non-moral considerations. So I agree with Mackie that the command and the motivation for complying with the command have to have the same source.

Having seen what objectively prescriptive values would be like if they existed, we can now ask: (1) why don’t objective values fit into a naturalistic world view?; and (2) why should this make us deny their existence?

The answer to (1) is that for objective values to fit into a naturalistic world view, they would have to be natural properties. But although natural properties are suitably objective, it is difficult to believe that features of the empirical world provide us with reasons for acting, or motivate us to act, independently of our mental states. How things are in the external, empirical world tells us nothing about how we ought to behave. So defenders of the view that moral properties either are, or are constituted by, natural properties, deny that both Rationalism and Internalism are true (e.g. Boyd 1988, Brink 1984). But even though the case for Internalism remains very controversial, the case against Rationalism seems incredible. For the denial of Rationalism implies that one has a reason to comply with moral demands only if one has self-interested (or other-interested) desires that are satisfied by so complying, that moral demands do not have priority over competing considerations, and that one can be criticised for acting immorally, and that we have reason to regret the wrong that we do only when we have thwarted our own interests.
So the basic argument against ethical naturalism can be stated thus: To act morally – as opposed to acting merely in accordance with morality – is to act for the reasons provided by the moral features of one’s circumstances, and not because acting in that way happens to satisfy one of your contingent desires. So if moral facts are natural facts, then some natural facts can give you reasons to act irrespective of your desires, which is absurd. This is the grain of truth in Blackburn’s claim that we have no conception of a fact that by itself determines practical issues (Blackburn 1998: 70). We have no conception of a natural fact that can do this. So moral facts cannot be natural facts – if they exist, they are non-natural.

The answer to (2) is a more delicate matter. One thing we should be clear about from the start is just what sort of argument Mackie is offering. Walter Sinnott-Armstrong has objected that Mackie moves illicitly from the claim that (a) moral values are utterly different from anything else in the universe to the conclusion that (b) there are no moral values (Sinnott-Armstrong 2006: 47). As he rightly points out, (a) does not imply (b). But Sinnott-Armstrong is attacking a straw man. The queerness argument is not a deductive argument, but an argument to the best explanation. What we want to explain are certain related phenomena: our propensity to make and voice moral judgements, and our claims to know that some of these judgements are true; the apparent objectivity of these judgements; and their behavioural and emotional influences over us. Mackie offer us two possible explanations of these phenomena (once we have dismissed noncognitivism, subjectivism, and ethical naturalism):
(i) Besides the sort of properties that we know about through sensory experience or scientific inference (natural properties), there are also a number of non-natural moral properties that belong to such things as actions, characters, and institutions, which, by definition, we cannot know about through ordinary sensory experience or scientific inference. We detect these properties by some *sui generis* epistemic faculty – let’s call it ‘moral intuition’ – and knowledge of their distribution (i.e. of which actions are right or wrong or permissible) provides both an authoritative guide to how to behave, and, to the extent that we are rational, motivates us to behave in that way.

(ii) We have moral beliefs that entail the existence of non-natural properties that would require a non-sensory epistemic faculty if we were to have knowledge of them. However, we can adequately explain why we have these beliefs, and why we are behaviourally and emotionally influenced by them in the way we are, without supposing that non-natural properties exist.

We haven’t yet seen how error theorists have attempted to explain why we should have moral beliefs in the absence of a moral reality, but assume for the moment that they can give an adequate explanation of this. Then explanation (ii) looks superior to explanation (i) in three important ways.

Firstly it is more parsimonious. Explanation (ii) only requires us to posit entities of a type that we have independent reasons to believe in, while explanation (i) requires us to posit an *ad hoc* type of entity that is brought in only to explain moral phenomena.
Secondly, explanation (i) is incomplete because the faculty of ‘moral intuition’ is left unexplained. In addition, knowledge claims based on intuition are notoriously problematic. In the quote at the start of this section, Mackie talks about a faculty of moral ‘perception or intuition’. If we are thinking of the faculty as being literally one of perception, then we must have in mind some kind of causal mechanism. But non-natural properties would seem to be excluded from entering into causal relations, merely in virtue of their not being physical. If the possibility of non-natural properties entering into causal relations is admitted, then we are owed a demonstration that we actually have a mechanism for detecting non-natural properties, or at least some kind of indication of what such a mechanism might be like.

We are more likely to be told that we intuit moral facts in much the same way that we intuit mathematical facts (Ross 1930: 32), that when we carefully consider certain moral propositions, they just seem self-evidently true (Shafer-Landau 2003, chapter 11). The problem with this suggestion is that the most plausible realist epistemology for mathematical facts is *a posteriori* rather than *a priori* (as I discuss below). So tying up the fortunes of objective ethical facts with objective mathematical facts will only be persuasive if there is available an *a posteriori* epistemology for ethical facts as well. But that is just what non-naturalists are seeking to avoid by appeals to self-evidence.

Thirdly, is the problem of supervenience. As I said above, almost everyone accepts something like:

*Supervenience Thesis:*  
moral properties (if they exist) supervene on non-moral properties, such that the subvening non-moral properties wholly determine the supervening moral properties.
Ethical naturalists have an easy enough time accounting for supervenience, as the moral facts are just identical to, or somehow constituted by, non-moral facts. The error theorist doesn’t have to account for anything: there are no moral properties to supervene on anything else. But what about the non-naturalist realist? Mackie asks:

What is the connection between the natural fact that an action is a piece of deliberate cruelty – say, causing pain just for fun – and the moral fact that it is wrong? It cannot be an entailment, a logical or semantic necessity. Yet it is not merely the fact that the two features occur together. The wrongness must somehow be ‘consequential’ or ‘supervenient’; it is wrong because it is a piece of deliberate cruelty. But just what in the world is signified by this ‘because’? (Mackie 1977: 41; emphasis original)

Another way of pressing the point is to ask why it is not possible, if moral and non-moral facts are ‘distinct existences’, for two instances of causing pain for fun that are non-morally identical to differ in their moral qualities. Non-naturalists are, of course, free to postulate a brute, metaphysically necessary connection between certain non-moral facts and certain moral facts: it just happens to be the case that causing pain for fun is wrong in every possible world (Shafer-Landau 2003: 85). Such a move is distinctly unsatisfying, however.

Nevertheless, it may be argued that we have a very strong reason to accept (i) over (ii) because accepting (ii) would seem to require us to reject a large number of beliefs – that certain actions are right and others wrong, that certain character traits are morally admirable and others not etc. – that have both strong intuitive

---

14 This does seem to entail the possibility of facts that are both moral and non-moral. To avoid this, the ethical-naturalist version of Supervenience might be better phrased “moral properties (if they exist) supervene on properties that can be picked out with non-moral predicates, such that these subvening properties wholly determine the supervening moral properties”.

15 And may be insufficient, if, as Blackburn has argued, the real task is to explain why it is analytically necessary that there cannot be a difference in a moral properties without a difference in underlying non-moral properties (Blackburn 1993a).
appeal and seem to play an important (some would say vital) part in our social lives. Of course, very few people – and certainly not error theorists – would deny that many of our moral beliefs have strong intuitive appeal, if all that means is that we are very much inclined to say that certain acts are morally right or wrong. However, there is a way that we might appeal to the strength of our moral beliefs, in support of (i). We might claim that the beliefs we are justified in having are those that would survive the process of reflective equilibrium, and, given how confident we are of some of our moral beliefs, we can safely predict that they would survive the process. But whatever we think of such a coherentist approach to epistemology, we surely cannot claim to have justified moral beliefs on the basis of the prediction that they would survive the process of reflective equilibrium. We have to actually go through the process, and that involves considering the metaphysical implications of our moral beliefs (Joyce 2001: 170). And besides, our concern is not whether or not anyone’s moral beliefs are internally justified, but with whether or not any of them are true.

There are, of course, various other ways that we might avoid the wholesale rejection of our moral beliefs, such as, for example, by defending a noncognitivist, a subjectivist or a naturalist metaethics. But at this stage of the argument, these options have been foreclosed. Our choice is between (i) and (ii), so we must ask ourselves whether to accept (i), with its mysterious entities, and mysterious epistemic faculty; or (ii), which adequately explains the phenomena and introduces no entities that we don’t have independent reasons to believe in.

With regard to the second point – that moral beliefs play either an important or vital role in our social lives – all that can be said is that at most this gives us a pragmatic reason to retain our moral beliefs. That they play this vital role could
only provide evidential support to our moral beliefs if we thought that they were just beliefs about rules that played an important or vital role in social intercourse. But that conception of moral beliefs is not on the table.

Admittedly, the idea that morality requires mysterious properties and a mysterious epistemology may not, for many people, be a mark against it. After all, many people believe in gods or other supernatural phenomena, and many – perhaps most – people think that morality requires the existence of a god. Error theory isn’t likely to be persuasive for such people. However, it might still be objected that there are other, non-supernatural objects and properties that we all think exist, but which can be detected neither by the senses nor by scientific instruments.

What about, for instance, mathematical and logical objects and properties? We seem to know things about prime numbers and validity, but we have never seen five or seven, but only symbolic representations of them, and we do not detect the validity of an argument from the way that it smells. However, there is a way of defending the existence of such properties that is broadly naturalistic.

The argument goes that we should believe in all the entities that are posited by our best theories of the world. Our best theories are those provided by the natural sciences, and these not only posit natural objects, but mathematical objects, as the theories cannot be stated without quantifying over mathematical objects (such as numbers or sets). Our best theories are confirmed by observation and experimentation, but as mathematical terms are necessary to state those theories, successful scientific practice brings with it a confirmation of the existence of mathematical entities. This argument – often known as the ‘Quine-Putnam indispensability argument’\textsuperscript{16} – is controversial and I am not offering my

\textsuperscript{16} Its \textit{locus classicus} is Putnam (1979).
endorsement of it here. However, it does have an obvious appeal. Few of us doubt that we get some knowledge via empirical observation and experimentation, and the indispensability argument provides a way of showing how we can gain knowledge of the unobservable via knowledge of the observable.

The indispensability argument obviously has limited application, however. Even if it is sound, it can only be used to demonstrate the existence of those entities that are ineliminably posited by our best theories. And it does not seem that any of our scientific theories posit the existence of objectively prescriptive values. Of course, the failure of the indispensability argument to vindicate objective values hardly shows that they cannot be vindicated by some other means. But this discussion points the way to an explanation of how we can reject them.

When we start to examine morality at the metaethical level, there are various phenomena to be explained. There is our propensity to make and voice moral judgements; there are our claims to have moral knowledge or insight; and there are the effects that our judgements have on ourselves and others – the way they influence our behaviour and emotions. The non-naturalist seeks to vindicate at least some of this behaviour and some of these responses, but in order to do so she has to posit a type of property that is unobservable and does not feature as a posit in any scientific theory. The error theorist, on the other hand, seeks to show how moral discourse, moral behaviour, and moral emotions are explicable from within a naturalistic framework. One likely response is that this is objectionably scientistic;

17 Hartry Field, for instance, has argued that mathematics is, in principle if not in practice, dispensable from science (Field 1980).
18 David Enoch has provided an indispensability argument for the existence of objective normative properties (Enoch 2007). The key premise is that objective normative properties are indispensable for the purpose of deliberating over what to do, as that practice could not survive the realization that we have no more reason to act in one way rather than other (and the process of deliberation is itself an indispensable activity). We can find similar arguments in many of Nagel’s writings (e.g. his 1980 and 2003). The error theorist, of course, does not deny that there are norms to which we are sensitive in deliberation, but only that they are objectively prescriptive.
but the error theorist does not have to commit to the view that the only route to knowledge is via the methods of science, but only that where a credible naturalistic account of some phenomenon is available, it is to be preferred to an account that posits entities that are neither observable, nor inferable from observation.

Let us now take a look at two debunking arguments.

1.3.2 The argument from relativity

Mackie presents his relativity argument as being independent from his queerness argument, but I think it is best considered as a supplementary debunking argument, for reasons I hope to make clear. He writes that the relativity argument

… has as its premiss the well-known variation in moral codes from one society to another and from one period to another, and also the differences in moral beliefs between different groups and classes within a complex community. (Mackie 1977: 36)

The mere fact that communities (or groups and classes within communities) disagree on moral matters, and that moral opinions vary over time, does not by itself suffice to show that there are no objective moral truths. For instance, the fact that there is disagreement about history, biology or cosmology (Mackie’s examples) does not show that there are no objective facts in these areas. But, Mackie thinks, while disagreements of the latter type can be explained by “speculative inferences or explanatory hypotheses based on inadequate evidence” (36), moral disagreements are unlikely to be susceptible to this sort of explanation. Rather,

Disagreement about moral codes seems to reflect people’s adherence to and participation in different ways of life. The causal connection seems
to be mainly that way round: it is that people approve of monogamy because they participate in a monogamous way of life rather than that they participate in a monogamous way of life because they approve of monogamy. (36)

From these considerations, Mackie concludes that:

… actual variations in … moral codes are more readily explained by the hypothesis that they reflect ways of life than by the hypothesis that they express perceptions, most of them seriously inadequate and badly distorted, of objective values. (37)

So we have a ready explanation of why there is disagreement in paradigm empirical fields of inquiry: disagreement persists so long as there is inadequate data to rule out competing hypotheses. Of course, there’s more to it than that. For instance, the ability to take great leaps forward in scientific theorizing is rare, and scientific disagreements may persist simply until a Newton or an Einstein comes along who is able to take that leap. But the point is that we have a good deal of understanding of why disagreements occur within science, and of how they can be resolved.

When we compare paradigm empirical domains of inquiry with moral inquiry, Mackie says that disagreement in the moral domain is better explained by the ‘ways of life’ hypothesis, than by the hypothesis that different individuals are better or worse perceivers of objective values. This is where it becomes clear that the relativity argument is best considered as supplementary to the queerness argument, rather than being an independent argument. For it is with the queerness argument that Mackie shows that moral facts, if they existed, would be non-natural objects, and thus that they would require some special epistemic faculty – moral perception or intuition – to detect them. And this would mean that moral inquiry was a kind of empirical inquiry. It is just that, unlike with other empirical domains where all our
raw data is in the form of ordinary sensory information, in the moral domain we have to draw on additional information we receive via moral perceptions or intuition.

This makes it interesting to compare the scientific and the moral domains of inquiry, but it also points to a weakness with the relativity argument. It is certainly true that one of the best predictors of what any individual’s moral beliefs will be is the moral attitudes dominant in her culture or community. But the same is true when we consider almost any individual’s beliefs about biology or cosmology: few of us have the resources or expertise to investigate complex scientific matters. And although we may justifiably believe that specialists in these areas have found methods of biological and cosmological inquiry that lead us to closer and closer approximations of the truth in these areas, for many centuries the opinions that nearly everyone held on these matters – or the opinions that it was safe to be seen to hold – were constrained by what was acceptable to the dominant culture, especially its religious authorities. Just as scientific inquiry gradually freed itself from the clutches of religious dogma, so it might be hoped that moral inquiry will rise above both religious and parochial customs. This seems to be the view of Smith, who writes:

…where entrenched disagreements [on moral matters] seem utterly intractable we can often explain why this is the case in ways that make them seem less threatening to the idea of convergence in the opinions of fully rational creatures. For example, one or the other parties to the disagreement all too often forms their beliefs in response to the directives of a religious authority rather than as a result of the exercise of their own free thought in concert with their fellows. But beliefs formed exclusively in this way have dubious rational credentials … The fact that disagreement persists for this sort of reason thus casts no doubt on the possibility of an agreement if we were to engage in free and rational debate. (Smith 1994: 188-9)
Thus it may be true that actual divergences in moral opinions between different times and cultures is best explained by the ‘ways of life’ hypothesis, but that there is grounds for optimism for greater convergence if moral inquiry is allowed the freedom from cultural pressures that the natural sciences have to a great extent achieved. Parfit has said something similar. He claims that “Belief in God, or in many gods, prevented the free development of Ethics” (Parfit 1984: 454). We are now able to freely and openly consider moral questions from a secular standpoint, but such inquiries are in their infancy, and this makes Parfit optimistic for convergence:

… Non-religious Ethics is at an early stage. We cannot yet predict whether, as in Mathematics, we will all reach agreement. Since we cannot know how Ethics will develop, it is not irrational to have high hopes. (1984: 454)

T.M. Scanlon has similar things to say in response to the relativity argument (Scanlon 1998: 356). We might be sceptical about the likelihood of future convergence, but the relativity argument does nothing to discredit them.

However, just as diversity in moral opinion does not support error theory, neither will future convergence (if it occurs) necessarily undermine it. If there is such convergence, it might be better explained by different communities becoming more and more alike in their subjective values and lifestyles, rather than their alighting on the objective moral truth. If there is greater convergence in moral opinion in the future, we will need to ask, are the moral opinions of different cultures more similar than they were in the past because their lifestyles are more similar, or is it the other way around? We will not be able to answer that question merely by pointing to the existence of convergence in values. So neither actual
diversity nor possible future convergence can tell us anything in themselves about the objectivity of value.

1.3.3 An Evolutionary Debunking Argument

I think an evolutionary debunking argument is more likely to be successful. Such arguments take as a premise that our propensity to make and react to moral judgements has an evolutionary origin, as that propensity contributed to the reproductive success of our ancestors. They then aim to show that such origins give us strong reason to doubt that the moral judgements we make are responsive to an objective moral reality.

Theories about the evolutionary origins of psychological traits are notoriously controversial, which might make us wary of putting very much weight on an evolutionary debunking argument. Even if we had very good reasons to think that much of our psychology is innate, the details of how particular aspects of it developed may forever remain a matter of speculation. However, I intend to argue in a way that removes this difficulty. For what I aim to show is that it is the non-naturalist realist, rather than the error theorist, who is obliged to give an evolutionary account of our propensity to make and react to moral judgements, but that any account they do give will be less plausible than the one the error theorist can offer. I don’t have to claim that the evolutionary account I offer is true, just that it is more plausible than the one that the non-naturalist will be obliged to give.

I will proceed by first outlining a debunking evolutionary account. Then I will say why the non-naturalist is obliged to give their own account of the evolution of

---
19 Steven Pinker provides an overview of the controversies between evolutionary psychologists and their critics in The Blank Slate (2002).
morality, before showing that any account that they can give will be less plausible than that of the error theorist.

It will be worthwhile briefly discussing a couple of issues before getting to the debunking argument. The first is that it clearly does not follow from the fact that a propensity to make and react to a certain type of judgement has evolved that those judgements are not responsive to an objective reality. We have evolved the capacity to make spatial judgements, and it is obvious that if the spatial judgements of our ancestors – necessary for, amongst other things, feeding, mating and avoiding predation – had not been responsive to the actual location of objects in their environment they would not have survived long enough to be our ancestors. However, when it comes to moral judgements, the case can be made that their contribution to human survival has nothing to do with their truth.

The second issue is that it is plausible that our propensity to make moral judgements, and the actual moral judgements we end up making, depends on the environment we grow up in. We might not have developed the capacity at all if we did not receive a moral education. This may be taken to suggest that the capacity to make moral judgements is wholly learnt and not at all innate. But for the evolutionary debunking argument to work we do not need to claim that we would have developed the capacity for moral judgement regardless of our childhood environment, or that we inherit a particular set of innate moral judgements. The possibility of innate moral principles isn’t relevant here. What proponents of the evolutionary debunking argument claim, rather, is that there are certain psychological tendencies, necessary for the development of moral judgement, that are inherited. In a way, that has to be true. After all, our tendency to form beliefs is necessary for the development of moral judgement. For the tendencies in question
to be interesting, they must be aspects of our psychology that, when absent, prevent an agent from being able to develop the ability to make moral judgements, but do not prevent her from making other types of judgement.

What might such tendencies be? To answer that we need to consider the psychological difference between someone with the capacity to make moral judgements and someone who lacks it. Being moral involves sometimes making sacrifices for the benefit of others – be it of your time, your labour, your money or whatever – and sometimes demanding that others make sacrifices for you or some third party. It looks as though an entirely self-interested person could calculate that it would sometimes be beneficial to make such sacrifices if she hoped to gain some advantage by doing so, and she would certainly have an interest in making demands of others. What is distinctive about the moral person, however, is that she thinks of certain demands on her time and resources, or certain sacrifices of pleasure, as being non-optional. They are demands that she must abide by whether or not it suits her to. In addition she thinks that if she is not compliant with these demands she not only deserves to be blamed or shunned, but she endorses the blaming or shunning behaviour of others, and may even at times, out of guilt or shame, solicit these reactions from others. And she will think of guilt or shame as appropriate internal responses to non-compliance with these non-optional demands. The person without the tendency to make such judgements, by contrast, might accept blaming and shunning behaviour as an occupational hazard, but certainly would never want to be blamed or shunned. In addition, she would have no tendency to feel guilt or shame about any of her behaviour.

So the sort of psychological tendencies under discussion are the tendency to think of certain demands made of us to be non-optional, and to experience blaming
and shunning behaviour in response to transgressions as appropriate or ‘called for’, and likewise for emotional reactions of guilt and shame. Why think these tendencies are inherited genetically?

The proposal is that these psychological tendencies contributed to the reproductive success of our ancestors. An organism can increase its chances of leaving descendents by making sacrifices of its resources that benefit other organisms. Although many organisms that could not possibly be said to make moral judgements do seem to have evolved the propensity to make such sacrifices, the suggestion made by the debunking evolutionary argument is that humans developed the psychological tendencies underlying moral judgement in order to facilitate self-sacrificial behaviour that is beneficial from the point of view of reproductive success. Traits underlying moral altruism have been selected for, as they get us to behave in a biologically altruistic way, and biological altruism has survival value.²⁰

Many organisms display biological altruism, which is the disposition, under certain circumstances, for one organism to make sacrifices that enhance the survival prospects of another organism, or that enhance the survival prospects of the group to which it belongs. The social insects, comprising mostly of sterile workers working tirelessly for the good of the hive or nest, leaving no offspring of their own, provide one of the most impressive and extreme examples of this phenomenon. The explanation for this phenomenon is that the sterile workers, though not reproducing themselves, nevertheless help portions of their genome to survive into future generations by assisting in the nurture of their siblings.²¹

---

²⁰ I take the term ‘biological altruism’ from Philip Kitcher (2005), but I prefer ‘moral altruism’ to his ‘psychological altruism’.
²¹ Richard Dawkins has provided the classic popular account of this phenomenon in The Selfish Gene (1976).
Biologically altruistic acts are not always performed for the benefit of close kin. It can make sense, from an evolutionary point of view, for one organism to make sacrifices that benefit another, unrelated organism, if there is a good chance that the one making the sacrifice will be rewarded in some way. It may make sense for me to share my food with you, when you are hungry, if you are prepared to make a similar sacrifice in the future. Such behaviour is sometimes called ‘reciprocal altruism’.

Organisms are biologically altruistic when they make sacrifices that benefit other organisms. Moral altruism also involves sacrifice on behalf of others, but one is morally altruistic only when such sacrifices are morally laudable. The evolutionary debunking argument proceeds by arguing that, in our species, the benefits of biological altruism have shaped us into moral altruists. Certain types of sacrificial behaviour on an organism’s part increase the chances that parts of its genome will be reproduced, but how can its genes get it to behave in this way? Michael Ruse considers three options (Ruse 1986). Firstly, like the social insects, we could have been hardwired, like ants are, to work for the good of the nest:

There are great biological advantages to this kind of functioning: it eliminates the need for learning [and] it cuts down on the mistakes … Unfortunately, however, this is all bought at the expense of any form of flexibility. If circumstances change, individual ants cannot respond. This does not matter so much in the case of ants, since (biologically speaking) they are cheap to produce. Regretfully, humans require significant biological investment, and so … the production of [biological altruism] through innate, unalterable forces, poses too much of a risk. (98)

At the other extreme, we might have evolved superbrains “rationally calculating at each point if a certain course of action is in our best interests” (98). The problem here is that such a superbrain would be enormously costly – large brains require a
lot of energy – and probably inefficient: “By the time I have decided whether or not to save the child from the speeding bus, the dreadful event has occurred” (98-9).

Luckily, there is a third way. In order to get us to be biologically altruistic, nature has given us the capacity to be morally altruistic (or, better, to conceive of ourselves as such):

To make us cooperate for our biological ends, evolution has filled us full of thoughts about right and wrong, the need to help our fellows, and so forth. We are obviously not totally selfless. Indeed, thanks to the struggle for reproduction, our normal disposition is to look after ourselves. However, it is in our biological interests to cooperate. Thus we have evolved innate mental dispositions (what the sociobiologists Charles Lumsden and Edward O. Wilson call “epigenetic rules”) inclining us to cooperate, in the name of this thing which we call morality … 22 We have no choice about the morality of which we are aware. But unlike ants, we can certainly choose whether or not to obey the dictates of our conscience. We are not blindly locked into our courses of action like robots. We are inclined to behave morally but not predestined to such a policy. (99)

(When Ruse says that we have no choice about the morality of which we are aware, he just means that we conceive of moral rules as objective. We are not literally aware of any moral rules because there are none – or so Ruse contends.)

Suppose we did think that this, or something like it, was the most likely account of the origin of our capacity for moral judgement. 23 Then there is an easy move from this account to the error theory. If this capacity was selected for because it pushes us to behave in beneficially sacrificial ways – beneficial from our genome’s point of view that is – then we have an explanation of why we are inclined to make moral judgements that does not depend on any of those

23 I choose to present Ruse’s picture of the evolutionary origins of morality because it is clear and succinct. A more technical argument to the same conclusion appears in Ruse and Wilson (1986); Joyce argues along similar lines in his (2006).
judgements being true. And if none of the judgements need to be true for this explanation to work, then we have no reason to posit the existence of moral properties – the things that moral judgements would be true in virtue of, if they were true – in our explanation of why we have a moral sensibility. In this situation, the principle of parsimony tells us that we should not posit them.

However, it might reasonably be objected that accounts of the evolution of moral judgement, such as the one I have just sketched, are far too speculative to gain our assent. Isn’t the non-naturalist free to offer her own, alternative account of the evolution of the moral sensibility that is more congenial to her? We cannot deny them the opportunity. I am not aware of any such account, but in a paper in which she develops her own evolutionary debunking argument Sharon Street suggests that the non-naturalist might propose that “our ability to recognize evaluative truths … conferred on us certain advantages that helped us flourish and reproduce” (Street 2006: 126). Street objects that this scientific hypothesis fares poorly when considered against debunking hypotheses. I think that’s right, and the point will be even more forceful if we can show (as I intend to) that the non-naturalist is obliged to give some sort of evolutionary account of our capacity to make and react to moral judgements.

Why are they obliged to give such an account? It is because the epistemology of non-naturalism requires either us to have a faculty of either moral perception or intuition (the former being analogous to sensory perception, the latter being analogous to mathematical intuition), and the non-naturalist’s account will be incomplete if it lacks an explanation of how we have come to have these faculties. The non-naturalist would not have to give an evolutionary account if they could show that our ability to detect moral properties was an acquired ability, made
possible by other, non-controversial faculties, such as our ordinary sensory 
faculties or our general ability to reason. But as I shall show, neither moral 
perception nor intuition can be supported in these ways. Of course, it is always 
possible that the non-naturalist could argue that the epistemic faculties necessary 
for moral knowledge are God-given, but I take it that few metaethicists would want 
to argue in this way.

If the non-naturalist takes our moral knowledge to come via moral perception, 
and she believes in the Darwinian theory of evolution, then she will surely believe 
that our faculty of moral perception has, like our other sensory faculties, evolved. 
Where else could it have come from? It cannot be an acquired perceptual ability, 
like the ability a painter might have to make finer discriminations between hues 
than the rest of us, because moral perception would be sui generis, not a 
development of any of our ordinary sensory faculties. Unless it is God-given, we 
have to assume it has evolved. Her account of how it had evolved would have to 
begin by explaining just what it is that we perceive when we perceive that 
something is or would be wrong. It would also need to locate the sense organ 
involved in moral perception, and explain how ‘wrongness’ and ‘rightness’ 
impinge on this organ. Most critically, sensory stimulation must surely involve 
causal contact, and it is difficult to imagine how we could be in causal contact with 
anything non-physical. But, of course, it is central to the non-naturalist’s theory 
that moral properties are non-physical. The chief difficulty in accounting for the 
evolutionary development of moral perception is really just that of imagining what 
moral perception could be.

On the other hand there is the perhaps more plausible suggestion that our 
moral judgements are built on a foundation of self-evident moral truths. Here it
might be thought there is more scope for denying a part for evolution in the development of our capacity for moral judgement. For it might be claimed that it is just our ordinary reasoning capacities – such as those that underlie our ability to do addition, assess probabilities, construct a modus ponens argument etc. – that allow us to grasp self-evident moral truths. It might be admitted that our ordinary reasoning capacities have evolved, but the claim might then be made that it was only after we had become sufficiently sophisticated that we began to grasp self-evident moral truths. That is to say that our ability to grasp moral truths is a side-effect of our general capacity to reason. Analogously, although we have brains capable of grasping complex scientific theories, the mental capacities that underlie this ability did not evolve for that purpose.

But I don’t think this will do. Some metaethical theorists may be entitled to claim, consistently with their preferred theory, that our ability to reason morally is a side-effect of our more general capacity to reason. Most obviously ethical naturalists, relativists and subjectivists. But I don’t see that non-naturalists can make this claim. The explanation for this requires us to think about what it is to be rational, from the non-naturalist’s perspective. Many philosophers have thought, following Hume, that there are no rationally prescribed ends. Instead, behaving rationally is simply a matter of pursuing one’s ends in a coherent way. I’ll discuss this view further in the following chapter. I mention it here because non-naturalists believe the opposite: there are rationally prescribed ends, which means that an agent can act in ways that are internally coherent – for instance, she always takes the means she believes necessary to achieve her ends – and still fail to be fully rational, because she has adopted the wrong ends. Some of those ‘wrong ends’ will be those that are morally wrong, and part of what is supposed to be self-evident is
which ends are morally wrong. And now we are in a position to see why our
capacity to grasp self-evident moral truths cannot be a side-effect of the
development of our ordinary reasoning capacities: because it is quite clear that the
seeming ability to grasp ‘self-evident moral truths’ varies independently of one’s
competence in other areas of reasoning. When I drew a psychological distinction
between the moral and the non-moral person, I said that a person with a capacity
for moral judgement thinks she is required to make certain sacrifices, whether or
not it suits her to. This is just the sort of thing that the non-naturalist will think is
self-evidently true. But psychopaths tend not to grasp this ‘self-evident truth’ even
though they may be extremely intelligent in other ways, while other people with
relatively poor reasoning skills in non-moral areas do appear to grasp it. So, our
supposed ability to grasp self-evident moral truths does not appear to be a side-
effect of our ordinary reasoning faculty. If it is not God-given, then we must
suppose that it is an evolved ability, distinct from ordinary reasoning.

Someone proposing an evolutionary account of moral intuition has the
advantage that she would not have to posit a causal link between us and non-
natural moral properties. Unfortunately, that might also turn out to be the chief
defect of her account, for it might be held that the only sort of thing that could have
influenced our evolutionary development is something that made a difference to
the survival chances of our ancestors. We can understand how pressures from the
physical environment have sculpted our physique and our sensory capacities, but
what influence could non-natural properties – which cannot kill us, or nourish us,
or keep us warm – have had on our ancestors’ chances of survival?

The sort of explanation that might be given is that when our ancestors started
to live in large social groups, that included non-family members, they would have
fared better if they had been able to grasp certain moral truths, and that this is no more mysterious than that they would have done better if they had been able to grasp, say, certain counterfactual truths, with whose subject matter they were not in causal contact with either. So, we might think that it was those that grasped that you mustn’t take your neighbour’s spear without permission, and that you have to keep your promises, and return a favour for a favour, that did best in the new social environment. The ability to grasp moral facts was advantageous because it was socially advantageous.

But the problem here is that this is just the same as the evolutionary debunking account, except moral facts have been gratuitously added so that it is not debunking. It was advantageous to our ancestors to coalesce into larger social groups, and it was advantageous for members of those groups to be thought of as trustworthy, and to feel entitled to make certain demands on each other, in order that they could reap the benefits of cooperative endeavours. This may only have been possible if our ancestors had developed a psychology that allowed them to think that certain actions were non-optionally demanded of them. But that gives us no reason to think that the actions they judged to be non-optionally demanded of them really were. Rather, we should expect that the sort of evaluative judgements we would be disposed to make, if natural selection had a hand in shaping the disposition, would be just those that happened to increase our chances of passing on our genes. So we have no reason to posit objective values in order to explain the origins of our capacity to make moral judgments: the principle of parsimony requires us to reject them.

If the non-naturalist is forced to give an evolutionary explanation of how we developed the capacity to make moral judgements then she ought to accept that any
account she gives will have to be compared with rival explanations. In deciding which explanation is most likely to be true, we should give no weight to which explanation we would prefer to be true. But it would seem that the only reason to prefer the less parsimonious non-naturalist’s account would be that it allows us to say that there are moral facts. It is the non-naturalist, and not the error theorist, who is obliged to give an evolutionary account of our moral sensibility, which fares poorly in comparison to the sort of account the error theorist can offer.

Mackie’s argument for error theory proceeds by first ruling out each competing ethical theory except for non-naturalism. (I got us to this point in my discussion of the six metaethical theses.) Mackie does not prove that non-naturalism cannot be true, but instead asks us to consider its metaphysical and epistemological implications, and to consider how implausible these turn out to be. We can avoid commitment to these implausible implications in our metaethical theory by denying that any statement attributing a moral property to an object is ever true. Further support for error theory can be gained by considering how we have come to have a moral psychology. I argued that non-naturalists are obliged to give an evolutionary account of this, but the accounts that they can give will be less plausible than the evolutionary account the error theorist can give.

1.3.4 Richard Joyce’s argument for error theory

The queerness argument is designed to show that error theory is more plausible than non-naturalism, but it may be thought that the argument will not be effective against metaethical theorists who think that moral obligations can be discovered
merely by reflecting on our status as rational agents. Joyce’s argument for error
theory seems designed more with these theories in mind.

Joyce accepts *Assertion* (Joyce 2001: 9-16), *Objectivity* (95-99), and
*Rationalism* (30-39). As we will see, he would also accept that *Alethic* is true if
moral discourse is in good standing. I assume that he would accept *Supervenience*,
but it plays no part in his argument. I’m not quite sure whether he rejects
motivational internalism or if he is just agnostic about it (he discusses it between
pages 17 and 27). In any case, his argument does not depend on it.

The thrust of his argument is that nothing could be a reason for action that was
independent of our desires or interests, but as the concept of a moral reason just is
the concept of a reason that is independent of our desires or interests, that concept
has no instances. This is how he presents his central argument:

(1) If \( x \) morally ought to \( \phi \), then \( x \) ought to \( \phi \) regardless of whether he cares to,
    regardless of whether \( \phi \)ing satisfies any of his desires or furthers his
    interests.

(2) If \( x \) morally ought to \( \phi \), then \( x \) has a reason for \( \phi \)ing.

(3) Therefore, if \( x \) morally ought to \( \phi \), then \( x \) has a reason to \( \phi \) regardless of
    whether \( \phi \)ing serves his desires or furthers his interests.

(4) But there is no sense to be made of such reasons.

(5) Therefore, \( x \) is never under a moral obligation. (42)

The first premise follows from *Objectivity*, and the second from *Rationalism*. Joyce
gives a lot of interesting and, to my mind, persuasive arguments for why we should
accept those theses, but what I want to concentrate on is his argument for premise
(4). Is it true that there can be no reason to do anything that does not serve our desires or further our interests?

To answer that we need to know what it is to have a reason to do anything (a practical or normative reason). There doesn’t seem to be much that could be said about what a practical reason is to someone who didn’t already have the concept. T.M. Scanlon seems to be right when he says

Any attempt to explain what it is to be a reason for something seems to me to lead back to the same idea: a consideration that counts in favour of it. “Counts in favour how?” one might ask. “By providing a reason for it” seems to be the only answer”. (Scanlon 1998: 17)

But even if we cannot elucidate what practical reasons are, we can at least attempt to say under what circumstances we have them. The key issue for us here is whether reasons are always ‘internal’ or sometimes ‘external’, in the senses these terms are used in Bernard Williams’ famous paper “Internal and External Reasons” (Williams 1981). According to Williams, we have an internal reason to φ if and only if we would be motivated to φ after deliberating flawlessly from relevant true beliefs. An external reason to φ, on the other hand, would be one that we had whether or not we would be motivated to φ under those conditions. One of the chief claims Williams makes in support of his denial of external reasons is that a normative reason has to be something that could explain an agent’s action. In other words, it has to be possible that it would be the agent’s reason for acting. But any answer to the question of why an agent acted in a particular way will always make at least a tacit reference to her motivations. So any time an agent acts for an allegedly external reason, she is really only acting for an internal reason.
Joyce agrees with Williams that all reasons are internal, and that is how he supports premise (4). I’m not going to embark on a defence of this view, as the main virtue of Joyce’s argument is its effectiveness against realists who do accept that reasons are internal. However, it may be thought that we can reject (4), even if we accept that all reasons are internal. If we look at premise (1), it says that if \( x \) morally ought to \( \varphi \), then he ought to \( \varphi \) regardless of his desires (I’m going to ignore the mention of interests). But it is consistent with internalism about reasons that an agent can have a reason to \( \varphi \) without having an *actual* desire to \( \varphi \). All that the internalist insists on is that the agent would be motivated to \( \varphi \) if he deliberated flawlessly on the matter, in the light of all the relevant facts. By going through that process the agent can both lose previous motivations and gain new ones. So even if Williams’ thesis is true, a realist claim that such and such an activity is wrong would not be directly threatened by the fact that some people are not actually motivated to avoid that activity. They might simply be ignorant, lacking in imagination, or muddleheaded.

We could try to get around this objection simply by interpreting the desires mentioned in Joyce’s argument to be the desires an agent would have in relation to \( \varphi \)-ing, after deliberating flawlessly in the light of all the relevant facts. I’ll call these the agent’s ‘perfected desires’. That might be a good way to interpret the argument if we were directing it against a realist who rejects internalism about reasons. But it wouldn’t make sense to interpret it in that way if we were directing it against a realist who accepts internalism about reasons. If we were to interpret it in that way, premise (1) would be equivalent to:
(1') If $x$ morally ought to $\varphi$, then $x$ ought to $\varphi$ regardless of whether $\varphi$ing satisfies any of his perfected desires.

And the realist who accepted internalism about reasons would simply reject (1'). As it is this sort of realist I want to focus on here, I think it might be worthwhile to reformulate the argument.

In developing his argument, Joyce contrasts his view of practical reasons with the views of Michael Smith. Smith is an internalist about reasons, but he thinks that if we were all to perfect our desires, we would converge on what we would desire to do. That is precisely what Joyce denies. It might make things clearer, then, if we reformulate Joyce’s argument so that it is about everyone, rather than a single agent, and so that it includes the notion of a perfected desire:

(i) For any agent $x$ and for any action $\varphi$, if it is wrong to $\varphi$, then $x$ ought not to $\varphi$ regardless of what his perfected desires turn out to be.

(ii) For any $x$ and for any $\varphi$, if $x$ ought not to $\varphi$, then $x$ has a reason not to $\varphi$.

(iii) Thus, for any $x$, if it is wrong to $\varphi$, then $x$ has a reason not to $\varphi$, regardless of what his perfected desires turn out to be.

(iv) But there is no sense to be made of such reasons.

(v) Thus there is no $x$ such that for any $\varphi$, it would be wrong for $x$ to $\varphi$.

In the original formulation of the argument, the thing that wasn’t meant to make sense was the idea of an external reason. In this formulation the thing that isn’t meant to make sense is the idea of reasons that apply universally because of the convergence in our perfected desires.
Joyce claims we have no reason to expect that our desires would converge if we deliberated flawlessly. Explaining why he is sceptical of the claim that there would be any such convergence, he writes:

Why should there be? Actual agents have very different interests, desires, projects, ideologies, beliefs etc., and there is every reason to think that this variation will transfer to their idealized versions.

The idealized agent [i.e. the agent who has perfected her desires] certainly has desires that the actual agent does not, and vice versa. Still, the idealized agent is derived from the actual agent, and so the desires of the former will be influenced by the desires of the latter. The relation is rather like that of parent to child: what the parent desires for the child may differ from what the child desires, but the child’s desires have an important effect on what the parent desires for the child. (76)24

So from disparate starting points we are to expect disparate end points. That sounds like a reasonable prediction, but what can Smith say in reply?

Smith’s objection is that practical reasons25 are always non-relative, because the concept of a practical reason is non-relative. So, because reasons are grounded in perfected desires, when I am deciding what I have reason to do in my current circumstances, I am effectively deciding what anyone’s perfected desires would be if they were in my circumstances. One thing he says in support of this is that if we were to suppose that practical reasons are relativized to individuals then we would potentially be talking past each other when we expressed judgements about reasons:

If I say to you that ‘There is a reason for φ-ing’ and you deny this, we are therefore potentially talking about quite different things: reasons_{me} and reasons_{you}. (Smith 1994: 167)

---

24 These two quotes from page 76 are out of sequence, but I haven’t distorted their sense by presenting them in this way.
25 Smith, like Joyce, actually talks of ‘normative’ rather than ‘practical’ reasons.
This, he says, is undermined by the consideration that none of our other normative concepts display this relativity, such as truth, meaning, evidential support, entailment, desirability (Smith’s examples).

This seems a misguided strategy. It would indeed be disastrous if the word ‘reason’ changed its meaning in the mouths of different speakers. But what the advocate of the relativity of reasons claims is that if we take two individuals, Sarah and Bob, then Sarah may have a reason to \( \phi \) in circumstances C, while Bob has no reason to \( \phi \) in exactly the same circumstances, given contingent facts about their preferences. And of course, it is completely obvious that we do express judgements like these, without equivocating on the meaning of ‘reason’. Smith doesn’t think that this sort of case undermines his claim that practical reasons are non-relative, because he thinks we should take an individual’s preferences into account when we are describing their circumstances. So we might say that Sarah, given her preferences is in C\(_1\), while Bob, given his preferences, is in C\(_2\), and that anyone has a reason to \( \phi \) when in C\(_1\) but not in C\(_2\).

But this is all the relativity that the error theorist needs. As Joyce notes, a hardened criminal may well agree that if other people were in similar circumstances to himself, but had very different preferences, then they would have no reason to act as he acts:

And so the criminal carries on stealing, the law-abiding citizen carries on manner, and so one has a normative reason to steal, and the other has a normative reason not to steal. (Joyce 2001: 94)

Another reason to prefer Joyce’s account to Smith’s is because of a hazard Smith himself notes. If it should turn out not to be true that there would be a complete convergence in our perfected desires, then we would have to say that
there is no such thing as a practical reason (Smith 1994: 166). This because of his claims that (a) the concept of a practical reason is non-relative; and (b) the test of which practical reasons we have is what we would all desire to do if we all perfected our desires. If there would in fact be no convergence in any of our perfected desires, then there would be no practical reasons whatever. It is an empirical question whether our desires would converge under conditions of full rationality. And, unfortunately, it does not look as though this is a question we could possibly answer (irrationality being somewhat difficult to eradicate). Thus, if we follow Smith, the appropriate stance to take towards the existence of practical reasons would appear to be scepticism.

Of course what is likely to make us resist saying that the criminal has a reason to murder or steal is that it may seem to us that to say that he has this reason is to endorse his nefarious ways. It is partly from this consideration that Allan Gibbard develops a noncognitive analysis of reason talk (Gibbard 1990). I think it is clear enough, however, that we can talk about the practical reasons a person has without endorsing their acting on those reasons. For instance, if we are in business we can understand that our competitors have reasons to act in ways that damage our profits. And if we are at war, we can understand that our enemies have reasons to act in ways that damage our chances of winning.

In conclusion then, on a conception of reasons as internal, an error theory is far more plausible than a realist construal of moral reasons.

1.4 Conclusion

In this chapter I have taken a stand on a number of metaethical theses. On the most plausible view of the way we use moral language, indicative moral sentences are
used to assert that objects have moral properties; that those moral properties are
objective; that they are the source of reasons for us to behave in some ways rather
than others; that the moral judgements we make have implications about what we
are motivated to do; and that moral properties are determined by non-moral
properties. I then looked at the metaphysical implications of moral properties
having these characteristics. On Mackie’s view, the implication is that they will be
non-natural properties. I have tried to show that we can give an explanation of why
we think and speak in ways that entail the existence of these properties without
assuming there really are such properties, and that this is explanation is superior –
more credible and more parsimonious – than any explanation that assumes they do
exist. However, it is possible to accept the theses I have accepted and not think that
this entails the existence of non-natural properties. Smith and others have thought
that the moral reasons that apply to us do so because of the nature of reasons and
our status as rational beings. However, I have agreed with Joyce that reflection on
the nature of reasons and our status as rational beings does not support this
conclusion.

I am going to complete my defence of the error theory by considering a
number of objections that have been raised against it. This I turn to in the next
chapter.
2: Objections

In this chapter I look at five objections that have been made against error theory, and show how each of them can be dealt with.

The first objection is that error theory is self-undermining because it entails first-order moral propositions of the sort that it is meant to rule out (section 2.1). It is a consequence of error theory that all statements of the form ‘x is not wrong’ are untrue. Proponents of the self-undermining objection claim that ‘x is not wrong’ is equivalent to ‘x is permissible’, and that that is a moral judgement. I show that the alleged equivalence only holds if we are operating under the presupposition that objects have moral properties, which the error theorist obviously rejects. I also show why we must be able to discuss moral discourse at the meta-level (i.e. without making first-order moral judgements).

The second objection is that error theory defies commonsense (2.2). Even if we were unable to say what was wrong with the arguments for error theory, it might be thought that our confidence that there are moral facts is sufficiently robust that we can dismiss any sceptical argument for their nonexistence. I argue that while such ‘commonsense arguments’ can sometimes by a reasonable way of responding to sceptical theses, this is because the rival commonsense view is supported by strong independent evidence that is not undermined by the premises of the sceptic’s argument. I show that the opponents of error theory do not have such evidence.

The third objection is that the bare possibility of moral facts at worlds non-morally identical to our own, in conjunction with the supervenience of moral facts on non-moral facts, entails the existence of moral facts at our world, and thus the falsity of error theory (2.3). I show that, although it has not hitherto been noted, it
can be shown how each argument for error theory also entails that there are no moral facts at worlds non-morally identical to our own. From this it follows that the error theorist is not obliged to accept the premise that moral facts are possible at worlds like ours in non-moral respects.

My discussion of the fourth objection takes up the bulk of this chapter (2.4). The basic objection is that if moral error theory is true, then we ought, for consistency to, accept a corresponding epistemic error theory. Just as moral discourse commits us to categorical moral reasons such, so, the objection goes, epistemic discourse commits us to categorical epistemic reasons, which are no more or less queer than their moral counterparts. But, the objection continues, it would be absurd to deny that we have epistemic reasons, such as a reason to adjust our beliefs to the available evidence.

My response is to reject the claim that epistemic reasons are categorical. I accept that we should not construe epistemic imperatives as hypothetical, because their application to us is not contingent on our goals or desires. But denying that they are hypothetical need not lead us to thinking they are categorical. This is because for an imperative to be categorical, it must not only apply to us independently of our desires, but our reason to abide by it must be independent of our desires. In section 2.4.1 I define ‘non-hypothetical imperatives’, which are like categorical imperative in that they apply universally, but like hypothetical imperatives in that our reasons to abide by them are contingent on our desires. In section 2.4.2 I argue that attempts to motivate the claim that epistemic imperatives are categorical only succeed in showing that they are non-hypothetical. They apply universally, but our reasons to abide by them are either self-interested or moral reasons.
In section 2.4.3 I consider a further worry against the claim that epistemic imperatives are categorical. The formation of beliefs is widely held to be involuntary, but it is also widely held that we can only be obliged to do that which it is in our power to do. So it cannot be the case that we *ought* – in the normative sense of that term – to believe anything at all. Against this it has been argued that epistemic oughts apply to us in virtue of the fact that we occupy a certain role: the role of the believer (Feldman 2000; Cuneo 2007: 82). To say that an agent ought to believe that *p* is equivalent to saying that she would fulfil that role better if she believed that *p*, and this remains the case whether or not she can voluntarily believe that *p*. However, it seems pertinent to ask why the believer should attempt to fulfil her role as well as she can. I find no reason for thinking that she has an irreducibly epistemic reason to do so.

Finally, I consider a pair of arguments due to Terence Cuneo to the conclusion that reductionism about epistemic reasons contradicts both *externalist* and *internalist* theories of epistemic justification (2.4.4). I show how, in the course of his first argument, Cuneo erroneously equates an agent being externally justified with her responding appropriately to reasons. With regard to the second argument, I show how an internalist can understand her thesis in a way that doesn’t saddle her with irreducible epistemic reasons.

The last objection I consider is the claim made by Jean Hampton (1998) and Christine Korsgaard (2008) that our reason to abide by the hypothetical imperatives that apply to us is not wholly dependent on our contingent desires (2.5). The fact that it would be rational to abide by the imperative is supposed to provide a reason to abide by it that is independent of our desires. As reasons that exist independently of our desires are categorical, this objection is also meant to
show that we cannot object to moral claims on the grounds that they imply
categorical imperatives. However, I argue that the fact that an action would be
rational does not provide a reason for us to perform it independently of our desires.

I will now consider the objections in detail.

2.1 Error theory is self-undermining

Some philosophers have thought that error theory is self-undermining (e.g.
Dworkin 1996; Nagel 2003; see also Sinnott-Armstrong 2006: 33-35). To see why,
consider that it is a consequence of error theory that sex outside marriage is not
wrong (because nothing is wrong). If an activity is not wrong, the objection goes,
then it is permissible; but in that case error theory entails that an indicative moral
sentence is true: “extra-marital sex is permissible”. Thus error theory undermines
itself. The best way for error theorists to respond to this is to admit that,
conversationally, we would normally hear “x is not wrong” as implying “x is
permissible”, but to point out that this is because we presuppose that all actions can
be morally evaluated, i.e. that they all have some moral property or other. And it
indeed follows from the claim “x is not wrong” that x is permissible, because
everything that is not wrong is permissible, if we presuppose that some things are
wrong. The error theorist, however, simply rejects the presupposition that acts can
be morally evaluated (Shafer-Landau 2010; Olson 2011a).

The self-undermining objection suggests a kind of scepticism about
metaethics as a subject matter at all, because one of its implications is that we
cannot form opinions about our first-order moral discourse without engaging in
that discourse. But why should we be unable to talk about moral discourse at the
meta-level? Thomas Nagel writes that

... the only response possible to the charge that a morality of individual
rights is nothing but a load of bourgeois ideology, or an instrument of
male domination, or that the requirement to love your neighbor is really
an expression of fear, hatred, and resentment of your neighbor, is to
consider again, in light of these suggestions, whether the reasons for
respecting individual rights or caring about others can be sustained, or
whether they disguise something that is not a reason at all. And this is a
new moral question. One cannot just exit from the domain of moral
reflection: It is simply there. All one can do is to proceed with it in light
of whatever new historical or psychological evidence may be offered.
(Nagel 2003: 21; my emphasis)

It is a little unclear what Nagel means by the claim that one “cannot just exit from
the domain of moral reflection: It is simply there”. Is it that, as a matter of
contingent, psychological fact, human beings are unable to avoid moral reflection?
If so, it is false, as there are people who can reflect, but have little or no capacity
for moral reflection (psychopaths are the stock example). Perhaps it would be more
charitable to interpret him as saying that when we who are capable of reflecting on
moral matters do so, we cannot do so without thinking in ways that have as their
end a moral conclusion, i.e. a first-order moral conclusion. And the point is
presumably not that we cannot do this because of a contingent fact about our
psychology, but that to engage in such reflection just is to think in a way that has as
its end a conclusion about how things are, morally speaking. If this claim is true,
then the error theorist’s project is doomed to fail from the outset, for the error
theorist claims that there are no truths about how things are, morally speaking.

But Nagel is clearly wrong in saying that the only possible way to respond to
an attack on one of your moral judgements is with a further moral judgement. For
you just might be persuaded by your critic that, say, the requirement to love your
neighbour is really an expression of fear, hatred, and resentment of your neighbour; or that a morality of individual rights is nothing but bourgeois ideology. The first of these claims doesn’t sound at all like a first-order moral claim. The second might be interpreted as a first-order moral claim if ‘bourgeois’ is being used, as it usually is, to express disapproval. However, we are entitled to use it merely as a synonym for ‘capitalist’ (which not everyone takes to have negative connotations); and we can note that not all disapproval is moral disapproval (we might consider our disapproval of everything bourgeois to be purely subjective).

Whether or not any of the arguments for error theory are sound, it has to be admitted that the theory is not incoherent in the way suggested by this objection. To deny there are objective moral values is inter alia to deny that there is a particular set of rules that forbid or commend certain activities. And we must be allowed to deny that some $x$ is forbidden by some $y$, without our having to admit, at the same time, that $y$ exists. Look at it this way. I can deny that eating apples is forbidden by Zeus, by saying that the sentence “Zeus forbids apple eating” is not true. Surely everything that is not forbidden by Zeus is permitted by Zeus? So I cannot deny that Zeus forbids the eating of apples without implying that he permits something, which in turn implies that Zeus exists. Of course, that’s absurd. And just as we do not have to admit the reality of Zeus to deny that the sentence “Zeus forbids apple eating” is true, so we do not have to admit the reality of morality to deny that “Sex outside marriage is wrong” is true.

Sceptics concerning metaethics have sometimes claimed that there is no Archimedean point, from which we can stand outside of moral discourse, to discuss issues such as the truth or objectivity of moral judgements, without bringing moral concerns to bear on the issues discussed, and thus at least tacitly endorsing some
first-order moral claims (Dworkin 1996: 88). A quick – and, I think, effective – response to that line of reasoning is to point out that the view just described is itself an opinion about the nature of moral discourse, but it does not plausibly express a first-order moral judgement. So ironically it is the denial of metaethics that is self-undermining.\(^{26}\) It is also worth noting that philosophers who go on to become error theorists do bring first-order moral concerns to bear in their reasoning. It is just that their pre-theoretical intuitions about what is right and wrong do not survive the process of reflective equilibrium.\(^{27}\) There is at least one discourse that one cannot intelligibly talk about without tacitly endorsing. One cannot intelligibly talk about logic without following (and thereby tacitly endorsing) some logical rules, for one cannot intelligibly talk about anything without following logical rules. For instance, if I claim that, firstly, Plato was a philosopher, and secondly that he was an Athenian, you won’t know what to make of these claims if I go on to deny that Plato was both a philosopher and an Athenian. So I have to at least follow the rule for conjunction introduction if I am to be intelligible. But one can intelligibly talk about morality without following (and thus tacitly endorsing) any moral rules, or expressing any distinctively moral judgements. It is intelligible to deny that there are moral rules. But what moral rule is followed, or which moral judgement is expressed, by someone making the claim that there are no moral rules?

2.2 Error theory defies commonsense

The second objection I will consider is that error theory flies in the face of commonsense. The striking thing about error theory is that it denies that we should

\(^{26}\) Paul Bloomfield makes a similar point in response to Dworkin (Bloomfield 2009: 290).
\(^{27}\) Thanks to David Liggins for this point.
attribute rightness or wrongness to actions or events that are difficult to think about without those epithets coming to mind. The error theorist denies that genocide or sadistic torture are wrong, for instance, or that honesty or compassion are ever the right responses to a situation (in the moral sense of ‘right’). These are jarring, unpalatable thoughts. To many ears, they sound simply absurd. One might be tempted to resist them by echoing the ‘commonsense’ philosophy of G.E. Moore (Moore 1959).\(^2\)

Part of Moore’s complaint about some of the wilder claims of philosophers – that the mental is fundamental, or that time is unreal, for instance – was that the philosophers who made such claims frequently contradicted themselves, saying things that implied that time is real or that there is an extra-mental reality. I don’t think that amounts to much as a philosophical complaint (although it might as a complaint about particular philosophers). The important thought that Moore had was that there are some propositions that we know to be true, and that being so, we should give no credence to any sceptical argument that contradicts them, even if we are unable to say which amongst the premises of the sceptical argument are false.

This is certainly a tempting response to many philosophical arguments. A defender of the commonsense approach, David Armstrong, offers this example:

> It is a very fundamental part of the Moorean corpus that there is motion. Things move. Perhaps we have still not, after two and a half thousand years, got to the full bottom of Zeno’s brilliant arguments against the existence of motion. What motion is in its deep essence may be a mystery, and fully answering Zeno may be one part of the resolving of that mystery. (Though I myself would look to science for major illumination.) But certainly Zeno should not persuade us that things do not move. Neither should anybody else. (Armstrong 1999: 79)

\(^2\) Ronald Dworkin provides such an argument (Dworkin 1996: 117-8).
I reacted as most people probably do when I first heard Zeno’s tale about Achilles and the tortoise: I didn’t know exactly what was wrong with the argument, but I knew it had to be wrong. As Armstrong says, things move. We have a strong feeling that our intransigence in this sort of case is reasonable, even when we are unable to defeat the argument.

It is fairly obvious what gives rise to this feeling in the case of arguments against motion: it is that we have strong competing evidence for realism about motion. First of all, we have a variety of sensory evidence for motion: visual, auditory, tactile, and proprioceptive. These senses don’t just accidently give us the impression of movement: it is a large part of their function to detect movement. Secondly, the concept of movement is central to many of our best confirmed scientific theories. If we were persuaded by Zeno’s arguments we would have to make radical alterations to our ‘web of belief’. And thirdly, there do not seem to be any credible explanations available of why our sensory and scientific evidence for motion could turn out to be so misleading. It is not surprising, then, that most of us are not convinced by Zeno’s abstract logical argument, and rest content that *someone else* will be able to point out the flaw that it must contain.

In the moral case, by contrast, we don’t have the evidence of our senses to prop up a belief in moral values; nor do moral values figure in scientific explanations. The removal of moral beliefs from our ‘web of belief’ would require little adjustment to the rest of the web. And there are reasonably plausible debunking explanations for our use of moral discourse (see section 1.3.3) So whilst our confidence in our belief that movement is a real phenomenon is supported by one of our most trusted epistemic sources (sensory perception), by its use in our
most sophisticated epistemic enterprise (natural science), and by the lack of a credible explanation of how we might be fooled into thinking that movement exists, our confidence in the existence of moral values is not backed up in any of these ways.²⁹

Our confidence in certain of our moral beliefs may best be likened to the confidence that a great many people have in their religious beliefs. Religious and moral beliefs for many people are central to their identity, so it is hardly surprising that they give short shrift to sceptical arguments against them. But even if confidence in the truth of a proposition is some indicator that the proposition merits our belief, it is clearly a defeasible one.

### 2.3 Moral facts are possible, thus moral facts are actual

In a recent paper, Christian Coons has argued that once we acknowledge the bare possibility of moral facts, that is enough to defeat the error theory. For his argument to work it must be the case that the arguments for error theory do not exclude the possibility of moral facts. Otherwise, relying on the possibility of moral facts as a premise will beg the question against error theory. Coons does not think that the arguments for error theory exclude the possibility of moral facts:

> For if error-theorists felt they already had reason to indict the possibility of moral facts, then appeals to debunking explanations of moral beliefs and theoretical simplicity would be superfluous. (Coons 2011: 89)

²⁹ I take inspiration for the points made in this paragraph from a paper by Tristram McPherson (McPherson 2009). McPherson also observes that Moore himself did not apply his argument from commonsense in support of his moral realism, nor did any moral proposition appear amongst the list of propositions that Moore took us all to be certain of in “A Defence of Common Sense”.

92
The other main premise in Coons’s argument is the truth of *Supervenience*:

*Supervenience*: moral properties (if they exist) supervene on non-moral properties, such that the subvening non-moral properties wholly determine the supervening moral properties.

His argument then proceeds as follows:

(1) There is a possible world, non-morally identical to our own, where there are moral properties.

(2) Moral properties, where they exist, supervene on non-moral properties, such that the subvening non-moral properties wholly determine the supervening moral properties.

(3) Thus any two worlds $W_1$ and $W_2$ that are identical in the distribution of their non-moral properties will be identical in their distribution of their moral properties.

(4) Thus, if there is a possible world, non-morally identical to our own, at which there are moral properties, then our world is a world at which there are moral properties.

(5) Thus (from (1) and (4)) our world is a world at which there are moral properties.\(^{30}\)

One point to pick up on here is Coons’s move from the *possibility of moral facts at some world or other* to the *possibility of moral facts at worlds non-morally...* 

---

30 The argument he presents in his paper is actually a bit more ambitious than this, since he claims to demonstrate a substantive first-order moral claim. But the less ambitious argument that I attribute to him here is contained within the one he presents in his paper.
identical to our own. Error theorists needn’t claim that moral statements are necessarily untrue. They may be content to argue merely that they are untrue at worlds like our own. So why does Coons think he is entitled to (1)? All that Coons has to say on this matter is that he assumes that “no non-moral features of the actual world preclude the instantiation of moral properties” (87), and that “moral facts are coherent and … we cannot rule out apriori that they can be true” (87). As we’ll see, these are pretty big assumptions to make in this context.

I don’t think error theorists should be at all concerned by Coons’s argument, because there is no reason for them to accept premise (1). To see this, let us distinguish two modal claims:

Weak: There is a possible world where there are moral properties.

Strong: There is a possible world, non-morally identical to our own, where there are moral properties.

Coons needs Strong to be true for his argument to go through. As Strong entails Weak, the falsity of either thesis will be enough to defeat his argument. The question then is whether error theorists have the resources to argue for the falsity of either of these theses. I think it is clear that they do.

Take Mackie’s relativity argument first. The kernel of that argument is that the observed diversity in moral opinions at different times and places is “more readily explained by the hypothesis that they reflect ways of life than by the hypothesis that they express perceptions, most of them seriously inadequate and badly distorted, of objective values” (Mackie 1977: 37). Assume that Mackie is
right both about the diversity of moral opinions at different times and places and about the best explanation of this fact. Then any world that is a non-moral duplicate of our own will be one at which there is a diversity of moral opinions that is best explained in the way that Mackie explains it. Unless we think we should believe some explanation for the diversity of moral opinions other than the best one, then we should conclude that Strong is false. The argument from relativity is toothless against Weak, however. For we can imagine worlds where there is a complete consensus over moral matters; and we can imagine worlds where the diversity of moral opinions is best explained by some other factor, such as differences in the reasoning capacities of individuals or differences in their access to empirical information that bears on moral matters. Error theory might still be true at such worlds, but the argument from relativity would not establish it.

If the argument from queerness is sound, it also refutes Strong. As we saw in the previous chapter, Mackie thought that believing in moral properties would require us to believe in entities utterly different from anything else in the universe, which would require a mysterious faculty of intuition if we were to know about them (section 1.3.1). In light of this, Mackie thought it more reasonable to accept a debunking explanation of moral discourse. If Mackie is right in his analysis of what moral terms purport to refer to, and if he is right that it is most reasonable to accept a debunking explanation of our use of moral language in the actual world, then it is most reasonable to accept a debunking explanation of our counterparts’ use of moral language in any non-moral duplicate of our world. After all, the conditions that make the debunking explanation most reasonable to accept in our world will be present in the non-moral duplicate of our world. Does the argument from queerness, if sound, refute Weak? It doesn’t look like it. We might, for instance, be
the most diehard naturalists about the ontology of this world, but be prepared to
grant there are possible worlds with non-naturalistic ontologies. Or (perhaps
relatedly) we might be able to conceive of worlds at which there are no plausible
debunking explanations for the use of moral terms at that world, and so the most
reasonable thing to say about moral terms there is that they successfully refer.
Certainly these possibilities are not ruled out by anything contained within the
argument from queerness, though a defender of the argument might want to
provide supplementary arguments to rule out such possibilities and refute Weak.

Joyce’s argument for error theory also rules out Strong. As we saw in
section 1.3.4, Joyce argues that any reason that an agent has for acting in one way
rather than another must be ultimately grounded in her desires or motivations (so
all such reasons are ‘internal’ in the sense made famous by Bernard Williams
(1981)). But putative moral reasons are not supposed to be grounded in the desires
or motivations of the agents to whom they are attributed: we wouldn’t excuse a
murderous tyrant from moral criticism if the evidence suggested that, of the
choices available to her, the ones she made best satisfied her desires. So, Joyce
concludes, putative moral reasons are not real reasons for acting. There would be
reasons that applied to all agents, regardless of their particular desires or
motivations, if there were convergence in their desires and motivations. Although
there is not actually such convergence, Joyce seems to consider that it would be
sufficient if the convergence came about under ideal circumstances (e.g. of full
knowledge, full imaginative acquaintance with relevant possibilities, flawless
deliberation). But Joyce argues that there is no reason to expect that such a
convergence would take place, even in ideal circumstances, because there is no
principled reason for thinking that an agent doing something that most of us would
find morally abhorrent must be making an error in reasoning or be lacking some vital information (Joyce 2001, chapters 3 and 4).

Joyce’s argument does not seem to rule out Weak. Even if it is true that there is no reason to expect a convergence in the motivations of agents in this world under ideal circumstances, this is a contingent fact about the group of agents in this world, rather than a necessary fact about groups of agents. In a world where all agents care equally for the wellbeing of themselves and all other agents, and have an equal regard for all non-agents, there plausibly would be such a convergence.

There are a couple of arguments for error theory that do rule out Weak. Bart Streumer’s argument for error theory shares some similarities with the argument from relativity. He begins by defending Frank Jackson’s (1998) argument that normative properties are identical to descriptive properties (to be precise, Streumer defends a conditional form of Jackson’s conclusion: if there are normative properties, then they are identical to descriptive properties) (Streumer 2008). In a subsequent paper he argues that if normative properties are identical to descriptive properties this entails that it is possible (in principle) to say which descriptive properties they are identical to (Streumer 2011). But, according to Streumer, it is not possible (even in principle) to say which descriptive properties normative properties are identical to, because this is determined by how the folk apply normative terms and the folk don’t agree about how to apply normative terms. From these premises it follows that there are no normative properties at this world. Assuming that Streumer’s premises are true, any world that is non-morally identical to ours will be a world at which it is not possible to say which descriptive properties normative properties are identical to, because the folk at that world will not agree in their usage of normative terms either. So Streumer’s argument rules
out *Strong*. As I have sketched his thesis, it would not rule out *Weak*, because we can imagine a world where the folk *do* agree about how to apply normative terms. However, in Streumer’s view, this sort of agreement is not sufficient, because future members of the linguistic community might grow up to disagree with their parents on how to apply normative terms, and thus the normative beliefs of the parents and children would form an inconsistent set. So Streumer’s thesis seems to rule out *Weak* as well as *Strong*.

Lastly, Don Loeb has argued that the persistence of fundamental disputes between cognitivists and noncognitivists “is evidence that inconsistent elements – in particular, commitments both to and against objectivity – may be part of any accurate understanding of the central moral terms” (Loeb 2008: 357-58). He adds “If … moral vocabulary is best understood as semantically incoherent, the metaphysical implication is with respect to that vocabulary there is nothing in particular to be a realist *about* – no properties, that is, that count as referents of the moral terms”. If Loeb is right, then moral terms do not even potentially refer to anything – at this world or any other – so there cannot be a world at which there are moral facts. Loeb’s argument, if sound, would show that all propositions attributing moral properties are necessarily untrue, and thus that error theory is necessarily true (Streumer’s thesis also entails this).

So it seems that, although error theorists have not hitherto considered the modal status of their thesis, they have ample resources for rejecting the first premise of Coons’s argument. And we can see where Coons has gone wrong. In claiming that appeals to debunking arguments and theoretical simplicity would be superfluous if error theorists could demonstrate the impossibility of morality he
fails to appreciate that such appeals can be used as arguments against the possibility of moral facts at worlds non-morally identical to our own.

### 2.4 Partners in guilt arguments (1): The epistemic argument

Mackie wrote that the best way for those realists committed to objectively prescriptive values to counter the argument from queerness “is not to evade the issue, but to look for companions in guilt” (Mackie 1977: 39). I shall consider two of these companions in guilt arguments, in this section and the next. The general strategy is to find some non-moral area of discourse $D$ and show first that (i) speakers of $D$ acquire the same sort of ontological commitments that Mackie takes to be objectionable in the moral realm; and then either that (ii) $D$ cannot be dispensed with, without absurdity; or (iii) the argument for moral error theory makes essential use of $D$. The argument I discuss in this section highlights analogies between moral discourse and epistemic discourse. Proponents of this argument claim that reasons for belief are no less categorical than moral reasons, but that the prospects for an epistemic error theory look dim. Thus if we are forced to concede that there are categorical epistemic reasons, we cannot use the supposed queerness of categorical reasons as a reason to reject them in the moral realm.

#### 2.4.1 The epistemic argument

There is the following very tempting objection to moral error theory. Mackie presumably wanted to convince his readers that error theory is true and took his arguments to give them a good reason to believe that it is true. Now we can ask: is
this reason hypothetical or categorical? If we say it is hypothetical then that must be because we think that his readers would only have a reason to believe that error theory is true if they have some desire that is satisfied by their believing that proposition. But on the face of things, that doesn’t seem right. If I tell you that an essential part of the reason that I believe a particular proposition $p$ is that it in some way suits me to believe $p$, you will very likely respond that my reason to believe any given proposition has everything to do with the evidence available for that proposition, and nothing to do with what it suits me to believe. Believing what suits one, without proper regard for the evidence, is the vice of the wishful-thinker.

I might begin to redeem myself in your eyes if I tell that I am in fact very concerned about conforming my beliefs to the available evidence, and that it suits me to believe $p$, not because of anything special about that particular proposition, but because I have a general desire to believe all and only those propositions where the evidence in favour of them is greater than the evidence against them. That might be an admirable disposition, but you still might think that, nevertheless, this desire does not form an essential part of the reason that I have to believe any proposition, and that I would have a reason to adjust my belief in $p$ to the available evidence even if I did not desire to do so. These considerations make the following thesis looks highly plausible:

\[(\alpha)\] Any two agents in the same epistemic situation (same evidence, same background beliefs) have the same reason for believing any given proposition, regardless of possible differences in their personal goals.\(^{31}\)

\(^{31}\) Adapted from Railton (2003: 293).
And they make this one look implausible:

$$(\beta) \quad \text{For any agent } S \text{ and proposition } p, S \text{ has a reason to believe } p \text{ only if } S \text{ has a goal that is served by her believing } p.$$

This makes it look as though epistemic reasons are categorical rather than hypothetical. However, hypothetical and categorical imperatives differ not only in whether our reasons to abide by them are contingent on our goals, but also in whether their application to us – i.e. whether they are imperatives for us – is contingent on our goals.

A hypothetical imperative is a rule that applies to an agent $S$ only if $S$ has some goal that is satisfied by her abiding by that rule; and she has a reason to abide by the rule only if in that way she attains her goal.\(^{32}\)

A categorical imperative, by contrast, is a rule that applies to all agents, irrespective of their contingent goals. All agents have a reason to abide by the rule, irrespective of their contingent goals.

There is clearly room in logical space for what I will call a ‘non-hypothetical imperative’. Non-hypothetical imperatives are like categorical imperatives in that they are rules that apply to all agents, irrespective of their contingent goals. But they are like hypothetical imperatives in that any agent’s reason to abide by the rule is contingent on their goals.

I think we should accept that there are non-hypothetical imperatives. As Philippa Foot pointed out, there are systems of rules that seem to apply universally,\(^{32}\)

---

32 I respond to a challenge to this way of characterising hypothetical imperatives in section 2.5.
but are not categorical because one does not always have a reason to comply with them (Foot 1972). She used the example of rules of etiquette. Such rules do not usually make exceptions for people who do not desire to comply with them. For instance, you are breaching Japanese etiquette if you do not bow appropriately when meeting a Japanese person, whether or not you care about offending Japanese people. So rules of etiquette apply universally. But it is difficult to imagine a rule of etiquette that we would never have a reason to break. So while imperatives of etiquette are not hypothetical (because they apply universally) they are not categorical either (because our reason to comply with them is contingent on our goals).

If we are only thinking in terms of hypothetical and categorical imperatives, then it will be easy to make the mistake of moving from the observation that a rule applies to all agents, irrespective of their goals, to the conclusion that all agents have a reason to abide by that rule, irrespective of their goals. Once we have the notion of a non-hypothetical imperative on the table, we can see that this move is too swift. For a non-hypothetical imperative can apply to me, whether or not I have any reason to abide by it. I am going to argue that epistemic rules apply universally, but that any agent’s reasons for complying with these rules is contingent on their non-epistemic goals. To facilitate the discussion, I’m going to abandon (α) and (β), and introduce three new theses:

(1) For all agents in the same epistemic situation (same evidence, same background beliefs), there is a set of rules prescribing what to believe with regard to any proposition; and every agent has a reason to abide by those rules, irrespective of their contingent goals.
(2) For all agents in the same epistemic situation (same evidence, same background beliefs), there is a set of rules prescribing what to believe with regard to any proposition; but for any agent S, S has a reason to abide by those rules only if in so doing she satisfies some non-epistemic goal.

(3) For any agent S and proposition p, there is a rule prescribing that S believe p only if S has a goal that is served by her believing p; and S has a reason to abide by that rule only if in so doing she satisfies some non-epistemic goal.

Thesis (1) is the position of the categoricalist; (2) is the position of the non-hypotheticalist; and (3) is the position of the hypotheticalist.

We can now outline a partners in guilt argument. I’ll call my initial version of this argument ‘The Epistemic Argument*’, in order to distinguish it from a revised version of the argument (to be given shortly), which will be the principal focus of this section:

P1* If moral reasons do not exist, then epistemic reasons do not exist.
P2* Epistemic reasons exist.
C* So moral reasons exist.33

The explanation for P1* should be obvious by now. The claim is that epistemic reasons are categorical in the same way that moral reasons are categorical. If moral reasons are ruled out on the grounds that they are categorical, we cannot

---

33 I adapt this argument from Cuneo (2007: 6). Cuneo talks of moral and epistemic facts, rather than reasons, but as the quote at the top of 2.4.2 shows, he thinks that epistemic facts entail epistemic reasons.
consistently refuse to rule out epistemic reasons. The error theorist can refuse to accept P1* if she is prepared to give up the argument from queerness. Then she could provide a different argument for thinking that there are no categorical moral reasons, which does not generalise to reasons in other areas.

A less damaging strategy for the error theorist would be to reject P2*. However, simply rejecting P2* is obviously problematic. If ‘epistemic reasons’ is taken to include reasons for belief then the inevitable objection to whatever argument the error theorist gives for rejecting P2* will be that she is trying to give us reasons to believe that there are no reasons for belief.34 A better strategy would be for the error theorist to claim that the Epistemic Argument* equivocates on the term ‘reason’. As we saw in the previous chapter, the error theorist takes putative moral reasons to be irreducible. If we have a reason to be moral, then this is a sui generis moral reason, and not, for instance, a reason of self-interest. But if it can be shown that epistemic reasons are reducible, then the Epistemic Argument* will not go through. Epistemic reasons might reduce to either ‘instrumental reasons’ – that is reasons that we have in virtue of having some (possibly idiosyncratic) goal – or to moral reasons. Any epistemic reasons that are reducible to instrumental reasons will not be a problem for the error theorist unless instrumental reasons are a problem for the error theorist. In any case, they will not be a special problem for the error theorist. Any epistemic reasons that are reducible to moral reasons will not be a problem for the error theorist, as she will simply dismiss them along with all other moral reasons.

To keep things clear, I give the following as my considered version of the The Epistemic Argument:

---

34 See Cuneo (2007), chapter 4, for this and other arguments against ‘epistemic nihilism’.
P1 If irreducible moral reasons do not exist, then irreducible epistemic reasons do not exist.

P2 Irreducible epistemic reasons exist.

C So irreducible moral reasons exist.

I am going to reject P2 on the grounds that, although there are epistemic rules that apply universally, our reasons for abiding by those rules are always non-epistemic. In section 2.4.2 I look at attempts to motivate the claim that epistemic reasons are categorical. I find that while there is good reason for accepting the claim that epistemic rules are not contingent on our non-epistemic goals, there are no good reasons for thinking that our reasons to abide by those rules are non-contingent.

In 2.4.3 I put forward the argument that it cannot be the case that epistemic reasons are categorical because belief is not sufficiently voluntary for it to be the case that we ought to believe one way or another. I consider and reject responses to this claim from Richard Feldman and Terence Cuneo, who argue that virtue epistemology can save the categorical claim.

Finally, in section 2.4.4 I look at a pair of arguments from Terence Cuneo designed to show that reductionism about epistemic reasons is not compatible with either of two important theories of epistemic justification, internalism and externalism. I argue that Cuneo does not successfully demonstrate this.

Having defeated a number of lines of argument in favour of epistemic reasons being categorical, I conclude that there are no irreducible epistemic reasons. Epistemic reasons are not analogous to moral reasons in a way that is problematic for the error theorist.
2.4.2 Some attempts at motivating the claim that epistemic reasons are categorical

Cuneo is a categoricalist:

[Epistemic facts] are, imply, or indicate reasons for agents to behave in certain ways regardless of whether these agents care about conducting their behaviour in a rational way, whether they belong to a social group of a particular kind, or whether they have entered into social agreements with others … [W]hether I have an epistemic reason, say, to believe a proposition for which I have compelling evidence is not contingent upon whether I care about believing what is true. To this, realists add that epistemic reasons are authoritative inasmuch as the decisiveness for some such reasons for an agent is not a function of whether she wants to act in an epistemically commendable way, or belong to a certain social group, or has entered into certain agreements with others, and so forth. (Cuneo 2007: 59)

Unfortunately, Cuneo merely asserts that epistemic reasons are authoritative in this passage. We need something better than this.

Hilary Putnam also seems to endorse categoricalism in the following passage:

Mackie points out that no physical property has a built-in connection to action (or to approval of an action), and concludes that [properties such as] ‘being the right thing to do’, etc., are ‘ontologically queer’. But … this argument proves too much. For some epistemic predicates (e.g. ‘rationally acceptable’, ‘justified belief’) are also action guiding (taking ‘action’ in a wide sense, so that accepting a statement counts as an action). One can say, ‘X is a good thing to do’ and ‘There is a good deal of evidence that Y’ and not be committed to doing (or prescribing) X or to accepting Y; but if one says that ‘X is the right action to perform in this situation’, or ‘Believing that Y is completely justified’, then one is oriented to doing (or prescribing) X and to accepting Y. ‘Justified’ (in the case of beliefs) has the characteristic of being action-guiding as much as ‘right’ in the moral sense does (Putnam, 1981: 210).
This claim deserves a few comments. We sometimes think to ourselves “φ-ing is the morally right course of action in my situation, so I shall φ”. Doing so may serve to firm our resolve against taking some course of action that is tempting, but which we consider to be less morally admirable than some alternative. It is doubtful that we ever think to ourselves “having belief $B$ in my situation would be completely justified, so I shall have belief $B'$” as a way of firming our resolve against having some tempting but less admirable belief. We don’t do so, because unlike actions, which we can plan, beliefs are not usually acquired voluntarily. We think about whether our present beliefs are justified, rather than whether some beliefs that we are considering having would be justified.

Thoughts about justification might be epistemically action guiding in the following way. If I come to realise that a belief of mine is less well supported than I had supposed, then I may become motivated to think more carefully about the matter, look for more evidence, consider whether I am competent to examine the relevant evidence, consider whether I have made any logical errors etc. However, this exercise will cost me in terms of time and resources, and no one could have the time and resources necessary to examine all, or even very many, of their beliefs in this way. We are most likely to spend time examining our beliefs if they are either relevant to our current endeavours, or concern matters that are of particular interest to us. So the suspicion that a belief of mine is less well supported than I had supposed might stimulate me into inquiry, but only because of other contingent facts about me: facts about my goals and interests. This suggests that my reasons for carrying out these inquiries are not irreducibly epistemic, but are contingent on my non-epistemic goals and interests.
Thomas Kelly has also argued that epistemic reasons are categorical rather than hypothetical (Kelly 2003). However, I think that all he actually demonstrates is that they are non-hypothetical. Let us look at a pair of thought experiments that Kelly uses in an attempt to establish that epistemic reasons are categorical. In the first one, an agent appears to have a reason for believing a particular proposition, even though his forming that belief does not advance any goal of that agent (this is my own version of Kelly’s thought experiment (2003: 625)):

Ray doesn’t know or care what colour the carpet in his colleague’s office is. He has no goal that would be satisfied by having a true belief on the matter, or which would be frustrated by having a false belief on the matter. Still, should he walk into his colleague’s office and see that the carpet is blue, he would have a reason to believe that the carpet is blue, despite his indifference.

In the second example, an agent appears to have a reason for believing a particular proposition, even though forming this belief directly frustrates one of her goals (here I paraphrase Kelly (626)):

Sheila wants to see a film that her friends have already seen. As they are discussing it in her presence, she asks them not to reveal the ending, as to know would spoil her enjoyment of the film when she eventually sees it. However, someone thoughtlessly reveals the ending. Now it would seem that Sheila has a reason to have a particular belief about how the film ends even though having this belief directly frustrates her goal of not knowing how the film ends until she has watched it.
These cases look as though they provide good evidence for categoricalism. It does sound right to say that Ray has a reason to believe that the carpet is blue, and that Sheila has a reason to believe that the film ends as her friend says it ends, despite the fact that they have no goals that are satisfied by so believing. If we want to defend thesis (3) (the hypotheticalist’s thesis), we might try to accommodate this appearance by claiming that everyone (or nearly everyone) has the cognitive goal of having true rather than false beliefs, and so it will appear as if reasons for belief are not contingent on our goals (622). To this move Kelly responds that we typically have no such highly general cognitive goal (623). We do desire to have true, rather than false beliefs, but these desires are associated with specific matters that are of interest to us. We are frequently concerned to form specific true beliefs about, for instance, what the quickest route is from Manchester to London, where the best place is to get a coffee around here, what is the current exchange rate from pounds to Euros and so forth. As Kelly notes, sometimes our cognitive goals are not so specific. If I read a newspaper, for instance, my goal might just be to acquire any information about significant recent events (624). But even with a less specific goal like this, we can see a clear connection between a fairly narrow goal and a means of achieving it (not everyone cares enough about current events to pick up a newspaper).

I think Kelly presents a good case that the hypotheticalist cannot save the appearance of categoricalness in this way. However, we can explain why Ray and Sheila seem to have the reasons they have, despite their goals, without having to admit that epistemic imperatives are categorical. These cases seem compelling
because it is true that we do not withdraw our verdict that an agent is being
epistemically negligent in failing to comply with an epistemic imperative when we
find out that she has no goal that is served by so complying. That is evidence that
epistemic rules apply universally. But it is a further step to say that we always have
a reason to abide by the epistemic rules that apply to us.

There would seem to be something wrong with a person in Ray’s position who
did not form the belief that the carpet is blue. But maybe all that amounts to is that
there is something wrong with them from the epistemic point of view. From that
point of view, beliefs are negatively assessed if they are not well supported by the
available evidence, and positively assessed if they are well supported. These
assessments are perfectly objective: whether or not an agent’s beliefs are supported
by the available evidence has nothing to do with his goals or desires. This
objectivity explains why epistemic rules apply universally. What I ought to believe
from the epistemic point of view has nothing to do with contingent facts about me.
For comparison, consider actions assessed from the point of view of chess strategy.
From the point of view of chess strategy, moves are assessed positively if they are
conducive to winning and negatively if they are conducive to losing. Given the
rules of chess, it will be a perfectly objective fact that, at some point in a game, a
particular move will be better than any alternative from the point of view of chess
strategy. Still, no one is strictly obliged to care about this fact – not even someone
who is playing chess (they might not care if they win or lose). In a similar way, we
might have no reason to care about some of the objective epistemic rules that apply
to us.

To demonstrate that epistemic imperatives are categorical we would have to
show that we always have a reason to abide by them. How could a categoricalist
convince someone who was sceptical of this claim? Her most extreme opponent would be someone who denied that we ever have a reason to abide by them. How might we try to convince this person that she was making a mistake?

We could point out to her that if we abide by epistemic imperatives we gain justified beliefs, and having justified beliefs is, at least on the whole, tremendously useful. But then we would only be giving them an instrumental reason to have justified beliefs (Railton 2003: 294).

A more promising strategy for the categoricalist would be to try to show that it is incoherent not to care about your beliefs being justified. To show this we can start by pointing out that beliefs are subject to a particular form of criticism that other propositional attitudes are not subject to. One can be criticised for holding beliefs that fail to represent the world accurately, which is not the case for other propositional attitudes. For example, I might believe that the moon landings were a hoax, hope that the moon landings were a hoax, imagine that the moon landings were a hoax, assume that the moon landings were a hoax, and so forth. Should it turn out that the moon landings were not a hoax, then it looks as though I ought to revise my belief, but I can happily go on imagining or assuming that they were a hoax. I probably ought to give up hoping that they were a hoax as soon as I discover that they really happened, but that does not show that, before this discovery, I was mistaken for hoping they were a hoax. But the fact that the moon landings were not a hoax does show that I was mistaken for believing them to be a hoax. This seems to show that to believe any proposition, is inter alia to represent oneself as accountable for that belief. As Peter Railton puts it, “The distinctive attitude of belief is … one that not only represents its propositional content as true,
but also one that cannot represent itself as unresponsive to – unaccountable to – their truth” (2003: 297).

To the suggestion that one could abandon beliefs, and propose to get along only with the other propositional attitudes, Railton comments:

Consider how deeply implicated belief is in our notion of agency. An agent acts on intentions and plans, which constitutively involve beliefs and are formed deliberatively in part on the basis of beliefs. To replace all beliefs with (say) wishing would be to form no intentions at all. Moreover, our notions of ourselves as agents extended over time constitutively involve memories and expectations. These, too, involve beliefs. There is all the difference in the world between believing that one is the father of John, or believing that one will experience the pains of an unattended toothache, and pretending or merely supposing these things. To delete all forms of belief from your mental repertoire would leave you with no recognizable notion of identity. (2003: 298; emphasis original)

Railton’s conclusion, then, is that I must have beliefs in order to act at all, or even to have a sense of a personal identity that extends over time; and I must represent my beliefs as being accountable to the truth, for them to count as beliefs. 36

This is all quite persuasive, but there is a difference between representing one’s beliefs as accountable to the truth, and actually holding them accountable to the truth. If we never changed our beliefs to accord with the available evidence, then we would no doubt fail to act effectively, but we can certainly fail in this regard some of the time, without such failures being damaging to our goals.

Returning to Kelly’s examples, we cannot conclude that Ray and Sheila have the reasons that they seem to have because they would not be able to act at all, or they would lose their sense of personal identity, if they failed to have these particular beliefs. All that Railton establishes is that we must try to believe some true things if

---

36 Railton is himself only arguing that reasons for belief are non-hypothetical. In this sub-section, I am exploring ways in which the categoricalist might try to build on Railton’s arguments for their own purposes.
we are to be effective agents, and not that we must always adjust our beliefs to the evidence, if we are to be effective agents with a sense of personal identity.

And not all true beliefs seem to be equally useful to us in deciding how to act. For most of us, if we falsely believed that Jupiter has four moons, rather than at least sixty-three, it is unlikely that this false belief would have a negative effect on our plans and intentions. From a practical point of view, there are some beliefs that it is more important to be right about than others. And which beliefs it is important to be right about will vary greatly between people, in accordance with their different goals. The injunction to hold one’s beliefs accountable to the truth might always be in place – i.e. it might always be the case that a propositional attitude only counts as a belief if its bearer would acknowledge that she would be mistaken in holding it if it doesn’t represent the world accurately. But how scrupulous any particular agent ought to be, with regard to any particular belief, that that belief actually succeeds at representing the world accurately, is a different matter. It would seem that we have more reason –from a practical point of view – to be scrupulous with those beliefs that have a direct bearing on our goals, or concern matters that are of particular interest to us, than those that do not appear to have a direct bearing on our goals, or which we are not particularly interested in.

As a rejoinder to this last point, it might still be claimed that true beliefs will always be more valuable than false ones, even if in particular instances the difference isn’t easily discerned. It might be claimed that although what we believe on some matters may seem to have no practical effect on how we achieve our goals, this will turn out to be false if we note how our different beliefs are related to each other. On this matter Mark Schroeder writes:
Now there are plenty of topics which don’t bear directly on how to succeed in acquiring a new pair of shoes. For example, it is hard to see the connection between buying a pair of shoes and knowing that my brother lives in Los Angeles, or knowing how many moons circle Jupiter. But of those topics which don’t bear directly on buying shoes, some do bear directly on questions that are quite relevant to buying shoes. Being in error about them could lead one to form the wrong beliefs about matters that it is important to be right about, for one’s shoe acquiring to go smoothly. And now consider the class of propositions, being in error about which can affect whether one is right about one of these questions. And then the class of propositions, being in error about which can affect whether one is right about one of these questions. And so on. A relatively weak hypothesis about the holism of belief formation says that the closure of this process will include any proposition. Suppose that that is right. Then being in error about any proposition is ultimately relevant [to the goal of buying a new pair of shoes]. (Schroeder 2005: 114)

This doesn’t undermine my claim that we have more reason to be scrupulous about those beliefs that have a direct bearing on our goals. We don’t need to deny that being in error about any proposition is relevant to our goals, because it is quite clear that being in error about, say, the number of moons circling Jupiter is vanishingly unlikely to affect our ability to buy shoes. Most people can buy shoes; most people don’t know how many moons circle Jupiter. And the fact that our errors are always relevant in this way doesn’t mean that our errors will always be harmful. They may sometimes even do us more good than harm (I discuss this possibility in section 3.2.2). In any case, we are now back squarely with the idea that beliefs are valuable to us insofar as they help us to achieve our non-epistemic goals (a conclusion that I think Schroeder would support). What we have been unable to see, thus far, is why epistemic imperatives should be authoritative over us independently of our non-epistemic goals.
2.4.3 Epistemic Virtues

There is a further worry that we can push against the categoricalist. Returning to Ray, it would seem that he cannot help but form the belief that the carpet is blue, if he notices it at all. Suppose that his colleague is colour-blind and asks Ray what colour the carpet is. Ray will be lying if he says it is any colour other than blue, or says that he cannot determine what colour it is. In fact, it would seem that for anyone in Ray’s situation, it would be impossible not to form the belief that the carpet is blue, at least if the matter of its colour is put to them (and they have normal colour vision). This reflects the widely held view that most of our beliefs are not under our voluntary control. Whether the impossibility of forming beliefs voluntarily is conceptual (Williams 1970), or merely psychological (Alston 1988), needn’t concern us here, as the weaker, psychological thesis is sufficient for my purposes.

Everyone will agree that anyone in Ray’s position would form the belief that the carpet is blue, and would be unable not to form the belief that the carpet is blue. (Those who defend the idea that we have some voluntary control over our beliefs focus on more complex cases, such as when the evidence available for determining the truth or falsity of a particular proposition does not appear decisive on either side). Note that we would not only say that Ray has a reason to believe that the carpet is blue, but that he ought to believe that the carpet is blue, implying that he has a decisive reason for so believing. If Ray cannot help but believe as he does, however, then this latter claim might look peculiar, at least if ‘ought’ is interpreted as meaning that Ray is obliged to have that belief. For is otiose to say that someone is obliged to do that which they cannot fail to do. And if, for some reason, Ray had failed to form the belief that the carpet is blue, it would not be fair to rebuke him
for this failure, as it must be a result of some physical or cognitive disability. So we might be tempted to think that, given the involuntariness of belief, deontic epistemic judgements are out of place (Alston 1988).

An alternative possibility for what we could mean by saying that anyone in Ray’s position ought to believe that the carpet is blue is just that we would expect anyone in his position to believe that the carpet is blue. We might have instead said, “Anyone with normal colour vision, a normally functioning brain etc., would form the belief that the carpet is blue, if they were in Ray’s position”. If that is what we are getting at, then it doesn’t look as if we are really making a normative claim, but rather a purely descriptive claim. This doesn’t seem like a very good suggestion, however. As Richard Feldman comments (in response to a similar suggestion from Nicholas Wolterstorff (1997)):

Some researchers report that people typically make various unjustified inferences and predictably form unreasonable or erroneous beliefs. Whatever the proper interpretation of this research actually is, it is at least possible that people normally make epistemic errors. It may be that we epistemically ought not do what we normally do (Feldman 2000: 675).

However, instead of assessing agent’s beliefs against the background of what is normal, we could assess them against some ideal standard. It is along these lines that Feldman argues that deontic epistemic claims (claims about what we ought and ought not to believe) can be true, even if we don’t have direct voluntary control over our beliefs:

There are oughts that result from one's playing a certain role or having a certain position. Teachers ought to explain things clearly. Parents ought to take care of their kids. Cyclists ought to move in various ways. Incompetent teachers, incapable parents, and untrained cyclists may be unable to do what they ought to do. Similarly, I'd say, forming beliefs is something people do. That is, we form beliefs in response to our
experiences in the world. Anyone engaged in this activity ought to do it right. In my view, what they ought to do is to follow their evidence (rather than wishes or fears). I suggest that epistemic oughts are of this sort – they describe the right way to play a certain role. Unlike Wolterstorff’s paradigm oughts, these oughts are not based on what’s normal or expected. They are based on what’s good performance. Furthermore, it is plausible to say that the role of a believer is not one that we have any real choice about taking on. It is our plight to be believers. We ought to do it right. It doesn’t matter that in some cases we are unable to do so. (2000: 676)

In a similar vein, Cuneo tells us that our behaviour is assessable against at least two kinds of norms: responsibility norms and propriety norms (Cuneo 2007: 82).

Responsibility norms apply only to actions that are under our voluntary control, but propriety norms can also apply to actions that are not under our voluntary control. An epistemic propriety norm is to be defined in relation to how “an epistemically virtuous agent – one whose cognitive faculties were working well in the appropriate circumstances” (2007: 82) would behave.

Feldman doesn’t quite say what the connection is between his claims that (a) it is our plight to be believers and (b) that we ought to “do it right”. Surely the fact that we cannot help but occupy some role doesn’t mean that we ought to live up to that role to the best of our abilities. Otherwise we would have to say that incurable kleptomaniacs ought to develop the most effective methods for stealing. Nor does Cuneo tell us why we should be interested in emulating the epistemically virtuous agent. What Feldman and Cuneo appear to be suggesting is that saying that S has a reason to believe that p is equivalent to saying that she would be fulfilling the role of a believer better if she believed that p. That does nothing to show that we have a reason to occupy this role as well as we could. Nor does it tell us anything about the nature of the reason that we have to fulfil our ‘epistemic duties’. It seems to me most implausible that our reason to fulfil these ‘duties’ is irreducibly epistemic.
How far should we go to ensure that our epistemic behaviour matches that of
the epistemically virtuous agent? I suppose one might argue that we ought to
develop our cognitive abilities as far as we can in that direction. Maybe we should,
but for what reason? Two reasons spring to mind. Firstly, by doing so we will be
more likely to acquire true beliefs, and true beliefs are often highly valuable to us
when we pursue our goals. So we have a self-interested reason to become more like
the epistemically virtuous agent. Secondly, it might be claimed that we have a
moral reason to ensure that our beliefs are justified. We all rely on each other as
sources of information and it is important that we can trust each other to be reliable
in this respect. So it might be considered a moral failing to fail to do what one can
to be a responsible former of beliefs (Grimm 2009). If we look at the other roles
Feldman mentions, teachers can be assessed against the ideal teacher, parents
against the ideal parent, and cyclists against the ideal cyclist. But anyone’s reason
for being a good teacher, parent or cyclist will be either moral, self-interested, or
other-interested. It might plausibly be claimed that you should be a good teacher,
because it is good to educate children, or because you are more likely to be
promoted if you do your job well, or because you care about the wellbeing of your
pupils. But the mere fact that you are a teacher doesn’t give you a good reason to
perform that role well. If you don’t interact well with children, you may have a
good reason not to perform the role at all, as you would be better off pursuing some
other career. If there is not a peculiarly pedagogical reason to be more like the ideal
teacher, why suppose that there is a peculiarly epistemic reason to be more like the
epistemically virtuous agent? If there is one, then I, for one, need it pointing out to
me.
There is also the following problem with trying to be like the epistemically virtuous agent. Presumably we are not to ascribe to such an agent any particular non-epistemic goals or interests. But regular, non-ideal agents have multiple non-epistemic goals and interests. Given limitations on our time and resources, and given the narrowness of our interests, there is only so much that we can do, and only so much that we are interested in doing, to ensure that our beliefs are well justified. And how we decide how to allocate our time and resources will depend on what our goals and interests are. So, if I am particularly interested in becoming a doctor, I will have a good reason to develop skills in science. If I am interested in becoming a history lecturer, I will have a good reason to develop skills in historical analysis. If I devote most of my time to the study of medicine, I will have much less time to devote to developing skills in historical analysis, and I will become a less responsible believer with respect to history than I might have been. In what ways, then, ought I to try to approximate the epistemically virtuous agent, given that it is not possible to do so in all ways? Presumably I start to answer this question by examining my goals and interests, or my moral responsibilities.

As we saw when we considered Railton’s views about belief, it is plausible that responding to epistemic imperatives is a vital part of what makes us agents. This lends support to Feldman’s claim that being a believer is a role that we cannot help occupying. But we cannot go from these claims to the conclusion that we ought always to abide by all the epistemic imperatives that apply to us, as what justifies our doing so, in any particular instance, would seem to be that it furthers some non-epistemic goal.
Finally, I will examine a pair of arguments from Cuneo against hypotheticalism. This argument focuses on two broad theories of epistemic justification: internalism and externalism. He attempts to show that, if reductionism is true, then neither of these theories can be correct. He understands internalism to be the thesis that “the only sort of thing that can justify a (human) person’s belief is mental states that have their source in sense experience, introspection, memory, broadly logical intuition, and the like” (Cuneo 2007: 192); while externalism is the thesis that things other than an agent’s mental states can justify an agent’s beliefs. For example, a leading form of externalism, reliabilism, has it that justification is a matter of whether our beliefs are produced by reliable processes. Cuneo argues that for each of these theories, we can discern a supervenience thesis, but that if reductionism is true, then neither of these supervenience theses are true. In the case of externalism, Cuneo gives the supervenience thesis as:

**The Epistemic Externalist’s Supervenience Thesis:** The E-justificatory status of a person’s belief strongly supervenes on how it was produced in an appropriate environment: Necessarily, if the beliefs of any two possible individuals are produced in the same way (and if they have no undefeated defeaters for their beliefs) at t, then these beliefs are E-justificatory alike at t, i.e. E-justified to the same extent. (195)

(Where ‘E-Justified’ means ‘externalist-justified’.) The internalist version is:

**The Epistemic Internalist’s Supervenience Thesis:** The I-justificatory status of a person’s belief strongly supervenes on that person’s mental states: Necessarily, if any two possible individuals are mentally alike in the relevant respects at t, then they are I-justificatory
alike at t, i.e. the same beliefs are justified for them to the same extent. (200)

(Where ‘I-justified’ means ‘internalist-justified’.) Assuming that ‘the relevant respects’ referred to in the internalist’s supervenience thesis exclude the goals of agents, these two theses are incompatible with a trio of theses that Cuneo takes to be endorsed by (or to follow from the views of) the reductionist:

**The Application Thesis:** S has an epistemic reason to Φ at t only if and because S has cognitive goal(s) Ψ at t. (189)

(Where ‘cognitive goal’ is defined in this way: “S has [a] cognitive goal at some time … just in case S is, at that time, dispositionally or occurently motivated to implement a doxastic policy whose end concerns the aim of believing what is true” (2007: 189).)  

**The Generation Thesis:** S’s attitude p has epistemic merit X at t if and only if and because p is an appropriate response to epistemic reasons at t. (189)

**The Conditional Claim:** S’s attitude p has epistemic merit X at t only if and because S has cognitive goal(s) Ψ at t. (190)

The Application Thesis is the distinctive claim of the reductionist, the Generation Thesis is supposed to be common ground between the reductionist and the categoricalist, and the Conditional Claim is entailed by the conjunction of the other two theses. With these theses in place, it isn’t going to be very difficult to tell the following pair of stories; the first designed to show how reductionism clashes with externalism, the second designed to show how it clashes with internalism.

---

37 I’m not certain what work the “and because” is doing in this and the other theses. In ignoring it, I am presumably discussing weaker theses than Cuneo intends.  
38 This is a narrow definition of ‘cognitive goal’. If cognitive goals are just defined as aiming at forming beliefs, then not all of these goals need be aimed at forming true beliefs, as Cuneo acknowledges (Cuneo 2007: 188-9).
In the first story there is an agent $S$ at a world $W$ who forms a particular belief $B$ which is produced by process $P$ in environment $E$, and in forming that belief, $S$ advances some cognitive goal $G$. In a different world $W^*$, $S$’s counterpart $S^*$ forms the same belief $B$ by the same process $P$ in the same environment $E$, but $S^*$ doesn’t have any cognitive goal that is advanced by her forming that belief. By the Externalist’s Supervenience Thesis, $S$ and $S^*$ are justified in forming $B$ to the same extent. But by the reductionist’s Conditional Claim, $S^*$’s belief cannot have any epistemic merit, and so cannot be justified to any extent, because she lacks any relevant cognitive goal. Thus reductionism is inconsistent with externalism (195-199).

In the second story there is an agent $S$ at a world $W$ who forms a particular belief $B$ while in mental state $M$ (where $M$ includes only the sort of mental states which, for an internalist, can justify belief), and in forming this belief, $S$ advances her cognitive goal $G$. At another world $W^*$, $S$’s counterpart $S^*$, forms an identical belief $B$, whilst in an identical mental state $M$, but $S^*$ has no cognitive goal that is advanced by her forming this belief. By the Internalist’s Supervenience Thesis, $S$ and $S^*$ are justified to the same extent in having belief $B$. But by the reductionist’s Conditional Claim, $S^*$’s belief cannot be justified to any extent, because she lacks any relevant cognitive goal. Thus reductionism is inconsistent with internalism (199-204).

I don’t think Cuneo’s argument that reductionism is incompatible with externalism is successful. This is because it does not seem that the Generation Thesis should be endorsed by externalists. At least from a reliabilist perspective, an agent’s beliefs do not acquire merit or demerit – i.e. are not justified or unjustified – according to how well she responds to reasons, but on how reliable the processes
are by which the belief is formed. Note that it is a consequence of reliabilism that two agents who are psychologically identical may differ greatly in how far their beliefs are justified (Steup 2005). A brain in a vat being fed an ersatz sensory input by an evil scientist may have an enormous number of false beliefs, which lack any justification, on a reliabilist model, as they are produced by extremely unreliable processes, whereas another individual might have qualitatively identical beliefs that are extremely well justified, based as they are on extremely reliable processes (he is not a brain in a vat, but a regular person with a well-functioning set of senses). It is difficult to make sense of the idea that, concerning these two individuals and from a reliabilist perspective, the brain in the vat is failing to respond appropriately to reasons while the regular person is successfully responding appropriately to reasons. Granting the possibility that brains in vats can have empirical beliefs at all, it would seem that there is nothing that either of these individuals can do to ensure that their empirical beliefs are better justified – it is merely bad luck that the brain in the vat lacks any justified empirical beliefs, and good luck that the regular person has many justified empirical beliefs. Are we supposed to equate being the fortunate recipient of reliably processed sensory input with responding appropriately to reasons? We cannot equate these things, and in the light of this, I would suggest that when we are considering the Generation Thesis in relation to reliabilism, we should have an alternate version of it that reflects this:

**The Generation Thesis (Reliabilist Version):** S’s attitude p has epistemic merit X at t if and only if and because *p has been produced by reliable processes.*
But, of course, the conjunction of Generation Thesis (Reliabilist Version) and the Application Thesis do not entail the Conditional Claim. With this alternate version of the Generation Thesis available, the reductionist can claim that, from a reliabilist perspective, epistemic reasons and epistemic merit can be decoupled. It is possible to assess how well an agent’s beliefs are justified without considering whether they have any reason to have those beliefs.\textsuperscript{39}

If epistemic merit and epistemic reasons can come apart in this way, then why not moral merit and moral reasons? If moral merit and moral reasons can come apart then we would be able to defend moral naturalism. The response to this is that moral naturalism, as we saw in the previous chapter (1.2.4), equates moral motivation with self-interested motivation, but these cannot be equated. There seems to be no corresponding problem with equating ‘epistemic motivation’ – i.e. the motivation to inquire, check one’s facts and logic etc. – with self-interested motivation. This is because we would seem to be behaving in an epistemically responsible way by doing such things, even if our motivation for doing so was self-interested. By contrast, we don’t deserve to be called moral if we are only looking out for ourselves.

What about the argument that reductionism is incompatible with internalism? This may seem to be more persuasive, as it looks less obvious how we could amend the Generation Thesis with respect to internalism (I don’t think we can amend the Application Thesis). For internalist justification it is not enough that I have a particular belief $B$ and a particular set of mental states $M$ that justify that belief, for my belief to be justified. In addition, I must believe $B$ because of $M$. If I haven’t

\textsuperscript{39} Railton agrees: “The epistemic warrant of an individual’s belief may be disentangled from the rationality of his holding it, for epistemic warrant may be tied to an external criterion – as it is for example by causal or reliabilist theories of knowledge” (Railton 2003: 9 [Facts, Values, Norms]). If epistemic merit and epistemic reasons can come apart in this way, then why not moral merit and moral reasons?
considered $M$ in forming $B$, but have just acquired $B$ through some epistemically non-respectable route, such as prejudice, then my belief will not be justified. My belief must be formed as an appropriate response to my evidence if my belief is to count as justified from an internalist perspective. These considerations make it look as if that the Generation Thesis is sound when it comes to internalist justification. However, if reasons for belief are being equated with evidence here, we can rephrase the Generation Thesis in the following way:

**The Generation Thesis (Internalist Version):** S’s attitude $p$ has epistemic merit $X$ at $t$ if and only if and because $p$ is an appropriate response to the evidence available at $t$.

Now it is open to the reductionist to deny that evidence ought to be equated with reasons for belief in this way. To be sure, evidence may give us reasons to believe when it is combined with an appropriate desire, but the reductionist may claim that evidence for $p$ is not, in itself, equivalent to a reason to believe $p$, irrespective of anyone’s desires. In everyday English ‘evidence for’ and ‘reason to believe’ are used interchangeably, but this can be accounted for by noticing that when we are sufficiently interested in the truth of some proposition, we will also be interested in which items are and which are not good evidence in support of that proposition. We can then think of items of evidence not as reasons to believe, but as indicators: $q$ is evidence that $p$ just in case $q$ reliably indicates that $p$ (Olson 2011c).

If we have the notion of ‘indicator evidence’ in play, then we can see how a reductionist can still be an internalist. Following the Generation Thesis (Internalist Version), S’s attitude $p$ will be justified if and only if it is an appropriate response
to the available evidence at t (i.e. if the evidence available at t really does indicate p). And this can be so even if there is no reason for S to have that belief.

2.4.6 Summary of section 2.4

Epistemic reasons appear at first sight to be categorical, because we would not withdraw our claim that someone was being epistemically negligent if they failed to abide by some epistemic imperative that applied to them. However, once we notice that categorical imperatives differ from hypothetical imperatives not only in the fact that their application to us is not contingent on our desires, but our reason to abide by them is not contingent on our desires, we can recognise a third category of imperative – the non-hypothetical imperative – that is like a categorical imperative with respect to its application, but like a hypothetical imperative with respect to its reason giving force. If we do not recognise the possibility of non-hypothetical imperatives, then we will be easily persuaded (by examples like Kelly’s) that epistemic reasons are categorical. But once we introduce the notion of a non-hypothetical imperative, the onus is on the categoricalist to show us that we always have a reason to abide by the epistemic imperatives that apply to us, and that that reason is irreducibly epistemic. Throughout this section I have attempted to show that (1) given our limited time and resources, we do not always have a reason to abide by the epistemic imperatives that apply to us; and (2) when we do have a reason to abide by them, it is always for some non-epistemic reason.

I also looked at a pair of arguments from Cuneo, designed to show that epistemic reductionism is incompatible with two dominant theories of epistemic justification: internalism and externalism. However, I showed how the reductionist
could reject both of these arguments as she is not obliged to accept Cuneo’s rendering of the Generation Thesis.

So, at the end of this discussion it looks as though the alleged categorical nature of epistemic reasons has not been demonstrated, and the error theorist has nothing to fear from the epistemic argument.

2.5 Partners in guilt arguments (2): The argument from the categorical requirements of practical reasoning

Another way of arguing that error theorists are committed to categorical imperatives is by claiming that our reasons for abiding by hypothetical imperatives are not wholly contingent on our desires. Rather, it has been claimed, the fact that it would be rational to perform the action commanded by the imperative gives us a further, non-desire based reason to perform it. However, I intend to show that the rationality of an action does not give us an irreducible reason to perform it.

I claimed in the previous section that hypothetical imperatives apply to us, and that we have a reason to abide by them, only if we have some desire that is satisfied by our being guided by them. This seems to be the natural way to think about a claim such as “You ought to do some exercise”. It is because I want to be healthy, and physical exercise is essential for health, that the imperative *do some exercise* applies to me, and I have a reason to abide by it. This simple view of hypothetical imperatives has been challenged, however. According to Jean Hampton (1998) and Christine Korsgaard (2008), my reason for abiding by any hypothetical imperative does not wholly depend on the fact that, by acting in accordance with it, I satisfy some desire. I am criticisable for failing to abide by a hypothetical imperative that applies to me, not because (or not only because) I fail to get something that I desire, but because I manifest irrationality. So there is a
reason to abide by a hypothetical reason that applies to me which is not grounded by any of my desires. And reasons that are not grounded by desires are categorical.

One way to get to the heart of this challenge is to consider that there must be at least one rule that it is necessary to follow if one is to be considered rational. Let me explain why. We prove ourselves to be rational or irrational by the way that we act. Moreover, it seems as though we have a choice over whether to act rationally or irrationally. (Doubts about the existence of free will might make us worry that this choice is illusory, but let us put such doubts aside.) Now, as we are judged rational or irrational by our choice of actions, there must be some way of distinguishing between rational and irrational actions. We cannot usually do this just by looking at a description of the action itself. Some may hold that there are types of actions that are always irrational, but usually a bare description of the action is not enough. For instance, if we are just told that someone has cut off their own, healthy arm with a penknife, we might think they were crazy. But if we are then told that the arm was stuck in a crevice of a rock, in some isolated spot, and the person would have died of thirst if they had not taken the desperate measure, then we will be inclined to revise our initial reaction. It seems rational to sacrifice a limb to avoid death by thirst.

If the mere description of an action is not enough for us to determine whether it was rational or irrational, then it must be something about the way we chose how to act. That suggests that there is a way of deciding which, amongst a range of alternative actions, is the most rational one to take – a *rational principle*, which if followed will lead to rational action, and which if departed from will lead to irrational action. What form will this principle take? Our first thought might be that it is rational to make the choice that secures the greatest good. We needn’t think of
‘the greatest good’ as an objective matter, but only what seems to be most good from the point of view of the person making the choice. For most people, a long life, minus one limb, will seem preferable to a much shorter life in which their body stays intact, so it seems to make sense to sacrifice the limb for the sake of a longer life. When we factor in the fact that we have to remove the limb ourselves, with a penknife, and without any pain-relief, we might be less sure that it would be irrational to forgo the operation. So let us borrow a similar, but more mundane example from Korsgaard, in which declining to take the means to secure the greatest good seems more clearly irrational:

Howard, who is in his thirties, needs medical treatment: specifically, he must have a course of injections, now, if he is going to live past fifty. But Howard refuses to have this treatment, because he has a horror of injections. Let me just stipulate that, were it not for his horror of injections, Howard would have the treatment. It’s not that he really secretly wants to die young anyway, or anything fancy like that. Howard’s horror of injections is really what is motivating him. (Korsgaard 2008: 39)

Howard seems irrational because a prolonged life, secured by a course of injections, seems preferable to a much shorter life without injections, and having injections is not really something he should be frightened of (unlike, say, auto-amputation). If Howard agrees with this, but still refuses his treatment, he prefers a course of action that results in a lesser good (short life, but no injections) to one which results in a greater good (much longer life, involving injections). Now Hume famously claimed that it is not

… contrary to reason to prefer even my own acknowledged lesser good to my greater, and have a more ardent affection for the former than the latter. A trivial good may, from certain circumstances, produce a desire superior to what arises from the greatest and most valuable enjoyment … (Hume 1739: 416)
Hume thinks this is the case because, in his view, we are moved to act solely by our passions, and neither our passions nor the actions resulting from them, can be rationally assessed:

Reason is the discovery of truth or falsehood. Truth or falsehood consists in an agreement or disagreement either to the real relations of ideas, or to real existence and matter of fact. Whatever, therefore, is not susceptible of this agreement or disagreement, is incapable of being true or false, and can never be the object of our reason. Now 'tis evident our passions, volitions, and actions, are not susceptible of any such agreement or disagreement; being original facts and realities, compleat in themselves, and implying no reference to other passions, volitions, and actions. 'Tis impossible, therefore, they can be pronounced either true or false, and be either contrary or conformable to reason. (Hume 1739: 458)

It seems clear that what Hume means by “conformable to reason” is what we would mean by “rational”. If we follow Hume, then, we will have to think that we are mistaken to think that Howard – or anyone else – is irrational. But it is hard to accept this conclusion. It seems apt to describe Howard as irrational for giving up the chance of a decently long life for the sake of avoiding a minor medical procedure.

However, I think we can preserve the fact that our desires ground our reasons, without giving up on the idea that there is such a thing as irrational behaviour. To do this we need to demonstrate two things. Firstly, we have to show is that there is at least one rational principle that can be used to determine whether or not an action is rational. Anyone who flouts this principle is rightly described as irrational. Secondly, we need to show that our reason to abide by this principle is wholly given by the fact that, by abiding by it, we satisfy some desire.
I have suggested that a rational course of action is one that secures more good than alternative actions, so let us try to derive a principle from this observation. Howard seems irrational because, although he wants to live a good long life, his fear of needles makes him refuse to take the necessary means to that end. It can sometimes be rational to give up on a goal once you realize what you will have to do to secure that goal, but this will be because you come to realize that the end is not worth the means. This doesn’t seem to be one of those cases. It seems that, even to Howard, his greater good would be secured by having the treatment he needs; but what he chooses to do is to forgo that treatment. It is in refusing to take the means to his self-acknowledged greater good that Howard manifests irrationality. And it seems that any agent who refused to take the means to their self-acknowledged greater good deserves to be called irrational.

So we could say that our rational principle is *take the means to your self-acknowledged greater good*. Alternatively, we could think of our ‘self-acknowledged greater good’ as our (true) end, i.e. what we all really want to pursue. So we could formulate our principle as *take the means to your ends*. Let us call this the ‘instrumental imperative’.

It might be objected that what an agent *perceives* to be their greater good may not in fact be their greater good, and that the rational thing to do would be to pursue their actual greater good. I want to avoid getting into a lengthy discussion on this matter. There seems a clear sense in which we can assess the rationality of someone’s behaviour *given how things seem to that person*, even if our assessment of what is rational *from that person’s perspective* would differ from our assessment of what is rational from a more objective perspective. Here we just focussing on what is sometimes called ‘subjective rationality’.
With the instrumental imperative in play, we have a way of distinguishing (subjectively) rational behaviour from (subjectively) irrational behaviour. That was the first part of our task. To complete our second task, we need to ask about the status of the instrumental imperative. Is it hypothetical, non-hypothetical, or categorical? I think we can quickly dismiss the idea that it is hypothetical. If it was hypothetical, then it would apply to me only if I had some end that would be promoted by abiding by it. So we could make the hypothetical nature of the imperative clear by saying, “If you have some end that would be promoted by taking the means to your ends then take the means to your ends”. The problem with this suggestion is that all ends would be promoted if we took the means to them. When we try to construe the imperative as hypothetical it reveals itself not to be.

So perhaps we should say that the instrumental imperative is categorical, i.e. an imperative that applies to all of us and that we have reason to abide by, irrespective of our contingent desires. Now it does seem that the instrumental imperative applies to all of us. This is so because anyone’s actions can be assessed as rational or irrational. But as we observed in the previous section, this is not enough to show that the imperative is categorical. Anyone’s behaviour can be assessed as polite or not from the point of view of Japanese etiquette, but we do not think the rules of Japanese etiquette are categorical imperatives, as we don’t always have a reason to abide by them. The question we need to ask is not, do we all have a reason to abide by the instrumental imperative, whatever our contingent desires? That might be the case because we all have some desire or other that grounds a reason to abide by the instrumental imperative. Rather we should ask whether

---

40 This, I think, is the gist of Schroeder’s response to the Hampton/Korsgaard challenge (Schroeder 2007). He discusses the challenge in chapter 3 of his book.
there is a reason to abide by the imperative over and above the fact that, by abiding by it, we achieve some goal or other.

What we are asking, in effect, is whether the reason to follow the instrumental imperative is like the reason to follow moral imperatives. As we saw in the previous chapter, we cannot be considered to have acted morally unless we were motivated to act (at least in part) by the thought that the action was right (and not merely by the fact that it satisfied some self- or other-interested desire). Similarly, if the instrumental imperative is categorical, then that will mean that we cannot be considered to have acted rationally unless we were motivated (at least in part) by the thought that the action was rational.

Hampton claims that for an action of mine to be considered rational, I must do it because it is rational. To use one of her examples, suppose that your friend Sally has a sore throat and you tell her she should go to the doctor. What you mean to imply is that it would be rational for her to go to the doctor, given her end of good health. But Sally doesn’t register this. Instead your words remind her that her doctor has done very well in his investments, and this causes her to desire to see the doctor in order to get some investment advice. And it is this desire that moves her to visit her doctor. Hampton comments:

Your hypothetical imperative was causally efficacious in generating a motive in Sally to see the doctor. But it is not true that she acted from the reason given her by the hypothetical imperative. That is, your imperative did not give Sally a reason, which, by virtue of its authority, created a motive in her to go to her doctor. Instead, the words you uttered conjured up ideas that initiated a line of thought that, as it happens (and it could have happened otherwise), eventuated in the generation of a desire to perform the action your imperative directed. (Hampton 1998: 158)
All this shows is that Sally did not act on the imperative that you gave her, even though, by chance, the words you said got her to do the rational thing. But if we are to convict Sally of irrationality here, then isn’t it because she went to see the doctor for financial rather than medical advice, when what she most needed, and what her doctor was best placed to give her, was medical advice? Is she even less rational because the imperative you utter fails to motivate her?

Consider Howard again. Is it any part of his reason to undergo his treatment that, if he does so, he will manifest rationality? One would have thought that his reason to undergo the injections would be to prolong his life, not to manifest rationality. It is true that we might sometimes be motivated to act in a certain way by the thought that, in so acting, we will manifest rationality. An employee desirous of promotion may want to display to her boss that she always takes the best means to her professional ends. But her ultimate motivation is clearly that, in so acting, she will impress the boss and gain promotion. It is not impossible that some people are motivated to manifest rationality at all times, and that this end is as precious to them as the end of good health or financial security is to other people. It is doubtful that this is a widely held end, however, and it doesn’t seem necessary to have this end if we are to be considered to (sometimes) act rationally.

What we should say, then, is that the instrumental imperative is non-hypothetical. It applies to all of us, but our reason to abide by it is always that, by so doing, we get more of what we value. If we always have a reason to abide by it, that is because, by so doing, it will always get us more of what we value. The conditions for rational action are not the same as the conditions for moral actions. I act irrationally if I fail to abide by the instrumental imperative, not if I fail to be motivated by it. With moral action, by contrast, I must not only act in accordance
with the moral imperatives that apply to me to be considered to have acted morally; I must in addition be motivated to act by the concern to do what is right.

Both of the partners in guilt arguments that we have considered attempt to demonstrate that error theorists are committed to categorical imperatives, despite their failure to notice this. These arguments proceed by demonstrating that there are imperatives that apply to all of us, irrespective of our goals or desires, and conclude from this that those imperatives are categorical. But once we notice the possibility of non-hypothetical imperatives, it becomes clear that the proponents of these arguments have more work to do. They must show that we all have a reason to abide by these imperatives, irrespective of our goals or desires, otherwise the imperatives are not categorical, but merely non-hypothetical. When we examine matters, however, it seems that our reason to abide by epistemic imperatives and the instrumental imperative is just that, in so doing, we acquire things that we contingently desire. That an action would be epistemically praiseworthy, or that it would manifest rationality, is rarely our ultimate reason for doing it. And our action does not fail to be epistemically praiseworthy or rational because we fail to do it for the sake of epistemic or instrumental imperatives.

2.6 Conclusion
For many people, the error theorist’s thesis that there are no true moral statements is highly counterintuitive. This has led a number of philosophers to claim that error theorist’s arguments are, in the end, incoherent: they commit us to things that error theorists specifically say we should not be committed to. The self-undermining argument is designed to show that the error theorist is committed to true moral propositions of the form “x is permissible”. But it was fairly easy to show both why
this wasn’t the case, and why we must be able to talk about first-order moral claims without making any of our own. The partners in guilt arguments were designed to show that error theorists are committed to categorical imperatives. But all they actually show us that we are committed to non-hypothetical imperatives. This shows that error theory is no incoherent in the ways suggested by these arguments.

The other two arguments we considered in this chapter trade on our confidence in the existence of moral facts, or our confidence in their possibility. The commonsense argument was unpersuasive because it seems reasonable to appeal to commonsense to reject the conclusion of an argument, despite an inability to fault any of its premises, only when we have strong independent evidence for the negation of the conclusion. Merely appealing to confidence in your moral opinions is not enough. And Coons’s modal argument was shown to be unsound, as various arguments for error theory entail that moral facts are impossible at worlds like our own. This shows that appeals to confidence in either the existence or the possibility of moral facts will not be effective strategies for opponents of error theory.
3: Eliminativism and Relativism

In the preceding chapters we looked at the arguments for error theory and at how error theorists can respond to various objections. It is time now to consider a different matter. What, if anything, should we do with moral discourse and practice if we are persuaded by the error theory? By ‘moral practice’ I mean such activities as making moral judgements about the behaviour of other people; treating people differentially because of their perceived moral character; creating, implementing and defending social policies on moral grounds; blaming, punishing, praising or rewarding people on the grounds of the moral qualities of their actions; and assessing one’s own past behaviour and deliberating about one’s future behaviour in moral terms. One can ask both “How, if at all, should I alter my own attitude to moral discourse and practice?” and “How, if at all, should I attempt to alter the attitudes of others?”

These might appear to be peculiar questions, for the following reason. If I am just going to answer these questions for myself, accepting that others might reasonably decide to answer them differently for themselves, then I cannot expect the answers I give to be of any philosophical interest. But if I give these questions a more general form – How should we alter our attitudes to moral discourse and practice if we are persuaded by error theory? – then I would appear to be presupposing that there are general answers to questions that are obviously normative. One of the possible answers to the question is: “Stop engaging in moral discourse and practice”. Another is: “Change the way in which we use moral terms, in order to eschew unwanted ontological commitments”. But what right could I claim to cast judgement on how everyone ought to act, if error theory is true? Isn’t
it a consequence of error theory that there isn’t any way that people, taken collectively, ought to act? It’s important to see that it isn’t.

All that the error theorist claims is that there are no ways that people, taken collectively, morally ought to act. However, it is consistent with this that there are ways in which everyone ought to act. For instance, it sounds plausible to say that everyone has a self-interested reason to avoid unbearable pain being inflicted upon them. Even if that is not actually true, it is consistent with error theory being true. In the same way, it might be true that everyone ought to stop engaging in moral discourse and practice, for self-interested reasons, if error theory is true. If it’s not true that everyone has a self-interested reason to avoid unbearable pain, it is surely true that almost everyone has this reason. And judgements about what almost everyone has a reason to do can be interesting. For instance, current medical evidence suggests that almost all mothers have a good, other-interested reason to breastfeed their babies. For medical reasons, it may be inadvisable for some mothers to do so, but that doesn’t detract from the importance of the advice to most mothers. Of course, we cannot guarantee that the question “What should we do with moral discourse and practice if error theory is true?” will have either a completely general answer, or an answer that applies to almost everyone. But I think there are the following grounds for thinking that the right answer to give to this question will have a high degree of generality: (1) there aren’t too many answers to the question which sound even prima facie plausible; (2) of the answers that sound prima facie plausible, not all will turn out to be coherent upon reflection, or practically implementable. Thus by thinking through the options, we can narrow the range of possibilities.
In the next section I am going to lay out some options for what to do with moral discourse if you are convinced of error theory. Then in section 3.2 I shall look in more detail at the option of purging moral discourse, which I will call ‘eliminativism’. I consider two main motivations for eliminativism. The first is that morality is actually harmful to most individuals, partly because it fosters inegalitarian social structures, and partly because it is emotionally damaging (3.2.1). These claims have been made most forcefully by Ian Hinckfuss (1987). I argue that while it is not difficult to find examples of repressive moral practices, morality is not inevitably repressive. Indeed, it can be spur to liberation. I also argue that while negative emotions such as guilt and shame may only be experienced by persons with a moral outlook, it is not worth abandoning that outlook to rid oneself of the propensity to experience these painful emotions (if that is even possible). People who lack the capacity for guilt and shame are not the sort of people that most of us would want to emulate.

The second motivation I consider for eliminativism is the claim that it is always better, from a practical point of view, to have true beliefs than false ones (3.2.2). Thus if we are able to replace falsehoods for truths, we ought to do so. I argue that it is not the case that it always practically preferable to have true beliefs over false ones. In fact, there is evidence from psychology that this is not the case (Taylor and Brown 1988).

In section 3.3 I turn to consider whether error theorists could decide to become relativists about moral discourse. I argue that relativism is inherently unstable, as there will always be pressure to universalize our moral opinions.
3.1 Some Options

The most obvious option is just to purge moral discourse, to no longer make any moral claims. I shall call this position ‘eliminativism’. I suppose one might decide to abandon moral discourse – which only involves refraining from speaking and writing in moral terms – without committing to abandoning moral practice entirely. One might, for instance, be pessimistic about the possibility of entirely ridding oneself of the propensity to evaluate human behaviour under moral concepts. But I take it to be more likely that a defender of eliminativism would aspire to eliminate not just their habit of expressing themselves in moral terms, but also of thinking and behaving as if certain moral claims were true. (Of course, it may be expedient for an eliminativist to sometimes behave as if certain moral judgements were true, just as it may occasionally be expedient for a moral believer to behave as if a moral judgement that she did not believe to be true were true.) If we take this option we will have reason to retain moral vocabulary only insofar as we have need to talk about the beliefs and attitudes of those who remain under morality’s spell, or about our own benighted past, just as we retain terms for witches and other supernatural entities that we no longer believe in.

Another option is to retain moral discourse not just for the purpose of talking about morality, but for some other practical purpose. Machiavellian types might consider it wise to continue to use moral discourse as if they sincerely believed in it, as a way of manipulating others to their advantage. This option would not involve the retaining of moral practice, although it would involve retaining the appearance of moral practice. Less sinisterly, some might fear that if morality was abandoned the world would plunge into an egoistic chaos, a Hobbesian war of all against all. This might motivate one to suppress error theory, for the good of
humanity. Following Terence Cuneo and Sean Christy (2011), I will call this option ‘propagandism’. Even less sinisterly, it may not be necessary to keep anyone in the dark about error theory. Perhaps moral thinking is so ingrained that even being convinced of the truth of error theory will not stop people continuing to have moral beliefs. Following Jonas Olson, I will call this option of letting our natural propensity to think in moral terms prevail, ‘conservationism’. Whether this is a good thing or a bad thing is something I take up in the following chapter.

Conservationism has the consequence that we end up believing a lot of untruths. But there may be a way for us to reap the benefits of moral discourse and practice without incurring this cost. It may be that moral discourse can perform useful functions whether or not we actually believe the moral judgements we are apt to make. Revolutionary fictionalists argue that we can amend our attitude to such judgements, so that we no longer believe their literal content. This will be worth doing, they argue, because morality can serve a useful function in our lives, whether or not it is believed. For instance, Richard Joyce argues that thinking in moral terms makes us more likely to behave prudently than non-moralized prudential thinking (Joyce 2001); and Daniel Nolan, Greg Restall, and Caroline West argue that there are things that we want to say about the non-moral aspects of our world that cannot easily be said without using moral vocabulary (Nolan et al. 2005). I consider their arguments in the following chapter.

At least two further revisionary options are available. One of these is revolutionary expressivism. The revolutionary expressivist is an error theorist who, as such, thinks that moral discourse is assertoric, but thinks that we should transform our practice in such a way that our moral judgements become expressions of our attitudes, rather than assertions of propositions. Mark Schroeder
suggests that Mackie can be interpreted as motivating a revisionary semantics for moral discourse, in the second part of *Ethics* “…since he continues to theorize about morality, even after having claimed to establish that moral claims are systematically false” (Schroeder 2007: 119, n. 24). This, he continues, “…might be one of the better ways of motivating expressivism. A revisionary expressivism wouldn’t have to accept the general semantic commitments necessary to defend expressivism against the Frege-Geach problem and other problems”. It is a bit odd of Schroeder to say that Mackie can be interpreted as motivating a revisionary semantics for moral discourse, as he makes no explicit attempt to do so. However, it is true that the second part of *Ethics* would have benefited from some discussion of how we are to understand the moral vocabulary that Mackie advocates we continue to use even after we have been convinced of error theory. Revolutionary expressivism is a possibility.

Revisionary semantics need not be expressivistic. We could retain a standard semantics by amending the meaning of moral terms, so that they did successfully refer. I’ll call this last option ‘reformist realism’. The reformist realist agrees with the error theorist that there is something amiss with our moral concepts as we currently employ them, but suggests that we can make a small modification to the concepts so that the modified concepts can be used without embarrassment. There is an extant metaethical thesis that might be thought of as a kind of reformist realism, namely metaethical relativism. Metaethical relativism did not emerge as a response to moral error theory (the former predates the latter), but it is possible that one persuaded by error theoretic concerns might think that we could salvage moral discourse by turning to relativism.

The options I have mentioned are:
I’m not going to discuss propagandism separately, although I shall address it briefly within my discussion of eliminativism. Revolutionary expressivism is potentially interesting, but I’ll pass on the opportunity to develop it here. In this chapter I will discuss eliminativism and reformist realism (in the form of relativism). As I have already indicated, I don’t think that either of these options are attractive. In the following chapter I will discuss revolutionary fictionalism and conservationism.

### 3.2 Eliminativism

The eliminativist is, at the very least, someone who proposes to stop using moral language to express moral judgements. However hard that might be to implement in practice, it is clearly possible. A more thoroughgoing eliminativist would also want to stop making (and being influenced by) unvoiced moral judgements. As I presume that anyone attracted to eliminativism would want to be a thoroughgoing eliminativist, that is the position that I will discuss here.

It *sounds* reasonable to say that if all positive moral judgements are false, it would behove us all not to make them, on the grounds that it is always best, if possible, to avoid making false judgements. It is a natural first thought that it is
always in our interests to make accurate judgements about matters that concern us. At the same time, a cognitive policy of always avoiding false judgements where possible is not the best policy to have. There are some judgements which, though false, are not in themselves harmful to our interests, while there may be other false judgements that are beneficial to make. If it is true, for example, that religious people are, on average, happier than atheists, then this must be counted as something in favour of having religious beliefs. And given finite time and energy it would impede our interests to do all that we could to ensure that our judgements were as sound as they possibly could be, as this would be at a the cost of spending time and energy on other important matters.

It might be thought, however, that once someone has been convinced of error theory they will quite automatically stop making moral judgements. That sounds naive to me. Most of us were taught morality from earliest childhood, and have had this training reinforced by the moral presumptions implicit in our culture: in art, in politics, in our personal lives, and perhaps in our religious practice. There is no guarantee that someone who accepts error theory when she is thinking about metaethics won’t slip back into moralizing at other times. An eliminativist might congratulate herself for consciously avoiding expressing, in word or thought, positive moral judgements. But it might reasonably be asked whether she is still motivated by characteristically moral patterns of thinking. Those authors who have argued in favour of eliminativism have always, at the same time, been eager to emphasize that they are not ogres or rampant egoists, but have concerns for the welfare of humanity at large, which, in other contexts, could easily be mistaken for moral concern.\footnote{Shafer-Landau makes this point in reference to Mackie and Richard Garner (Shafer-Landau 2003: 66)} Their defence would no doubt be that their concerns merely
reflect their subjective valuing of humanity, which does not (by an error theorist’s lights) equate to moral concern. We might be tempted to give the eliminativist the benefit of the doubt here. However, there may be less direct, and more revealing, ways of discerning whether someone harbours moral patterns of thought than simply asking them their own opinion about the contents of their judgements.

One way we might do this is by considering whether their emotional repertoire is affected by their decision to eliminate moral patterns of thinking. If self-proclaimed eliminativists continue to experience, or even worry about experiencing, shame or guilt – which are often thought of as moral emotions – this may be good evidence that they have not eliminated moral thought. It seems likely that we could not, without taking drastic measures, rid ourselves of the propensity for shame or guilt, and it seems right to me to say that these emotions somehow involve the thought that one has done wrong. However, arguing this would probably require us to endorse a cognitivist theory of the emotions, where emotions such as guilt and shame are at least partly constituted by the judgement that one has done wrong. I am not averse to endorsing a cognitivist theory of the emotions, but I would prefer to argue against eliminativism on less controversial grounds.

Another way to show that eliminativists had not rid themselves of moral thought would be to examine their vocabulary. While it might, perhaps, be relatively easy to avoid thin terms, such as the moral senses of ‘good’, ‘bad’, ‘right’, and ‘wrong’, there are a whole host of thick moral terms, the use of which arguably reveals the speaker’s moral outlook. I am thinking of such terms as fair, just, cruel, kind, considerate, generous, miserly, cowardly and courageous. Judgements involving these terms often seem to be judgements about moral rights and duties. We treat people fairly or justly when we give them what is due to them.
We are kind or generous when we exceed our duties to other people. And we are miserly or cowardly when we fail in our duties to others. It would be difficult to do away with these terms, but if the eliminativist does not merely use them to discuss the judgements of others, but to express her own judgements about the character or behaviour of others, then we might suspect that she is deceiving herself if she thinks she has rid herself of moral thought. I am not sure how effective this would be. The eliminativist might continue to use these terms, not to express thoughts about moral rights and duties, but about rights and duties that arise out of conventional social arrangements, which she does not conceive of as being moral. She might think of fairness and justice as just a matter of treating people according to conventional social norms; kindness and generosity as a matter of doing more than those norms require; and cowardice and miserliness a matter of failing to meet certain social norms.

I would rather steer clear of trying to demonstrate the difficulty of becoming an eliminativist and of finding ways to trip them up. I think it is better to focus on whether eliminativism is attractive, whether or not it is psychologically plausible that we could become eliminativists. It is to this matter I turn to now.

### 3.2.1 Is morality harmful?

The most forceful presentation of the view that morality is harmful is given by Ian Hinckfuss in *The Moral Society: Its Structure and Effects*. He writes:

Few who are reading this will disagree that they live in a moral society. Few will disagree that the society that they live in is elitist, authoritarian, intellectually dishonest in its social decisions, lacking in esteem for most of its members, inefficient in the resolution of its conflicts, inefficient in maximising human happiness, satisfaction or
self-esteem, and because of the threat of war with other societies, physically dangerous. (Hinkfuss 1987: 20)

Even if we agreed with Hinckfuss’s bleak appraisal of contemporary societies (and societies clearly differ markedly in the extent to which they display these qualities) it would not follow that they have these qualities because they are moral societies. To be fair to Hinckfuss, he is not so crude as to suggest that the amoral society which he advocates would necessarily be a better place (better, that is, in being less elitist, authoritarian etc., and better at maximising the happiness, satisfaction and self-esteem of its members). But he does try to show that morality plays an important part in introducing and maintaining the bad qualities he highlights.

He paints a vivid picture of how morality fosters an unequal, hierarchical society. In a moral society, he writes:

[c]hildren will vary in the way they react to condemnation and praise, and the quantities of condemnation and of praise will vary from child to child. One child will be held up as an example – good or bad – thus giving some children a moral boost at the expense of others. The children who receive most moral boosts from their parents or guardians are likely to believe what they are continually told, namely that they are very good. These will be the children who succeed in pleasing their moral mentors most. Other children get the inverse treatment and go into adulthood with an inferiority complex and a tendency to seek continual moral guidance and leadership from their ‘betters’. Most people end up somewhere on the spectrum in between.

But those who are convinced of their own goodness will be those most likely to become the moral leaders of society. In fact such moral self-confidence is a necessary condition for entry into the moral elite. For with such self-confidence, it is easy to believe that what one wishes for oneself is morally permissible, and how one wants others to behave is morally obligatory. A good person will not want what is wrong. (27)

To complete the picture, the moral elite will not only enjoy telling all the rest of us what to do, they will also reap material and political rewards:
…if morality is to keep going, the moral carrot and stick must be displayed or applied continually to most people throughout their lives. The punishments include frowns, snubs, deprivations of income, deprivation of possessions, imprisonment and physical violence. The rewards include smiles, honours, property, economic security, power and privilege. This is the system of moral deserts. Again it is the trainers, not the trainees, who determine who deserves what. (28)

And the non-elite are supposed to acquiesce in this “because they will be trained to assent to the proposition that people should get what they deserve – and of course better people deserve more” (28)

Hinckfuss raises a number of other criticisms of the moral society. He tells us that moral societies damage their members psychologically, for in an amoral society “there is no possibility of feelings of guilt, guilt complexes or moral inferiority complexes with all the sadness, madness and suffering that these feelings and complexes entail” (40). And he says that morality frustrates, rather than aids us, in resolving conflicts – indeed, it fuels those very conflicts. After reminding us of the blood shed in Ireland, Lebanon, Palestine, Vietnam and in the two World Wars he comments:

Think of any one of these conflicts and 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. 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. (45-6)

If Hinckfuss is right, then he seems to have a strong case that morality is harmful to most of us:
1. It is used to politically disempower, physically abuse, and materially disadvantage the majority.

2. It causes us unnecessary psychological harm by creating the conditions under which we can experience negative moral emotions such as guilt, and a perception of ourselves as bad.

3. It does not help eradicate conflict, but in fact perpetuates and exacerbates it on a grand scale.

With regard to the first point, Hinckfuss scores an easy goal in showing us that morality can be harmful to people. Everyone will agree with that. But Hinckfuss’s ideas about why morality is harmful are hard to swallow. For one thing, it is hard to believe that the (or even a) major driving force for the material and political stratification of society is the differing amounts of moral praise and blame that people receive as children! It is also difficult to believe that most people are fooled into thinking that those with power, property and influence are as a rule morally better than those who lack it. That is not to say that ruling or ‘upper’ classes have never attempted, or even succeeded, in convincing ‘lower’ classes that they were morally superior. But such a mindset is hardly inevitable. Indeed, many people today are highly sceptical about the moral values of politicians, business leaders and members of other ‘elites’ (as they were in the 80’s, when Hinckfuss was writing). People of wealth and power may feel in some way morally entitled to their wealth and power – and we who lack it may agree or disagree with them on that – but surely very few people think that wealth and power are indicators of or even fitting rewards for living a morally good life.\footnote{Hinckfuss’s view that morality advantages elites at the expense of the majority is the mirror opposite of Nietzsche’s view, in The Genealogy of Morality (1887) and other places, that our Judeo-}
The view that morality is overwhelmingly oppressive may be countered by citing examples of how people have fought to make their societies more egalitarian and more respectful of their less advantaged members. Movements to recognise and promote the rights of women, ethnic minorities, homosexuals, and the physically and mentally disabled (to name just some of the more prominent causes) have had remarkable, if regrettably incomplete, success over the last century or so, and these causes have been fought under a moral banner. It certainly appears as if large groups of people can be inspired to demand radical changes to society – the sort of changes that Hinckfuss would presumably have welcomed – because of their conviction that they have right on their side.

With regard to the second point, I have already said that I am not going to use the fact that eliminativists are likely to continue to experience emotions such as guilt against them. But Hinckfuss, suggests, implausibly in my view, that we would be able to rid ourselves of this emotion just by coming to accept the error theory. However, as I am hardly in a position to prove that Hinckfuss is wrong about this, let us grant him the possibility and see whether guilt is worth getting rid of. Since guilt is a painful experience, and the thought that our actions were morally wrong is always untrue, why not, if it is possible to do so, give up the thought and save yourself the grief?

It is not uncontroversial whether guilt is an exclusively moral emotion. Prinz notes ‘survivor guilt’ – a sort of guilt suffered by survivors of tragic incidents because they have survived while others have perished – as an example of how this emotion may manifest itself in contexts where the sufferer of guilt cannot be construed as having done something wrong (Prinz 2010: 535). And Joyce claims

Christian ‘slave morality’ is the majority’s way of stymieing the elite. Both of these views strike me as extreme.
that “We all know that people can feel guilty over non-moral and self-regarding transgressions” (Joyce 2006: 103). However, even if guilt can be elicited by thoughts other than moral ones, Hinckfuss could still claim that eliminating moral thought would reduce the amount of guilt that we are liable to suffer.

The aim of reducing the amount of guilt in the world is admirable, but I am not sure that the most desirable way to do this is to remove our capacity for guilt. Presumably the eliminativist will think that, for reasons grounded in our subjective values, or self-interest, most of us ought not, at least in normal circumstances, to lie, cheat, break promises, steal or murder. How will she feel if, despite herself, she does end up doing any of these things? If not guilty, then perhaps sad, regretful, angry at herself and eager to make amends. But it is tempting to say that if she only feels like that, especially about the more egregious acts, then she doesn’t feel bad enough. Her being only disposed to feel these milder emotions would seem to indicate that she attributes less value to those who are harmed by her actions than she would if she were disposed to feel guilt. I suspect that this would be a disturbing consequence for most of us. Add to this the fact that those amongst us who lack the capacity for guilt (such as infants and psychopaths) do not seem promising role-models, and we have a powerful consideration in favour of retaining a capacity for guilt in response to behaving in ways that would elicit guilt in a sincere moral believer. I would submit, from the point of view of most of us who are encumbered with a conscience, it seems like a burden worth bearing, and one that we would wish others to bear. We want others not only to act in ways that we approve of, but to have emotional responses that are indicative of how serious they take certain of their actions to be. A person who was not disposed to feel guilty if she perceived herself to have caused great harm may not always be able to
meet this expectation, even if she generally acted in ways that we approve of. The aspiring eliminativist, if she did succeed in ridding herself of moral emotions, would find herself emotionally adrift from her fellows – and what psychological suffering might that cause her?\footnote{Blackburn emphasizes the motivational costs of getting rid of guilt and shame, and also comments on how the retention of moral emotions is compatible with campaigning for a less moralistic and more forgiving society (Blackburn 1998: 20-1).}

If there is no moral truth, then the third point might seem strong. We cannot deny that conflicts are moralized: that is, their protagonists take themselves to be on the side of right, and soldiers and citizens are enjoined to make their sacrifices for moral reasons. But should we be so willing, as Hinckfuss claims to be, to dispense with feelings of duty or loyalty towards a group or cause to which we are deeply attached? And while conflicts are moralized, movements for peace, international cooperation and inter-group tolerance are also moralized. If morality is such a powerful motivator of action, why should we give it up when we can attach it to causes that actually promote human flourishing?

Looking back over history, one can concede to Hinckfuss that morality has been used to sustain practices – such as slavery, the oppression of women, the disenfranchisement of the majority of men and women from politics, torture, capital punishment (often by horrific means), religious persecution, genocide and much, much else – which most of us now abhor. And looking across at the world at large, we see countless instances of the same sorts of thing going on today. As the Hinckfuss saw things, morality serves as an obstacle to improving the human condition: it is because individuals see the moral propositions they accept as being authoritative that they feel entitled or obliged to maintain the status quo, or impose their view of the right way to live on others. But there is a difference between seeing morality as authoritative on the one hand, and claiming infallible access to
the content of morality on the other. Citizens of the democratic nations of the world really are free, to an unprecedented degree, to discuss and challenge prevailing moral opinions, and can do so in the light of an unprecedented access to information about – and personal contact with – the conditions and attitudes of people from other cultures. Of course war, oppression and the rest are still very much with us, and the warmongers and oppressors may feel morally justified in what they do. But it would seem hard to sell the idea that a good way of combating war and oppression would be to give up on ideas of rights, fairness, justice and moral goodness, as Hickfuss wanted us to do.

I don’t think that Hinckfuss has a convincing case that morality is worse for us than its eradication would be. It is not difficult to discover cases where moral practice is harmful in the ways that Hinckfuss indicates. But we cannot discount the palpable beneficial aspects of moral practice either. This might suggest that we should simply perform a cost-benefit analysis when deciding whether or not to purge moral discourse and practice, but it is hard to know how this could be done with any claim to exactness. Individuals such as Hinckfuss may chose to abandon morality if they want to (and if they can). The rest of us, who think that moral thought and practice is or can be beneficial, can draw from Hinckfuss cautionary tales of how moral codes can be put to inhumane ends, which can only help us to ensure that our morality works for everyone’s benefit, and not just privileged ‘elites’.
3.2.2 Should we eliminate moral discourse just because moral judgements are untrue?

It might be thought that it is in all circumstances bad, or at least likely to be bad, if we have false beliefs where we could have true ones, especially over matters that are of central concern to us. And it might seem even more obvious that continuing to go in for moral reasoning, i.e. continuing to partake in a process that generates false beliefs, should certainly be dispensed with if possible. It might seem obvious, but I’m not sure it’s right.

Richard Garner, who advocates eliminativism, compares the case of retaining moral beliefs, though they are false, with belief in Socrates’ ‘noble lie’. Plato has Socrates saying that the citizens of the Republic will be told that the earth is their mother: “So now they must think of the land they dwell in as a mother and nurse, whom they must take thought for and defend against any attack, and of their fellow citizens as brothers born of the same soil” (Plato 1941: 74). Garner comments:

If the members of any society should come to believe Socrates’ fable, or any similarly fabricated radical fiction, the result would be a very confused group of people, unsure of what to believe, and unable to trust their normal belief-forming mechanisms. It is not wise to risk having a society of epistemological wrecks in order to achieve some projected good through massive deception. (Garner 1993: 96)

Any time we do not see things as they are, we are to some extent made incapable of appropriate behaviour. Further, if fanciful doctrines are widely believed, people will become confused about what it takes to have a true belief, and will be encouraged to indulge in many spurious modes of cognition. (97)

Richard Joyce, though a revolutionary fictionalist rather than an eliminativist, has similar concerns about retaining moral belief:
In the vast majority of cases having a true belief to act upon is more likely to bring satisfaction of desire than having a false belief on the matter, and given that we don’t know in advance how and when we are going to employ a particular belief, the safest bet is to have the true over the false one … A seemingly useful false belief, moreover, will require all manner of compensating false beliefs to make it fit with what else one knows. This is what is so hopeless about Plato’s “medicine of deception”: for the citizens of the republic sincerely to believe that their origin lay underground (to say nothing of being partly made of metal), they would need to interpret large amounts of the evidence of their senses in dramatically eccentric ways. (Joyce 2001: 179)

I think Garner and Joyce do little to advance their case by bringing in Socrates ‘noble lie’. As Garner himself says, “That story is so preposterous that it is hard to see how anyone might be led to believe it” (Garner 2007: 88). Accommodating that belief might require all sorts of radical reinterpretations of our other evidence, but that does nothing to show that accommodating moral beliefs will require such reinterpretations. Indeed, most of us already have moral beliefs, and it would seem that these can exist happily alongside, and without distorting, our empirical beliefs about our physical origins and constitution. So the claim that having false moral beliefs will leave us as ‘epistemological wrecks’ is just untrue.

Garner seems to be objecting not just to the position I labelled ‘conservationism’ but to the one I labelled ‘propagandism’. Someone who advocates propagandism thinks that we should hush up the evidence for error theory lest society collapses. Garner suggests that this will be achieved through “mass deception”, achieved by the promulgation of “fanciful doctrines” and “radical fictions”. But morality hardly seems like a fanciful doctrine or a radical fiction to most people, so it is hardly likely that proponents of propagandism would have to expend a lot of effort on mass deception (Cuneo and Christy 2011: 94).
There may be some types of belief with the potential to wreak havoc in the way that Garner and Joyce suggest that false moral beliefs will. An obvious example would be certain forms of religious belief. Galileo’s struggles with the Catholic Church over cosmological matters are well known. Galileo had empirical evidence on his side, the Church had its dogma, and we all know who was in the better position, epistemologically speaking. In our own day biblical literalists, to maintain their belief that earth is a few thousand years old and that Adam and Eve are our earliest ancestors, must deny the vast amounts of evidence from geology, biology, palaeontology and cosmology that shows that the earth is billions of years old and that our species has evolved from simpler life forms. So religious beliefs do have the potential to disrupt our empirical beliefs. The same goes for all manner of groundless pseudoscientific beliefs. Could moral beliefs have a similar effect?

Religious texts make factual claims about the history and future of the universe, how it came to be and how it will end, so there is an obvious way in which religious belief can interfere with scientific reasoning. I can think of a couple of ways that a belief in objective moral values might have the potential to interfere with our other beliefs. One is with the scientific investigation into the origin and nature of moral belief itself. If we believe that there are objective moral values which we have access to via some form of moral intuition, then this could affect what we say about the origin and nature of moral belief when we try to investigate the matter empirically. As I argued in chapter 1, the fact that we do not have to posit objective moral values in order to explain the origin of moral discourse and practice is a powerful reason for thinking that they don’t exist. However, if we are unable to give up our commitment to objective values, it may be feared that this will distort what we say about the scientific study of morality.
And it would threaten the integrity of science and philosophy to allow ourselves to pick and choose the scientific or philosophical findings that fit our prejudices. My solution to this, which I expand on in the following chapter, is that we can retain both our moral beliefs and our scepticism that any of those beliefs are true – it is all a matter of the context in which we have them. Moral beliefs are appropriate for certain contexts in which we are deciding what to do and what to admire. By contrast, it is widely held that most philosophical beliefs are not very important for practical decision making. Scientific beliefs can be very important to get right when we engage in decision making, but it is doubtful that scientific beliefs about the origins of moral discourse and practice will often be relevant to our day-to-day decisions. As I will argue in the following chapter, we can compartmentalize beliefs without becoming epistemological wrecks. In any event, the threat here is only to those who are persuaded by error theory, not to the mass of people who are entirely ignorant of it.

Another way that morality might be thought to have a distorting influence on our other beliefs is if we explain the causes of people’s behaviour largely in terms of their moral character. If we think that criminal and antisocial individuals behave as they do because they are evil, without investigating the non-moral causal factors that have shaped their behaviour, then we might fail to understand criminal and antisocial behaviour adequately, and so fail to create effective policies to prevent them. I think that this is a significant worry. It can seem to some people, and to large sections of the media, that to explain a criminal’s behaviour with reference to the social environment they grew up in, or to their genetic makeup, is to excuse the criminal of her crime. I can’t say much about this issue here, because it involves the issue of whether we can ever genuinely be said to be responsible for our
actions, which in turn involves getting into the free will debate – not something to be dipped into lightly. What I will say is that if we think that criminal and antisocial behaviour will be more effectively reduced if we focus on the non-moral, causal factors that underlie it, then we have a good reason to shape social policies accordingly. If such policies are resisted by popular opinion because they seem to excuse the criminal, then it may be wise to do what we can to change popular opinion to the view that, when designing social policy, it is better to focus on the overall outcome – a safer and more pleasant society – than on whether some people escape the moral condemnation they seem to deserve.

The last thing I want to look at in this section is Joyce’s claim that

In the vast majority of cases having a true belief to act upon is more likely to bring satisfaction of desire than having a false belief on the matter, and given that we don’t know in advance how and when we are going to employ a particular belief, the safest bet is to have the true over the false one. (Joyce 2001: 179)

We can happily grant Joyce his ‘vast majority of cases’, as that’s compatible with their being a minority of cases where having a false belief is better (or no worse) than having a true one. Some psychologists claim that psychologically healthy individuals are more likely to have a slightly inflated sense of self. Shelley Taylor and Jonathon D. Brown describe as dominant the view in psychology that “…the psychologically healthy person is one who maintains close contact with reality” (Taylor and Brown 1988: 193) and has “…an awareness of both the positive and negative aspects of self” (195). In contrast they found evidence that most people’s view of the self is distorted, in that they have unrealistically positive opinions of their own abilities; unrealistic appraisals of their ability to exercise control over what happens to them; and to be unrealistically optimistic about the future. By
contrast, those individuals who were depressed or suffered low self-esteem were
more likely to have a more balanced view of their own skills and ability to exercise
control over their circumstances, and be more realistic about their futures. In their
conclusion they state that:

The mentally healthy person appears to have the enviable capacity to
distort reality in a direction that enhances self-esteem, maintains belief
in personal efficacy, and promotes an optimistic view of the future.
These three illusions … appear to foster traditional criteria of mental
health, including the ability to care about the self and others, the ability
to be happy and contented, and the ability to engage in productive and
creative work. (204)

Might moral belief have similarly positive effects? In considering the
advantages of morality to the individual, Hume speaks of

Inward peace of mind, consciousness of integrity, a satisfactory review
of our own conduct; these are circumstances very requisite to
happiness, and will be cherished and cultivated by every honest man,
who feels the importance of them.

Such a one has, besides, the frequent satisfaction of seeing knaves, with
all their pretended cunning and abilities, betrayed by their own maxims;
and while they purpose to cheat with moderation and secrecy, a
tempting incident occurs, nature is frail, and they give into the snare;
whence they can never extricate themselves, without a total loss of
reputation, and the forfeiture of all future trust and confidence with
mankind. (Hume 1777: 283)

Seeing ourselves in a morally positive light is impossible if we eliminate morality.
It sounds plausible to suggest that seeing oneself as morally good is conducive to
psychological wellbeing. The other point Hume makes in this passage is perhaps
more immediately persuasive. If we think about our own behaviour in purely self-
interested terms, then we may be motivated to behave ourselves in a pro-social
manner when all eyes are upon us, but why should we continue to do so when we
think we can get away with a little anti-social, but self-serving behaviour, when no
one seems to be looking? Hume’s answer, of course, is that we will eventually overreach ourselves and will rob ourselves of the benefits of a good reputation. Joyce agrees with Hume that it is a genuine benefit of moral thought that it can help to prevent us to avoid getting into the situation of Hume’s knaves. It would seem that we would not be able to enjoy this benefit if we abandon moral belief. Joyce disagrees. He thinks that we can retain this benefit even if we regard morality as a fiction. It is part of my purpose in section 4.4 to argue that this is not the case.

Neither Garner nor Joyce have a persuasive case that false beliefs inevitably do more harm than good. But before we leave this issue behind, it will be worthwhile considering whether it would be possible to reform moral discourse to the extent that it is possible to engage in it without believing or uttering untruths. In the final section of this chapter, I consider whether adopting relativism might give us a way of doing this.

### 3.3 Reformist Realism: Moral Relativism for Moral Error Theorists

Moral relativism (henceforth just ‘relativism’), as I will use the term, is the thesis that the truth of a moral claim can only be assessed relative to a particular system of morals, and no system of morals can be judged morally superior to any other. (This does not preclude there being some other standard of evaluation by which moral systems might be ranked, e.g. by how well different moral systems promote health, happiness, or economic growth.) So I am using the term to denote a metaethical thesis, rather than the descriptive thesis that what is considered morally
right and wrong varies with time and place. As it stands, relativism is incompatible with the error theory. The error theorist holds that the truth of moral claims is not to be judged relative to contingently adopted moral systems, but by an objective standard that applies to every agent. So if error theorists are to turn to relativism as a way of salvaging moral discourse, we will have to see them as advocating a reform of moral semantics.

What might motivate an error theorist to turn relativist? Two motivations suggest themselves. The first is the idea that without morality, things are liable to go badly for us. Indeed, this seems to be Mackie’s chief motivation for advocating the retention of morality in the second part of *Ethics* (see especially chapter 5). The second is the idea that moral discourse and practice are so central to our conception of our characters and actions that it would be desirable to avoid any major change to our discourse and practice. We usually want to see ourselves as good, and our actions as right, and to see our attitudes of praise or dispraise towards the actions and characters of others as justified. Relativism might appear to offer a way of preserving much that is valuable in our former conception of our characters and actions, without our having to engage in fictions or retain false beliefs.

The interesting question to ask about moral relativism, once we have accepted moral error theory, is whether relativism can be adopted as a revision to our current moral practice. Can moral relativism help us to preserve the socially useful function of morality?

---

44 A more precise term for the thesis I am discussing would thus be ‘metaethical moral relativism’. Noncognitivist and response-dependent theories are sometimes thought to have relativistic implications, though prominent defenders of these theories have argued that this is not the case (Blackburn 1999; Smith 2002).
3.3.1 Is moral relativism even coherent?

It has often been questioned whether relativism is even a coherent thesis. Gilbert Harman, who has been the most prominent defender of moral relativism within the analytical tradition, has complained about the hasty dismissal of relativism by many philosophers. In his earliest article on the subject, “Moral Relativism Defended” (Harman 1975), he writes that moral relativism is often defined in such a way that it is obviously incoherent. Relativists are characterised as holding that “(a) there are no universal moral principles and (b) one ought to act in accordance with the principles of one's own group, where this latter principle, (b), is supposed to be a universal moral principle” (1975: 3). As we saw in the first chapter, Harman’s own view is that moral judgements are true and false relative to some (tacitly) agreed standards, that function to regulate social behaviour. Thus he can reject (b). There is no objective rule saying that you ought to act in accordance with the principles of your group. You are only obliged to act in accordance with your group’s principles if you agree with those principles.

Another problem that relativism is often thought to face is that it seems to easily generate contradictions. If the standards of one group say that a certain activity is morally forbidden, while the standards of another group say it is permissible, then it would seem that, according to relativism, an activity can be both permissible and forbidden. Whether or not relativism does generate contradictions in this way depends on the details of the theory. David Lyons identifies two types of relativism, which have different things to say about when a moral judgement is appropriate, which he labels ‘agent’s-group relativism’ and ‘appraisers-group relativism’ (Lyons 1976) (for short I’ll call these ‘agent relativism’ and ‘appraiser relativism’). Agent relativism says that an individual’s
conduct can only be appraised using the norms of the group to which she belongs.

Appraiser relativism says that the norms of a group govern all the normative
judgements of that group, including those made about those who do not share the
group’s values.

Appraiser relativism seems to be obviously logically flawed. Suppose that
culture X says that a particular act of cannibalism, committed by a member of
group X, was right and culture Y says that the same act of cannibalism was wrong.
There is no problem with saying that it is true-according-to-X that the act was right
and true-according-to-Y that the act was wrong. That just says again, in different
words, that X says the act was right and Y says that it was wrong. But the appraiser
relativist holds that the truth of a moral claim is determined by the standards of a
group, so the claim that the act of cannibalism was right turns to be both true and
false.

We could try saying that the words ‘right’ and ‘wrong’ have different
meanings for X and Y. This would also remove the contradiction, but at the price
of saying that adherents of different moral systems did not disagree over whether
the action was wrong. This is unacceptable. Even if relativism did turn out to be the
correct theory of moral discourse, members of different groups are obviously not
just talking past each other when they make such claims.

Agent relativism doesn’t have this problem with contradictions. But it does
sound implausible as stated. In calling an action ‘right’ we are usually taken to be
endorsing that action. Culture X thinks that the act of cannibalism was right, and as
it was committed by one of the members of X in accordance with the principles of
X, their opinion is true according to agent relativism. To be good agent relativists,
culture Y should revise its opinion and agree with X that the act was right, even
though we can imagine the members of Y being revolted by the act, sincerely wishing that it had not happened, hoping that such acts won’t happen again in the future, and even planning an intervention to prevent members of X committing cannibalism in the future. It would seem bizarre to hold this combination of attitudes and intentions toward a type of action that you considered morally right, and this seems to suggest that the members of Y would be meaning something different by ‘morally right’ than we usually do.

Harman gets around this problem by making agent relativism even more restrictive. As I said during my discussion of the Objectivity Thesis in chapter 1 (1.2.3), he thinks that there are some moral judgements that we can only make about members of the moral group that we belong to, which he calls ‘inner judgements’. These are the judgements that an agent ought or ought not to behave in a certain way, or that it was right or wrong for her to behave in that way. There are two considerations that make Harman think that we can only make inner-judgements about members of our own group. The first is that he thinks that normative reasons are ‘internal’ in the sense of ‘internal’ associated with Williams (1981): our normative reasons “have their source in [our] goals, desires, or intentions” (Harman 1975: 9). The second is that Harman also accepts that to say that someone ought to act in a certain way is to approve of them acting in that way. The idea, then, is that I can make judgements about what some other member of my group ought to do because we both fundamentally agree on what it is that he ought to do – i.e. behave in a way that is consistent with the standards that we have agreed to. We both intend to keep to our agreed principles, so long as the other members of group intend likewise. I cannot use my moral principles as a guide to
what someone from another moral group ought to do, however, because members of other groups do not have this agreement in intention.

Harman seems to give us a coherent thesis, but as I argued in chapter 1, it is not convincing as a theory about how we actually use moral language. It seems perfectly acceptable to morally criticise members of other societies who are not party to tacit agreements with us over how to live. But perhaps we can use Harman’s ideas about how moral discourse actually operates as a template for thinking about how moral discourse could operate once we stripped it of its claim to objectivity.

### 3.3.2 So should we turn relativist?

Although Harman succeeds in making relativism coherent, it is doubtful that it is a practical option for the error theorist to take.

What we might do is to decide to use moral terms like ‘ought’ and ‘ought not’ only to give advice to or to admonish other people who share our values – who intend to act on the same principles that we intend to act on. Why should someone who shares my values need advice from me about what to do? It may be that she is unsure over which of two competing reasons is stronger, that she needs some words of encouragement, that she is not fully informed about the relevant non-moral aspects of her situation, or that she has been unable to reason from her values and her beliefs to a conclusion about what she ought to do.

There is a fairly obvious problem with this proposal. What if someone shares only, say, half of my basic principles? Another person might share a third of them, another five sixths of them, and so on. If I were concerned about giving correct moral advice to a particular person, then I should have to work out exactly where
her moral principles were in agreement with mine before I could even know in which areas I was entitled to give her advice. I could not even form an opinion of what any person morally ought to do until I knew which moral principles we shared. This will be mitigated to some extent by the fact that we can expect there to be a certain degree of convergence in core moral values, especially within a particular society. But in large, complex societies there will be significant disagreements as well (think of the readership of The Daily Mail versus the readership of The Guardian). The turn to moral relativism is supposed to give us a way of preserving moral discourse without falling into error, but it looks as if a determined attempt to avoid error will be hard work indeed, if it is possible at all.

This problem could be overcome at a stroke, if we decided not to care whether the person to whom we address a moral judgement shares the principles on which the judgement is based. If she does, all well and good: she is ‘one of us’. If she does not, she may still be fooled into thinking that a moral judgement we make about her applies to her, and this may result in her acting in a way that we approve of. Harman seems to endorse this tactic: “a speaker may pretend that someone is susceptible to certain moral considerations in an effort to make that person or others susceptible to those considerations”, adding that “[i]nner judgements about one’s children sometimes have this function” (Harman 1975: 8). Now, surely our success in convincing others that they ought to act in a certain way will have much to do with the conviction with which we express our judgement, and most of us are only able to display conviction when we are saying something that we actually believe. So it would seem that we are better off believing false judgements about what other people ought to do after all.
The situation this leaves us in is this. We can be honest relativists, committed only to expressing moral opinions about those who share our principles. The problem is, it would be extremely difficult to work out to what extent we shared our principles with anyone. On the other hand, we could be dishonest relativists, expressing moral judgements without regard for their truth, but only with regard to whether they were in line with our personal principles. And if we take this second option, it will be more effective if we actually believe what we say, and thus we will not avoid error after all.

Relativism is not only unconvincing as a metaethical theory, but unconvincing as a practical proposal for revising moral discourse in the light of the error theory.

3.4 Conclusion

Proponents of eliminativism do not have a convincing case for the harmful effects of moral belief. However, there is an understandable tendency amongst philosophers to want to eschew false beliefs, and systems of thought that generate false beliefs. Turning relativist is one suggestion for reforming moral discourse in order to avoid error. But it looks as if it will be pretty much impossible to be an honest relativist, while the best way to be a ‘dishonest’ relativist will be to believe your own false moral judgements. So relativism turns out to be either useless or unstable.

In the next and final chapter, I examine revolutionary fictionalism and non-revisionism.
4: Fictionalism and Conservationism

If we come to accept error theory that would seem to oblige us to adjust our attitude to moral discourse. Eliminativism and relativism seemed unattractive, but there is a further option that has received discussion over the last decade. In one of its varieties, moral fictionalism is the thesis that we should treat morality as a useful fiction (in its other variety it is the thesis that we already do treat it that way). In the bulk of this chapter I focus on fictionalist responses to error theory.

Fictionalism has become an increasingly prominent antirealist strategy in recent years, and has been applied to a broad range of discourses, including mathematics (Field 1980; Leng 2010; Melia 1998; Yablo 2000a; 2000b; 2005), possible worlds (Rosen 1990), (allegedly) empty names (Walton 1990), unobservable theoretical posits of science (van Fraassen 1980), truth (Woodbridge 2005), composite material objects (Rosen and Dorr 2002), religious discourse (Lipton 2007), as well as ethics (Joyce 2001; Kalderon 2005; Nolan, Restall, and West 2005). Fictionalism is a broad church, diverse both in its motivations and in its strategies; but two things that the different varieties of fictionalism have in common are their approach to semantics and their recommendation, or observation, that we treat the subject matter of the discourse as a fiction.

In sections 4.1 and I will explain what fictionalism is and how it might be a solution for the error theorist. Then in the subsequent sections I will look at attempts to develop versions of moral fictionalism. I consider two key motivations for moral fictionalism. (1) moral discourse gives us the resources to describe certain non-moral states of affairs that we could not describe without it (or at least allows us to describe them with greater economy than we could without it) (section
4.2. (2) pretending that we are morally obliged to do or refrain from certain activities, which we have prudential reasons to refrain from anyway, will help motivate us to do or refrain from those activities (4.5). Regarding (1), I argue that there is some truth to the claim that moral discourse can provide us with some economy in description, but I find no good reason for thinking that we need moral discourse to describe anything non-moral. And regarding (2), although I agree that thinking of our behaviour in moral terms can have positive motivational effects, I disagree that it can have those effects if we treat it as a fiction. In section 4.3 I discuss whether, if we were to adopt moral fictionalism, we should adopt a pretence or a non-pretence fictionalist theory. And in section 4.4 I discuss whether we should employ the fiction self-consciously or non-self-consciously.

The final option is to refrain from – or perhaps resist – changing our attitude to moral discourse in any way. To go right on believing that characters, actions and institutions can be morally evaluated, and that people can be rightly praised or criticised for the moral qualities of their actions. I argue that this option – *conservationism*\(^{45}\) – has all the same advantages that fictionalists have claimed for their option, but avoids its most pressing problems (4.6).

**4.1 What is fictionalism?**

In section 1.1 I said that metaethical theses are either revisionary or non-revisionary. Most at least aim to be non-revisionary, although there are notable exceptions, such as those of Anscombe (1958) and Harman (Harman and Thomson1996). Although these philosophers think that some of the concepts involved in current moral discourse are defective, they believe that the discourse can be reformed in such a way that it is descriptive of a real ethical subject matter.

---

\(^{45}\) I take this use of the term from Jonas Olson (2011b), who also defends the thesis.
I also said that error theory is not, in itself, either revisionary or non-revisionary. Once we have been convinced by error theory, there is a further question of whether or not we should continue to use the discourse in its current state.

Most realist theories are non-revisionary, but so are most antirealist theories. Although non-revisionary antirealists are sceptical of the apparent subject matter of their target discourses, they reject the idea that ordinary utterances in the discourse are systematically untrue. One way of doing this is by arguing that despite their surface grammar, apparently indicative sentences of the discourse do not have a representational function, but serve to express noncognitive mental states (expressivism). Another way, developed by Crispin Wright, is to argue that commitment to metaphysically contentious entities can be avoided by denying that the notion of truth is uniform across discourses. On this view (which Wright calls ‘minimalism’), truth for realistically conceived discourses equates to correspondence with an external reality, but a weaker notion of truth as ‘warranted assertability’ serves for other discourses (Wright 1992).

Over recent years hermeneutic fictionalism has appeared as a rival non-revisionary view. Earlier non-revisionary views have faced tough criticisms of the non-standard semantics that their theories employ. Hermeneutic fictionalists interpret the semantics of their target discourses in exactly the same way that realists do. They avoid commitment to contentious entities by claiming that indicative sentences of the discourse are not used to make assertions – or, at least, not used to assert the literal content of the sentence – but for some other purpose. This is not a recommendation about how we ought to use the discourse, but purports to describe how it is used. This does not mean that hermeneutic fictionalists cannot be error theorists, as they may yet think that, if indicative
sentences of the discourse were used to make assertions, those assertions would always be untrue.\textsuperscript{46}

Revisionary antirealists are typically error theorists, who interpret indicative sentences of the discourse as being systematically untrue. There are three basic options for revisionists. You can just stop using the discourse (eliminativism). You can reform the discourse by tweaking the meaning of its distinctive terms in such a way that they do refer to something you believe in (reformist realism).\textsuperscript{47} Or you can retain the discourse with the meanings of its terms intact, but change the way you use it. The \textit{revolutionary fictionalist}, taking the third option, treats indicative sentences of her target discourse as fictitious, but for some reason useful to preserve. She absolves herself from error (or at least attempts to) by switching from believing the sentences of the discourse to ‘accepting’ them, where ‘acceptance’ indicates a willingness to use the sentences for whatever reason they are deemed to be useful, but not to assert them as true (more on this later).

For most of this chapter I will have revolutionary fictionalism in mind. It is more natural to think that someone who is tempted by fictionalism as a response to error theory will think of it as a reform rather than an interpretation of how we’ve been using moral discourse all along. I don’t think that hermeneutic fictionalism is any less viable an option for an error theorist, as such, than revolutionary fictionalism. It is just that most error theorists seem to think that ordinary speakers

\textsuperscript{46}Hermeneutic fictionalists have taken a variety of views about the semantics of their target discourses. Joseph Melia (1995) appears to think that mathematical terms do not refer to anything real; Stephen Yablo appears to be officially agnostic on that point (2001: 72; n. 1). Although he defends hermeneutic moral fictionalism as a theory about how moral discourse is actually used, Mark Kalderon considers whether we ought to use the discourse realistically (Kalderon 2005: 157).

\textsuperscript{47}We considered this option in chapter 3. Another example of it is Trenton Merricks’ suggestion that if we accept his arguments that no non-living composite material objects exist, we should reform the way that we use terms such as ‘table’ and ‘statue’ in such a way that they no longer refer to composite objects (Merricks 2001: 186).
are committed to a moral reality. That claim is not, in my view, an essential component of error theory (as I said at the end of 1.2.1). For some of the following discussion, it isn’t necessary to distinguish between hermeneutic and revolutionary fictionalism, but I’ll make the distinction clear when it matters.

Is fictionalism plausible? I think it is not just plausible, but clearly true that we sometimes utter indicative sentences without the intention of asserting the content of the sentence, but for some other purpose.48 For example, an actor playing Cordelia in a performance of *King Lear* makes the following utterance (directed at the actor playing Lear):

```
Good my Lord,
You have begot me, bred me, loved me.
I return those duties back as are right fit,
Obey you, love you, and most honour you.
```

This actor has most likely not been begotten by the actor playing Lear, and even if she has been, her intention in uttering these lines is obviously not to communicate that fact. So she must have some other linguistic intention. Some philosophers have defended realism about fictional characters,49 but even if that thesis were true, this utterance is clearly not intended to convey that the fictional character Cordelia was begotten by the fictional character Lear. For one thing, fictional characters cannot really be parents (that’s strictly the privilege of biological creatures). Nor should she be interpreted as saying that Cordelia is the daughter of Lear in the fictional

---

48 This distinguishes fictionalist utterances from simple lies, where the intention of the speaker is to assert something which she does not believe, with the intention of getting her audience to believe it. I don’t want to preclude there being lies within fictionalist discourses. Indeed, these should be easy to produce. The following sentences, if uttered with the intention to deceive, seem to be lies: “Hamlet lived at 221b Baker Street”, “Sherlock Holmes was the Prince of Denmark”.

49 E.g. Thomasson (1999). Thomasson construes fictional characters as created, but abstract, objects.
world of the play. For that would be to talk about the play, which is clearly not what actors are doing during a performance.

I think the best thing to say is that she has not asserted anything. She doesn’t assert that she is Lear’s daughter, or the daughter of the actor playing Lear, or that Cordelia is Lear’s daughter in the world of the play. What else could she be asserting? We might say that these lines are meant to assert something about the proprieties of father-daughter relationships: that daughters ought to love, honour, and obey their fathers. But although it may have been part of Shakespeare’s intention in writing these lines to make this assertion, and even if the actor playing Cordelia was in agreement with Shakespeare on this point, we should not interpret her as asserting this during a performance of the play. After all, whether she agrees with it or not is immaterial to her performance. In delivering her speech she cannot be making an assertion, as it is essential to assertion that the speaker purports to believe what she is saying, and this is not essential to the actor’s performance. Indeed, it is essential that she does not indicate in any way to the audience that she believes anything that she says during the performance, as that would be to break out of character.

One thing we can say with absolute confidence about the actor’s performance is that it is intended (in collaboration with her co-performers) to convey a story, and with only slightly less confidence, that the author of this story has something to say about the proprieties of father-daughter relationships. But as we have seen, we had better not interpret the performers as making such assertions. What we can see them as doing, at least in part, is offering Shakespeare’s thoughts for the audience’s consideration. Their utterances are the vehicle by which this is

50 Of course what playwrights are saying with their texts cannot be just ‘read off the page’, but may require a considerable amount of interpretation.
achieved, but they themselves assert nothing, even when what they say has the form of an assertion.

A different type of fictionalist utterance can be encountered in contexts where we are talking about a fictional story, rather than acting it out. If when talking about *King Lear*, I say that Cordelia is Lear’s youngest daughter, I clearly should not be interpreted as asserting that a real person, Cordelia, was fathered by another real person, Lear. So how should my utterance be interpreted? As with the prior case, we could say that I am pretending that Cordelia and Lear are real people who stand in the daughter-father relationship. An immediate problem with this sort of interpretation is that, in the sort of contexts where such utterances are made, such as in English literature classes, it isn’t obvious that speakers are engaged in acts of pretence. Another interpretation would be to say that my utterance has an unvoiced story operator. When I say that Cordelia is Lear’s youngest daughter, what I really mean, on this view, is that according to the Lear-fiction Cordelia is Lear’s youngest daughter.\(^{51}\)

There is something quite intuitive about this view. When discussing the characters of a play, I could aim for maximum clarity by actually voicing the story-operator; but in practice, it would be tedious to do so. On this view, although my utterance conveys a belief to my audience – a belief about the content of the Lear-fiction – it does not do so by asserting the semantic content of the uttered sentence, because that sentence does not contain the story operator. So this would also count as a fictionalist utterance. Non-fictionalist interpretations of this sort of utterance have been defended, which would treat its subject matter as abstract objects, or non-existent objects, or merely possible objects.\(^{52}\) Perhaps a minimalist (à la Wright) interpretation of such utterances could be defended; but an

\(^{51}\) See Lewis (1978) for a defence of this view.
\(^{52}\) See Sainsbury (2010) for a review.
expressivist interpretation would seem to be ruled out, as I am clearly, in one way
or another, expressing a belief when I say that Cordelia is Lear’s youngest
daughter. In any case, I think the fictionalist interpretation is immensely plausible._

These cases nicely illustrate two types of speech act that might be performed
in making a fictionalist utterance. (1) When in an English literature class I say that
Cordelia is Lear’s youngest daughter, I am not asserting the literal content of that
sentence, but I can most plausibly be interpreted as asserting something. Those
familiar with the play would probably agree that I had said something true, and that
if I’d said that Regan was Lear’s youngest daughter, I’d have spoken falsely. To
keep things clear, fictionalists distinguish between assertion and quasi-assertion.
To quasi-assert is to make a statement with the form of an assertion, where the
speaker’s intention is not to convey the literal content of that statement, but for
some other purpose. In this case, I quasi-assert that Cordelia is Lear’s youngest
daughter, but my intention is to convey that in the Lear-fiction, Cordelia is Lear’s
youngest daughter. In Stephen Yablo’s terms, that is the real content of my
utterance, and is what I assert by making the utterance (Yablo 2001). However, not
all instances of quasi-assertion are also instances of assertion (2) Although the
actor playing Cordelia quasi-asserts she was begotten by Lear, she does not
actually assert anything. (1) is an instance of what Matti Eklund calls content
fictionalism (Eklund 2007), where the fictionalist utterance is used to assert
something other than its literal content. In the present case, the context cues the
audience to understand that the speaker’s utterance is qualified by an unvoiced
story operator, but there are other ways that the asserted content (what is actually
expressed) can differ from the literal content. (2) is an instance of what Eklund
calls force fictionalism, where an indicative sentence is not used to make any
assertion, but to perform some other function, such as an alternative speech act. I suggested that we can interpret the actor’s utterance as a vehicle for conveying Shakespeare’s thoughts about the proprieties of father-daughter relationships. They also, as part of the overall performance, serve to arouse an emotional response in the play’s audience. According to Yablo, Hartry Field thinks we should ‘quasi-assert’ various things about mathematical objects because this allows us shorten proofs concerning only concrete objects, but gives no indication that in doing so we are really asserting anything else (Yablo 2001: 75). This might tempt us to say that Field advocates a variety of force fictionalism – that is how Eklund (2007) appears to see things – (although another option is to complain that his account is just incomplete). According to Kalderon’s moral fictionalism, the function of indicative moral utterances is fully explained by their being expressive of noncognitive attitudes. When I quasi-assert that tax-avoidance is wrong, I don’t really assert anything. As we shall see, moral fictionalism could be a form of either content or force fictionalism.

This brings us to the question of motivation. Why use a sentence which, on its surface says X to assert something other than X, or nothing at all? In the case of performances of plays and tellings of stories, the full answer may be very long indeed, but we can offer a little of what it will contain. I have claimed that in performing plays, actors sometimes act as vehicles for presenting the playwright’s views. But, there must be some point to playwrights’ expressing themselves in dramatic form, and for actors to perform their plays. The answer is in part that plays (if they are any good, and if they are well performed) are often more emotionally engaging, more memorable, and more vivid than other media. Fictional stories in all forms can help us imagine situations we have never
confronted; they can help us to imagine what it is like to be other people; they can help us simulate the sort of emotions that we would have in situations that we have never confronted, or that we would have if we had different personalities, were of the opposite sex, had a different cultural history, or had abilities or disabilities that we don’t actually have, and in these ways, stories can help us to gain a greater insight into the human condition than is available from our own narrow experiences. The ability of stories to engage our emotions makes them a particularly powerful medium. If fictionalist moral utterances also packed an emotional punch – which it was in some way desirable to deliver, or be on the receiving end of – then this might provide us with a reason to retain moral discourse, even if its subject matter is fictional.

Moral issues are emotive, and moral discussion is often highly emotionally charged. And while this feature of moral discussion might sometimes be regretted, appeals to feeling are ubiquitous, and apparently acceptable by the norms of moral discourse. Most of the discourses for which fictionalism has been suggested are much more sober, however. Whilst discussions of mathematics, possible worlds, and the unobservable posits of scientific theories may be passionate, it is clear enough that appeals to feeling are not considered an acceptable tactic in such debates. So it is implausible that if we are going to be fictionalists about these discourses it will be because of their emotional power. Much more likely, it will be because the discourse regarded as fictional in some way helps us to communicate

53 Some have denied that fictions elicit genuine emotions (Walton). Others have argued that while they do elicit genuine emotions, emotions elicited in this way are irrational (Radford). Richard Joyce argues that such emotions are both real and rational (Joyce 2000).
54 Walton has similar things to say about engaging with plays and other fictions: such engagement “…provides practice in roles one might someday assume in real life, … helps one to understand and sympathise with others, … enables one to come to grips with one’s own feelings, [and] broadens one’s own perspectives (Walton 1990: 12).
factual information more easily, or more vividly than we could if we restricted ourselves to speaking entirely literally.

Take mathematics. As we saw in section 1.3.1, the leading argument for mathematical realism goes like this:

(1) Our best guides to what to believe are our best theories of the world, so we ought to believe what our best theories say.
(2) Our best theories of the world are our best scientific theories,
(3) Our best scientific theories refer to mathematical objects.
(4) Thus we ought to believe in mathematical objects.

Hartry Field has argued that we can rewrite our scientific theories in such a way that they do not refer to mathematical objects, a process he calls ‘nominalization’ (Field 1980). If we can do this, then mathematics will not be part of the best possible theory of the world, and it would seem to be more rational to believe in the entities referred to in the best possible theory, rather than our current best theory. But despite this, Field does not think that we should give up mathematical talk. For one thing, he has only succeeded in nominalising a very small bit of physics (if, indeed, he has succeeded in this). For another the nominalised theories will inevitably be much longer and more cumbersome than their platonistic counterparts.

Other mathematical fictionalists have doubts that a complete nominalization of science is possible, even in principle (Yablo 2000a, Melia 1998). So their motivation for retaining the discourse is that it is necessary for a complete description of the world. However, Yablo and Melia deny that the fact that
mathematical concepts are necessary for description commits us to mathematical objects, because we can interpret our utterances containing mathematical concepts as fictionalist utterances. Moral fictionalism could be motivated in a similar way. We might claim (in the manner of Field) that moral talk allows us to talk more conveniently about certain aspects of the non-moral world; or (in the manner of Yablo and Melia) we might claim that moral talk is practically indispensable for talking about certain aspects of the non-moral world. Daniel Nolan, Greg Restall, and Caroline West have defended the view that moral discourse is, in certain situations, has greater expressive power than non-moral discourse (Nolan et al. 2005: 312). I'll discuss this claim in section 4.2.

Before examining the proposals of Nolan et al. and Joyce, it will be worthwhile completing my general discussion of fictionalism by distinguishing between pretence and non-pretence fictionalist theories, as this will make subsequent discussion clearer.

Some fictionalists think that to engage in certain discourses is to engage in pretence. But some forms of fictionalism do not seem to involve any form of pretence. One form of non-pretence fictionalism is the claim that some of our utterances are (or should be) interpreted as bearing an unvoiced story-operator, as already described. The sentence uttered is just an abbreviated form of the sentence that we want to assert. In making such utterances, we do not seem to be engaged in pretence, nor does there seem to be any reason to interpret them as instances of pretence. Field’s mathematical fictionalism might be another form of non-pretence fictionalism, but it is perhaps again best to say that Field has not given us a complete account of how we are to understand fictionalist utterances. Another form that non-pretence fictionalism might take is expressivist fictionalism. The
expressivist fictionalist claims that in quasi-asserting \( p \) I do not really assert any other proposition, but express some non-cognitive mental state. It doesn’t seem that the expressivist fictionalist is bound to say that quasi-assertion involves pretence. However, it is difficult to pin down what the chief advocate of expressivist fictionalism, Mark Kalderon thinks about this issue. He writes that “In accepting a moral sentence, a competent speaker does not so much believe the moral proposition expressed as he makes as if to believe that proposition, where making as if to believe, in this context, is to be disposed to respond affectively in the relevant manner” (Kalderon 2005: 151-2). ‘Making as if to believe’ sounds like pretending, but being ‘disposed to respond affectively in the relevant manner’ does not.

With a pretence theory, our utterances are to be interpreted as though we are speaking in a pretence, or a ‘game of make-believe’ as Kendall Walton often puts it. Sometimes it is obvious that people are speaking within a pretence, such as when we watch actors on stage or children playing. But the pretence theorist claims that we either do or can employ pretence in more subtle ways. Walton explains how fictions can facilitate description, by introducing the notion of *prop-orientated make-believe*:

> Where in Italy is the town of Crotone?, I ask. You explain that it is in the arch of the Italian boot. ‘See that thundercloud over there – the big angry face near the horizon’, you say; ‘it is headed this way’. (Walton 2005: 66).

It seems that in these examples, there is a systematic connection between what is said in a fiction, and how things are in the real world. Where things are, or how
they are shaped, make certain sentences true in a salient fiction. In each case we can construct biconditionals, such as:

\[
\text{Crotone is located } \text{there} \text{ iff in the fiction, Crotone is in the arch of the Italian boot.}
\]

\[
\text{I am referring to } \text{that cloud} \text{ iff I am referring to the cloud that, in the fiction, is angry.}
\]

Walton calls these biconditionals ‘principles of generation’. They tell you how to move back and forth between a fiction and reality. There is a clear reason for talking in this way: so long as one’s audience catches on, it is just easier to specify where Crotone is, and which cloud we have in mind, by going through the fiction.

Walton’s ideas have inspired Stephen Yablo to take a similar approach to mathematical discourse, construing numbers as playing a metaphorical role: “Rather as ‘smarts’ are conjured up as metaphorical carriers of intelligence, ‘numbers’ are conjured up as metaphorical measures of cardinality” (2000a: 214).

In Yablo’s view, there are descriptive sentences for which mathematical concepts are indispensable. He illustrates this by discussing the following attempt at nominalization (2000b: 295), starting with the sentence:

\[
(S) \text{The average star has 2.4 planets.}
\]

We seem to be able to avoid commitment to the dubious ‘average star’, by paraphrasing \(S\) as:
(T) The number of planets divided by the number of stars is 2.4.

But this sentence, taken literally, commits us to numbers, which Yablo wants to avoid. We could try replacing (T) with the infinite disjunctive sentence:

(U) There are 12 planets and 5 stars or 24 planets and 10 stars or …

But according to Yablo, the rules of English don’t allow infinitely long sentences, so (T), it seems, is the best we can do. Even if infinitely long sentences are good English, they are certainly of no practical use to us, as we could never successfully utter them. Yablo suggests further advantages for (S) over (U):

The formulation in terms of the average star is … on the whole hugely to be preferred – for its easier visualizability, yes, but also for its greater suggestiveness (“that makes me wonder how many moons the average planet has”), the way it lends itself to comparison with other data (“the average planet has six times as many moons as the average star has planets”), and so on. (Yablo 2000b: 296-97).

Even if we cannot actually succeed in expressing the metaphysically lightweight content that we were trying to express with (U), Yablo, I think, pretty successfully gestures at what it is (anyone competent in basic arithmetic could continue (U) at least a bit further), and offers further reasons to prefer (S) to (U).

The most obvious criticism to be made against pretence theories is that most people participating in the target discourse would deny that they are engaged in pretence. The pretence theorist can respond with reasons for thinking that it is not always possible to introspectively distinguish pretence from belief.55 A potentially

55 For a response of this kind see Kalderon (2005: 154).
more devastating complaint, along the same lines is that the discourses to which they have been applied can be successfully engaged in by autistic people who, it is alleged, lack the ability to engage in pretence (Stanley 2001).\footnote{56 For a response to this objection see Liggins (2010).} I shall not enter into these discussions. Whether or not the objections succeed against varieties of hermeneutic fictionalism that employ a pretence theory, versions of revolutionary fictionalism that employ pretence theory would appear to be immune. For the revolutionary fictionalist who employs pretence theory chooses to engage in pretence.

Whether fictionalists choose a pretence theory or a non-pretence theory will depend on what they want to do with moral discourse, once they regard it as a fiction. So let us turn to look at some actual proposals. I’ll discuss a paper by Daniel Nolan, Greg Restall, and Caroline West in section 4.2, then in 4.5 I’ll discuss Richard Joyce’s contribution.

4.2 Nolan, Restall and West

In this section I will discuss Nolan et al.’s attempts to motivate moral fictionalism over eliminativism. I will mostly focus on their most interesting suggestion: that moral discourse allows us to convey information about non-moral states of affairs that we could not convey with non-moral vocabulary. I find this claim unconvincing, but go on to consider (in 4.2.1) whether moral discourse might nevertheless allow us to make non-moral claims more economically. While I find this claim more plausible, it is doubtful that, on its own, this is a good enough motivation for fictionalism.
Let us begin by considering three reasons they give for thinking that fictionalism is a better option than eliminativism.

(1) While it may not be impossible to give up moral discourse, Nolan et al. claim that it would be difficult and inconvenient to do so, as moral concepts “pervade ordinary thinking and discourse” (Nolan, Restall, and West 2005: 311). “Ceasing to talk [morally] would be much like ceasing to talk of people having beliefs, desires, and emotions: possible, perhaps, but not an easy thing to do; and certainly not a consequence to be embraced lightly”.

There are two objections that might be made here. The first is to question why the relative ease of adopting fictionalism over eliminativism (if, indeed, it really is the easier option) speaks in favour of it. Normally the ease of taking one option over another – $A$ over $B$ – is a decisive factor only when $A$ and $B$ are means to the same end, but fictionalism and eliminativism are not best seen as means to the same end. As we saw in the previous chapter, eliminativism has been motivated out of a concern to rid ourselves of an institution – morality – that is allegedly harmful to most people. Fictionalists, by contrast, think that this institution is worth preserving. The two approaches might be thought to share the end of allowing us to avoid making untrue moral utterances. While it is true to say that both eliminativists and fictionalists have that end, I don’t think it provides a motivation for fictionalism. Rather, fictionalism will be worth pursuing only if moral discourse performs a useful function (or functions), even after we have abandoned moral beliefs. It is not worth making fictional moral claims merely because, in so doing, you won’t actually be making any moral claims. For these reasons, the relative ease of adopting one of these approaches over the other seems irrelevant.
The second objection is to deny that eliminativism is psychologically more difficult than fictionalism, and claim that either they are equally psychologically taxing, or that eliminativism is in fact the easier option. I think the latter claim has some plausibility. Both the fictionalist and the eliminativist have to give up their belief in morality, which is arguably psychologically difficult, given our moral upbringings and, perhaps, our psychological predisposition towards moral thought and behaviour. The difficulty of this step would seem to be the same for both eliminativists and fictionalists. However the fictionalist, and not the eliminativist, takes the further step of carrying on with the discourse, with an adjusted attitude to its content. So the fictionalist has two psychological hurdles to overcome, while the eliminativist only has one. Against this, it might be argued that the second step taken by the fictionalist mitigates some of the psychological harm done by taking the first step. Moral pretence might turn out to be a pretty good substitute for moral belief. This view is advanced in detail by Richard Joyce, so we’ll discuss it further when we get to his views. As we will see, Joyce thinks that it need not seem to the fictionalist, when she is using moral discourse in everyday contexts, that she is using it any differently than an ordinary user would. No such claim is made by Nolan et al. If the fictionalist that they envisage is using moral concepts in a way that is significantly unlike how they are ordinarily used, then it is not clear how fictionalism is more psychologically convenient than eliminativism, on their account.

(2) Nolan et al. claim that the fictionalist, unlike the eliminativist, does not have to worry about raising complicated metaethical issues every time she wishes to discuss an applied ethical issue (311). So, if someone asks the fictionalist whether she thinks that factory farming is wrong, for instance, she does not have to
preface her reply by saying that she doesn’t believe that anything is right or wrong, which might seem distracting in the context of a discussion about whether factory farming should be allowed under law. However, the eliminativist can reply that we can respond to normative moral questions, in a reasonably satisfactory way, without committing ourselves to any moral facts. To the question about factory farming, we could couple a factual claim – factory farming causes more suffering than free-range farming – with an expression of subjective preference for less suffering, or even the factual claim that most people would prefer free-range farming to factory farming, if they were aware of the amount of suffering involved in the latter. This is a way of dodging the precise question asked, but the questioner may well be satisfied that her question has been answered adequately (she might not even notice that the eliminativist’s answer lacks moral content, if it’s put subtly). It is not very unusual to object to presuppositions in questions that are put to us, but to want to provide some kind of response nevertheless, without explicitly challenging the presupposition.

(3) Nolan et al. also claim that eliminativism robs us of expressive power. This is the most interesting of the motivations they consider, but they fail to make a convincing case. They compare moral discourse to mathematical discourse. As they point out, saying that there are no prime numbers between 23 and 29 is more succinct than any nominalized alternative (and if Field-style nominalization is not viable for all scientific theories, then we will have no choice but to avail ourselves of mathematical concepts if we want to carry on with science). They seem to imply that moral discourse is indispensable for describing some non-moral states of affairs. I will dispute this, but then show how moral discourse might help to make
such descriptions more economical. First of all, let us look at what Nolan et al. have to say:

There are sentences with moral vocabulary, which we use to imply things about non-moral features of the world, where it seems difficult to identify those features in non-moral terms. Take a common sort of moral claim such as ‘the property rights of some farmers have outweighed the rights of the environment in this case’. Exactly what non-moral features of the situation does talk of ‘rights’ capture? (312)

They are presumably not suggesting that ‘rights’ might refer—however obliquely—to something non-moral. That would make them ethical naturalists, and if they were endorsing ethical naturalism, their fictionalism would be entirely unmotivated. Another option would be to deny that talk of ‘rights’ refers to anything non-moral, but claim that it in some way facilitates non-moral description.

We saw in the previous section how Walton and Yablo showed that pretence theory can facilitate description. Unfortunately, Nolan et al. attempt nothing comparable when they discuss their example. All they do say is that:

The fictionalist can avoid the difficulties of finding an appropriate literal non-moral paraphrase. We can say that it’s literally true that if there are rights then the property rights of the farmer are more important than the rights of the environment. (312)

It is unclear exactly what benefit there is in saying this, or how it makes one a fictionalist. Let’s consider a less curious example. An error theorist might be prepared to say that it’s true that if anything is wrong, then torturing the innocent just for the fun of it is. But almost everyone in the world will agree with her. Even an eliminativist could consistently agree to this. So the fact that the fictionalist can make this claim does nothing to distinguish her from the eliminativist, or show how fictionalism is superior to eliminativism.
If Nolan et al. cannot even gesture at what non-moral content we might be conveying with moral vocabulary, then the suspicion might be that there is none. However, I think we can come to their aid. The moral claims that we make do have non-moral implications, and, pace Nolan et al., it does seem possible to identify what these are (although the identifications will be controversial). I will discuss some of these next.

4.2.1 What might fictionalist moral utterances help us to express?

We can find a suggestion for this in another part of Nolan et al.’s paper. They make the claim that fictionalism is as well placed as quasi-realism to accommodate certain intuitions about motivational internalism, the thesis I formulated as (315):

\[
\text{Motivational Internalism} \quad \text{I judge that I morally ought (ought not) to } \phi \text{ in circumstances } C \text{ only if I am, } ceteris paribus, \text{ effectively motivated to } \phi \text{ (not to } \phi \text{) in } C.
\]

Noncognitivist theories are well placed to embrace Internalism, and it is, in fact, a leading motivation for these theories (e.g. in Blackburn 1998). For noncognitivists equate positive moral statements (such as, “φ-ing is right”) with expressions of

---

57 Quasi-realism is an expressivist project, that aims to show how expressivists can accommodate the realist-sounding trappings of moral discourse, e.g. talk of moral propositions, moral beliefs, and moral knowledge, largely by developing deflationary accounts of these concepts. Its main proponents are Blackburn (1993, 1998) and Gibbard (1990). It is not actually the quasi-realist aspect of Blackburn and Gibbard’s theories that allow them to accommodate Internalism, but the expressivist aspect (i.e. the claim that indicative moral utterances function to express noncognitive attitudes). By singling out quasi-realism, Nolan et al. are presumably just trying to show that fictionalism is as good or better than what they take to be the most sophisticated form of expressivism.
pro-attitudes, and it is not mysterious that having a pro-attitude towards φ-ing goes together with being motivated to φ, under favourable conditions. Nolan et al. say that fictionalism can do just as well here:

We can guarantee that it will be true that some state of affairs concerning morality obtains, according to the fiction, just in case people have certain motivations … by including bridge principles that, in certain circumstances, take one from facts about motives …, on the one hand, to facts about moral truth or beliefs genuinely about morality, on the other, and vice versa. (Nolan, Restall, and West 2005: 316)

The bridge principles referred to here are principles that allow us to move back and forth between statements in a fiction and statements of some discourse that we are realists about (which Nolan et al. call the base discourse). Walton’s principles of generation are thus a kind of bridge principle.

We cannot accept Nolan et al.’s suggestion for how we can construct our bridge laws to accommodate Internalism. They claim that for fictionalism to accommodate Internalism, we need it to be the case that it will be fictionally true that some moral states of affairs obtain just in case people have certain motivations. So they seem to be conflating motivational internalism with another thesis that was mentioned in section 1.2.5:

**Moral Internalism** We cannot have a [moral] duty to act in some way unless (if we knew the relevant facts, and deliberated rationally, we would be motivated to do this thing). (Parfit 1997: 100-102)

*Moral* internalism is a thesis about necessary conditions for the truth of moral judgements; *motivational* internalism, on the other hand, is a thesis about when
moral judgements are sincere or successfully-formed (see my discussion at 1.2.5).

It is *motivational* internalism that noncognitivists think they are best placed to accommodate.

So we need to be careful which variety of internalism we are trying to accommodate. We also need to be careful about what form the bridge laws can take. When I discussed Walton’s examples, the bridge principles took the form of biconditionals, but in this case the construction of a biconditional might not be possible. My being motivated to φ is at most a necessary, but not a sufficient condition of judging that I morally ought to φ, because I can be motivated to do any number of things that I don’t think I am morally obligated or forbidden to do, and I can be motivated to do that which I think I am morally forbidden from doing. We can easily construct a *conditional* bridge principle, such as:

\[
\text{Internalism BP: } \quad \text{If I sincerely judge, in the fiction, that I ought to } \phi, \text{ then} \\
\quad \text{I am – in reality – } \text{ceteris paribus} \text{ motivated to } \phi.
\]

This principle only tells us how to move from the fiction to the base discourse, but not vice versa. I do not think that we can turn this into a biconditional without altering the way moral discourse functions. Assuming the truth of internalism, we can go from facts about an individual’s sincere moral judgements to facts about her motivations, but we cannot go the other way. What this means is that the person who has embraced fictionalism – the “fictionalist”\(^{58}\) – cannot move mechanically from the thought that she is motivated to φ to the thought that she can sincerely judge, in the fiction that she ought to φ. So *Internalism BP* doesn’t tell the

---

\(^{58}\) ‘Fictionalist’ is ambiguous between (a) a philosopher who argues for fictionalism, and (b) a speaker who adopts this philosopher’s recommendations. It’s difficult to think of a better term for (b).
fictionalist what to say in the fiction, but rather puts constraints on what she can sincerely say in the fiction. This isn’t really a problem, however, for we would be in a similar position if moral discourse were in good standing. If that were the case, and Internalism were true, a speaker could not move mechanically from the thought that she is motivated to φ to the thought that it would be good or right for her to φ, for the same reason the fictionalist cannot do this.

We now at least have a suggestion on the table about what non-moral content we might use moral discourse to assert: using moral discourse is a way of talking about what would be motivated to do in certain situations. As things stand, however, we may wonder why we don’t just talk straightforwardly about what we are motivated to do and leave the fiction out of it.

Here are my suggestions. First of all we can, when appropriate, just talk about what we are motivated to do, without using moral vocabulary. Secondly, however, retaining moral vocabulary may allow for more economical expression. For example, if there is an a priori connection between moral attitudes and motivation, it needs to be stated with a ceteris paribus clause. I can think that φ-ing is right, but be unmotivated to φ because I’m depressed or squeamish. And I can lack motivation to φ, although I think it is right to, because the circumstances are not favourable for φ-ing, or I don’t yet have the skills to φ. A fictionalist might say, “It is right to save drowning children”, while an eliminativist might have to say: “I would be motivated to save drowning children, if the following circumstances held …”. Part of the intuitive appeal of Internalism is that making or agreeing to moral judgements is a way of telling people how we will be motivated in certain circumstances. Politicians, for example, stress their conservative or liberal values, and are called hypocrites when they fail to be motivated by those values. It is
useful to be able to let others know how, in general, we are motivated, and moral vocabulary is a useful for that purpose.

Further economy could be gained if we retained moral discourse not only to convey facts about motivations, but also facts about certain kinds of reasons. Recall the thesis I called ‘moral rationalism’:

**Moral Rationalism:**
If $\phi$-ing is the right (wrong) thing for an agent $S$ to do in a particular circumstance, then $S$ ought, for irreducibly moral reasons, to $\phi$ (not to $\phi$) in those circumstances; and there are no alternative considerations that can outweigh the moral ought.

Clearly the fictionalist, in quasi-asserting that some action is the right (wrong) thing to do in the circumstances, will not want to convey that there are irreducibly moral reasons to do that thing (not to do that thing) in the circumstances. But she may sometimes want to convey that there are very strong reasons to do or refrain from doing something – perhaps even some that she thinks that no alternative considerations could outweigh. This will be so if the fictionalist has certain strict principles she is committed never to breaking. Many of us have a mixture of lax and strict principles. For example, I think it generally makes sense to refrain from stealing and breaking promises, but I can think of situations where I would be prepared to do those things. I try not to do these things, on principle, but there are situations where I would be prepared to abandon my principles. On the other hand, I don’t think that, for me, there are any considerations in favour of sadistic torture or child abuse that could outweigh considerations not to do these things. And I
think that, for me, there is always a reason to help someone in imminent danger – if I am well-placed to do so and the risk to me in helping them is not too great – which will not be outweighed by other considerations. I thus have strict principles regarding these things. I can convey these facts about me by saying that it would be wrong for me to take part in sadistic torture or child abuse and that, in situations where I am well placed to do so and the danger to me is not too great, helping people in danger is the right thing to do. And if, in my opinion, you have reasons, that cannot be outweighed, for refraining from these awful activities, and for helping people in need, they I can convey this by saying what it would be wrong or right for you to do. We can construct the following bridge principle:

**Rationalism BP:** \( \varphi \)-ing is the right (wrong) thing for an agent \( S \) to do in the fiction if and only if there is a reason for \( S \) to \( \varphi \) (refrain from \( \varphi \)-ing) that cannot be outweighed by any consideration against \( (\text{for} \) \( \varphi \)-ing).

In saying that \( \varphi \)-ing is wrong, the fictionalist conveys that there are reasons against \( \varphi \)-ing that cannot be outweighed by competing considerations for \( \varphi \)-ing. The reason for going through the fiction, instead of talking straightforwardly about reasons that we do not think can be outweighed, is again that it is more economical to do so. We haven’t discovered anything that cannot be said without using non-moral vocabulary.

There is a danger that, if fictionalists mimic ordinary users of moral discourse too closely, there will be nothing to distinguish them from ordinary users.

*Internalism* and *Rationalism* are controversial metaethical theses, but let’s assume
that they are both true. So when Olivia, an ordinary user of moral language, says that helping people in danger is the right thing to do, she conveys that she is motivated, *ceteris paribus*, help people in danger, and that she has a reason to help people in danger that cannot be outweighed by competing considerations. A fictionalist, Fred, we are supposing, might also say, “It is right to help people in danger” to imply the same is true of himself. So why think that Olivia is speaking plainly, while Fred is involved in some sort of pretence? What differences might separate Olivia and Fred? Two answers suggest themselves. The first is epistemological. Fred works out that it is appropriate to say “It is right to help people in danger” by considering what he is motivated, *ceteris paribus*, to do, and which reasons statements apply to him. As he is using the sentence merely as shorthand for conveying these non-moral facts, he cannot begin with the thought that it is right to help people in danger and work out from there that he is motivated, *ceteris paribus*, to help people in danger and that he has a reason to help people in danger. When ordinary users of moral language make moral judgements, however, they do not usually – if ever – think about motivations and reasons in this way; and it may never occur to some ordinary speakers that their judgements have such implications. It might just seem intuitively true to Olivia that she ought to help people in danger.

The second answer is that Olivia means to imply more than Fred does. Olivia, but not Fred, means to imply that she has a reason to help people in danger regardless of her desires or self-identified interests. When Olivia attributes moral properties to actions and characters, she implies – and perhaps means to imply – that these are objective features of the world. Fred, on the other hand, just uses pre-existing moral terms as shorthand for talking about certain complex relations of
motivations and reasons to agents. (Of course, if his audience are not aware that Fred’s utterances are fictionalist, they may take him to imply everything that Olivia implies. I discuss this issue in the next section.)

I think, then, that moral discourse can make non-moral description more economical. However, this advantage will also be enjoyed by the conservationist, so we need to see which of these options is in the end most feasible or palatable.

To complete our discussion of fictionalism, we will need to discuss Joyce’s contribution to the debate, but before doing so I want to discuss two further matters. The first is whether fictionalism, if adopted, should be of pretence or non-pretence form. As we will see in 4.3, adopting a non-pretence form of fictionalism will not get the moral fictionalist what she wants. The second issue concerns whether or not we should employ the moral fiction self-consciously, which I discuss in section 4.4. As we’ll see, employing it self-consciously helps us to avoid one objection that has been made against fictionalism – that fictionalists are indistinguishable from realists – but problems arise if we are advocating fictionalism for its positive motivational effects (as I discuss in 4.5).

4.3.Pretence or non-pretence?

If we did decide to be fictionalists, should we adopt a pretence theory or a non-pretence theory? The two types of non-pretence theories we identified were expressivist fictionalism and the story-operator view. If you want your fictionalist utterances to be assertoric, then you can’t be an expressivist fictionalist. What about the story-operator view? There are three major problems with this.

If I say that Anna Karenina was killed by a train, my audience will probably understand me to mean that in the novel by Tolstoy, Anna Karenina was killed by a
train. But if I say that torturing the innocent for fun is wrong, then unless I do so in a sarcastic tone, or in some other way indicate that I’m not being entirely serious, then people will naturally think that I have made a sincere moral claim. It is common to discuss the content of literary fictions, and we understand from the context of discussion how to interpret utterances like the one about Anna Karenina. But most people do not expect to have to interpret moral utterances in a similar way. Thus if the suggestion is that the moral fictionalist makes moral utterances with unvoiced story-operators, she would need to explain that she was doing this to anyone who was not aware (at least if she didn’t want to be deceitful). This would be tiresome for both the fictionalist and her audience, and baffling for many people.

Another problem concerns disagreement. If I say that Anna Karenina was killed by a train, and you say that she drowned at sea, then at least one of us is wrong: in the novel, she only dies once. We could settle our disagreement by reading the relevant passage in the novel. But if I say that abortion is morally acceptable, but you say it is morally forbidden, then we may not be in disagreement if what we mean is In the fiction in which there are moral properties, abortion is acceptable/wrong. This is because we might not be referring to the same fiction. A pair of moral realists may be in disagreement over whether abortion is morally permissible, in the same way that a Muslim and a Christian can be in disagreement over whether Jesus is divine. A pair of atheists might argue over whether Jesus is divine in the fiction in which some entities are divine, but clearly there are many such fictions (even from a committed religious point of view, the

59 We might do this sometimes. Talking about the peculiarities of some of Kant’s substantive moral claims, you might say, “So it’s OK to kill bastards, but selling your own hair is morally questionable”. As it’s hard to say such things with a straight face, your audience would probably realise you weren’t being serious.
texts of other faiths must appear to be fictions). They can settle their argument by making clear whether they are talking about the orthodox Christian story, or the Muslim story, or some other story; and if they are still in disagreement, they can consult the appropriate texts. Are moral ‘disagreements’ between fictionalists going to take a similar form? Are fictionalists supposed to commit to one fiction over another (the pro-life or the pro-choice), and if so, why? It was suggested that going through ‘the’ moral fiction was meant to facilitate non-moral description, so I suppose that if we have that motivation, we should embrace the moral fiction that best accomplishes that. But how do we decide that? Perhaps fiction A will facilitate some descriptions better and some worse than fiction B. Should we alternate between them? How will we keep track?

And, of course, the biggest problem with the story-operator view is just that if all I intend to convey by saying that murder is wrong is that according to the moral fiction murder is wrong, then I convey nothing that could not be endorsed by an eliminativist.

What if we instead use a pretence theory? Does this help with the problems I identified for the story-operator type of fictionalism? The first problem was that most people would not know to interpret the fictionalist’s utterances as elliptical, unless they were told to. Arguably, we have a similar problem, in that most people would interpret you as genuinely asserting a moral proposition, if you were doing this in a context which did not prompt them to interpret it as a pretend-assertion. This consideration moves Joyce to say that there can be no honest “lone fictionalist” – his moral utterances would be systematically misunderstood. His solution is that “If [fictionalism] is to be viable it must be an attitude that a group may take towards a hitherto believed theory” (Joyce 2001: 204). The ‘may’ here is
a bit odd – why is fictionalism viable only if some group may adopt it? I think Joyce must mean that fictionalism will be viable only for and within groups of fictionalists, where each member will know how to interpret the moral utterances of the other members. But that would make fictionalism a very unattractive option. Firstly, the implication seems to be that if you are a fictionalist you should only make moral claims to people who don’t believe in morality, and scrupulously avoid doing so in the earshot of believers. That seems absurd. And secondly, it is pretty far-fetched to think that there is any time soon going to be enough people willing to join a fictionalist group to make the venture worthwhile. Error theory is not even a particularly popular thesis amongst philosophers, let alone the non-philosophical public.

I don’t think the problem of the ‘lone-fictionalist’ is any great problem, so long as the fictionalist can live with being to some extent misunderstood. Take a different discourse. Suppose that an error theory about mathematical objects is true, but that it is still useful – as it obviously is – to sometimes say things like “Seven is a prime number”. Perhaps most people commit themselves to something false when they make such utterances, but a revolutionary mathematical fictionalist, who only Pretends to assert it, escapes these commitments. Still, if the topic of conversation is not the ontology of mathematics, then what does it matter if the mathematical fictionalist is sometimes misunderstood? She has retained mathematical discourse because it has some useful function other than to assert the reality of numbers or sets or whatever. So long as it is this function which she is exploiting when she uses mathematical language with true-believers, then it is not clear that she is deceiving anyone. Is it her fault if her audience don’t realise that you don’t have to believe in mathematics to exploit it for descriptive purposes? I’m
not sure that the moral fictionalist will have it as easy as the mathematical fictionalist. Even if it is true that everyday mathematical utterances commit their utterers to a mathematical reality, it is hard to believe that people often make mathematical statements in order to assert the existence of such a reality. But people do often seem to really intend to attribute moral properties to actions and characters. So a moral fictionalist is likely to be interpreted as making similar attributions, by those who are not in the know about her pretence. But she has the option of just living with that. If Internalism and Rationalism are true, then she still communicates beliefs about motivations and reasons that she intends to communicate, even if her audience take her to be committed to something more.

The second problem that I raised above was that of disagreement. Here pretence theories clearly fare better than story-operator theories. Part of the problem with the story-operator view was that unlike most literary fictions, which at least contain a large body of propositions that are determinate in truth value within the fiction, there is no widely accepted body of moral propositions to serve as our canonical fiction. But with a pretence theory, we don’t need to know all the details of the fiction we are participating in. We can pretend that there is a set of objective moral truths (if we want to) but we don’t have to pretend that we know what all the members of this set are. This sort of situation can occur in games of pretence. If a group of us are playing Cluedo, we all know that it is true-in-the-game that the victim has been killed by someone, somewhere, and with something, but none of us know, at the beginning of the game, by whom, where, and with what. And at various stages in the game, different players will form different beliefs about these things.
Would disagreements between moral fictionalists be real disagreements, or pretend disagreements? If we are using the fiction to facilitate non-moral description, then we will want genuine disagreement to be possible. Fortunately, it seems that there can be genuine disagreement within a pretence. One of Walton’s examples is of a game of make-believe in which tree stumps represent bears (Walton 1990: 21ff). Any stump that you see when playing this game makes it fictional (i.e. true in the fictional world of the game) that there is a bear at that spot. Suppose a pair of children are arguing about how many bears there are in the churchyard: Tom says there are three, but Lucy thinks there are four. This is a genuine disagreement within a pretence. They are disagreeing about what they ought to pretend – which is determined by the number of stumps in the churchyard, in conjunction with a principle of generation connecting stumps with fictional bears – but they do not need to step out of their pretence in order to express their disagreement.

What would genuine disagreements within a moral fiction be about? If we say that it is true, in the moral fiction, that an agent has a moral reason to \( \varphi \) if and only if she has, in reality, a non-self-interested reason to \( \varphi \), then disagreements about whether it would be wrong for a particular agent to \( \varphi \) would reflect a genuine disagreement over whether that agent has a non-self-interested reason to \( \varphi \).

Finally, the third charge against the story-operator view was that those employing it would say nothing that could not be endorsed by an eliminativist. But obviously the eliminativist cannot endorse our saying that some acts are right or wrong, even within a pretence, as the eliminativist wants us to move away from using moral language, except to talk about our former beliefs and practices, or those of others. (Obviously, they may not object to our making pretend moral
claims when doing things like performing plays.) For all these reasons, it is clear that the moral fictionalist should adopt a pretence form of fictionalism.

4.4 The self-conscious fictionalist

Hermeneutic fictionalists do not usually think of typical participants of their target discourses as self-consciously engaged in pretence. It would be difficult to motivate the claim that maths teachers conceive of their role as getting pupils to pretend that certain things are true, such as that seven is a prime number. However, it may seem that if we adopt a revolutionary pretence fictionalism, we will obviously be aware that we are engaged in a pretence when we make our fictional claims. As we shall see in the following section, Joyce challenges this claim. But if we do conceive of the fictionalist as self-conscious, this will enable us to deal with an objection that has sometimes been made against fictionalism, namely that they cannot maintain the distinction between beliefs and acceptances, which they need to if their project is to succeed. So that we can appreciate this objection, let me go over what the distinction is supposed to be.

If we are realists about the subject matter of a discourse, then we can say that the following relationship holds between belief and assertability: a sentence is assertable, for me, if and only if I believe its propositional content to be true (where assertion is understood to be the act of putting a proposition forward as true). I am ignoring factors such as whether the sentence would be polite, or conversationally appropriate to utter. Fictionalists need an analogue of this relationship for fictionalist discourses. Fictionalists want us to avoid asserting the propositional content of the discourse’s sentences; instead, following Rosen (1990) they often talk of ‘quasi-assertion’. The speech act that is thought to be performed when
making a fictionalist utterance varies between different versions of fictionalism, so
‘quasi-assertion’ can be thought of as an umbrella term for whatever speech act the
fictionalist thinks is undertaken when one makes a fictionalist utterance of an
indicative sentence. We now need to specify the conditions under which a
fictionalist sentence is quasi-assertible. Fictionalists have sometimes claimed that
as assertions stand to beliefs, so quasi-assertions stand to ‘acceptances’, where an
acceptance is a propositional attitude distinct from belief in the proposition
expressed by the sentence. A sentence is quasi-assertible, for me, if and only if I
accept its propositional content. In *The Scientific Image*, for instance, Bas Van
Fraassen explains that acceptance of a scientific theory has both an epistemic and a
pragmatic dimension. Acceptance involves the belief that the theory is *empirically
adequate*, and a “commitment to the further confrontation of new phenomena
within the framework of that theory, a commitment to a research programme, and a
wager that all relevant phenomena can be accounted for without giving up that
theory” (van Fraassen 1980: 88).

Now for the objection. Paul Horwich claims that there is no distinction to be
drawn between belief and acceptance:

This is a distinction without a difference … If we tried to formulate a
psychological theory of the nature of belief, it would be plausible to
treat beliefs as states with a particular kind of causal role. This would
consist in such features as generating certain predictions, prompting
certain utterances, being caused by certain observations, entering in
characteristic ways into inferential relations, playing a certain role in
deliberation, and so on. But that is to define belief in exactly the way
instrumentalists [for which we can read ‘fictionalists’] characterise
acceptance. (Horwich 2004: 89)

And Gideon Rosen and John P. Burgess have raised similar concerns about the
notion of acceptance as used by mathematical fictionalists:
One doesn’t have to be a behaviourist to think that when a person understands a sentence $S$, confidently affirms it without qualification and without conscious insincerity, organizes serious activity just as would be done if $S$ were believed, and so on, then we have a powerful case for attributing to that person a belief that $S$ ... Attitudes cannot be brutally different. There must be some difference between a believer and a mere accepter, and if there is such a difference, surely it is reasonable to expect that its presence will be betrayed somehow in behaviour. (Rosen and Burgess 2005: 526)

There seems to be an easy response available to the Rosen and Burgess objection. There will be a behavioural difference between accepters and believers if, when asked if they believe that the theory they put forward is true, accepters say ‘no’ and believers say ‘yes’ (Daly 2008: 428). Horwich attempts to forestall this move. The difference here is not, he says, one between believers and accepters, but “between belief on the part of those who are not confused about their psychological states, and belief on the part of those who have been so muddled by philosophical double-talk that they are mistaken about the right way to describe their psychological state” (Horwich 2004: 90). However, as Chris Daly has noted, this puts the fictionalist in a no-win situation. Her opponent says that there is a difference between believers and accepters only if there is a behavioural difference; the fictionalist offers one, but is told it doesn’t count. Indeed, any behavioural evidence that the fictionalist cites will likely get the same response (Daly 2008: 429). Not only does this start to look dogmatic, Daly points out that it isn’t a line that Horwich can consistently defend. The fictionalist wants to distinguish the truth of a theory from its utility. If she is successful, the following claims can differ in truth value:

1. $T$ is true
(2) T is useful.

Horwich thinks this distinction is illegitimate: to believe (1) is to believe (2). Identical beliefs have identical contents, so if Horwich is right, (1) and (2) are identical in content. But this contradicts Horwich’s rejection of pragmatism about truth:

Although there is indeed an association between the truth of a belief and its tendency to facilitate successful activity, the tightness of the association should not be exaggerated. After all, actions based on true belief can none the less work out badly. (Horwich 1998: 9)

Convincing though Daly’s responses are, they will only work if the speaker in question is aware that she is speaking fictionally. Nolan et al.’s fictionalism seems to be one where such an awareness is present, but this is not the case for the other forms of moral fictionalism we discuss in this chapter. As we shall see, Joyce holds that there may be no behavioural or phenomenological distinction between a moral believer and a moral fictionalist, at least when the fictionalist is immersed in the fiction; and Mark Kalderon concedes that even if fictionalism is the correct account of moral discourse, participants in that discourse mistakenly think that they are asserting sentences they believe, rather than quasi-asserting sentences they accept. So the Horwich/Rosen and Burgess objection may have some bite in these cases, if we cannot find some other way to distinguish acceptance from belief.

I have argued that two of Nolan et al.’s attempts to motivate fictionalism over eliminativism don’t work. It is not the case that the relative ease of adopting fictionalism over eliminativism counts in its favour (even if it is actually easier). Nor is it the case that eliminativists need find themselves entangled in distracting metaethical debates, at least no more so than the fictionalist will. I did find some
potential for their claim that moral discourse may allow us to express certain non-
moral facts more conveniently than if we dropped moral discourse altogether.
Note, however, that we would get all the advantages that Nolan et al. think
fictionalism gives us if instead of adopting fictionalism, we simply decided to go
on believing the moral judgements we were inclined to make. This would seem to
be psychologically convenient, to allow us to avoid distracting metaethical debates,
and allow us to communicate certain non-moral beliefs. I will defend taking this
option in section 4.6. But first let us turn to look at Joyce’s proposal.

4.5. Richard Joyce’s moral fictionalism
Like Nolan et al., Joyce defends revolutionary moral fictionalism, but unlike them,
he emphasizes the positive motivational qualities of morality. According to Joyce,
thinking about our behaviour under moral concepts has practical benefits, because
it makes us more likely to behave rationally than non-moralised practical
deliberation. So Joyce needs to defend the following claims:

(1) Thinking about our behaviour under moral concepts makes us more likely
to behave rationally than non-moralized practical deliberation;

(2) Moral thinking can retain this useful function, even if we have given up
moral beliefs.

As we saw in the previous chapter, Joyce also defends the claim that:

(3) It is irrational to go on believing a proposition we have discovered to be false.
Note how important (3) is in motivating fictionalism for Joyce. Someone persuaded by (1) and (2) might still wonder why we should change our attitude to moral discourse: “Moral thinking is useful and we currently think morally? So why rock the boat?” (3) sounds plausible at first blush, but we saw in the previous chapter how the importance of ridding ourselves of false beliefs can be overstated. And we saw evidence that false beliefs – such as inflated estimates of one’s own abilities, and of how one is perceived by others – can have positive effects on mental health. But I shall not dispute (3) further in this section, because even if (3) were true for a great many propositions, (2) will turn out to be highly dubious.

I shall not dispute (1), because I think Joyce presents a very plausible case for it. I shall get to what that is presently. The main focus of this section will be (2), and I shall argue that Joyce does not present a compelling case. In particular, I shall argue that if we continue to use moral discourse in the way that Joyce recommends, then it will not be plausible to describe us as taking a fictionalist attitude to the discourse.

In 4.5.1 I set out Joyce’s motivation for fictionalism. I begin 4.5.2 by explaining why, according to Joyce, we needn’t believe the fiction in order for it to have a positive motivational effect, before showing how Joyce fails to provide a convincing case for this. Then in 4.5.3 I show why we cannot get the motivational advantages of the moral fiction if we employ it self-consciously. This allows us to conclude that Joyce’s fictionalism fails.

4.5.1 Joyce’s motivation for fictionalism
Let us first look at the case for (1). Joyce begins his case by reminding us of a well-known fact about the Prisoner’s Dilemma (PD): in a one-off game, one gets a greater payoff by defecting, no matter what one’s opponent does, despite the fact that the payoff for each cooperating is greater than the payoff for each defecting (Joyce 2001: 206 ff.). Agents in a Hobbesian state of nature appear to be in a PD situation, for which Hobbes’s solution is that each agent transfer their natural right to self-preservation to an all powerful sovereign. Having done this, the payoff for defecting severely diminishes. Cooperative behaviour becomes more attractive, not because subjects become more altruistic, or develop a sense of duty, but purely because it is in their own self-interest to cooperate. One of the problems with this solution is that however powerful the sovereign is, she will not be omniscient. It will always be possible to get away with a little self-serving, uncooperative behaviour without getting caught.

In iterated PD games, cooperative strategies are more successful. However, such games lack a feature present in the real world: the possibility of agents defecting without getting caught. Given that this possibility is available to us, we might be tempted to think that the most rational thing to do is to be cooperative when all eyes are upon us, but to defect whenever we get the chance. However, Joyce quotes with approval Hume’s warning that those tempted by this path are betrayed by their own maxims; and while they purpose to cheat with moderation and secrecy, a tempting incident occurs, nature is frail, and they give into the snare; whence they can never extricate themselves, without a total loss of reputation, and the forfeiture of all trust and confidence with mankind. (Hume 1777: 283)

60 Joyce cites Axelrod (1984) to support this claim.
Even if it would be most rational to defect when we could be sure of getting away with it, given our epistemic fallibility, and our tendency to risk long term benefits for paltry short-term gains, the most rational thing for us to do is to develop robust cooperative tendencies, which continue to constrain our behaviour even when we think we are not being observed. This is where moral thought comes in. Joyce claims that morality acts as a sort of internal sovereign:

If one has a tendency to think of the valued action in terms of something that “just must be done,” that is somehow required regardless of whether it suits you, then one is less likely to yield to the temptation of refraining. The stronger the terms in which the value of the action, or its sense of “requirement” is conceived, the less likelihood there is that akrasia will win out. This, I believe, is an important instrumental value to moral beliefs: they are a bulwark against the temptations of short-term profit. (Joyce 2001: 212-13).

Thinking of certain forms of behaviour as categorically forbidden (or categorically required) is conducive, on the whole, to a more successful life. I find all of this quite plausible (and if this sounds like an unpleasantly self-centred way to defend moral thought in this way, we needn’t insist that this is the only value that moral thought has; we can, for instance, observe that something that helps us to avoid self-destructive behaviour may also help us to avoid behaviour that is harmful to others we care about, including, if we like, the whole of humanity). So let us grant that (1) is true. But Joyce devotes over a hundred pages of The Myth of Morality to convincing us that no actions are categorically forbidden or required, and also claims that it is always instrumentally rational to have true over false beliefs, when the truth is accessible to us (178-79). So doesn’t believing error theory rob us of the practical benefits of moral thought?
4.5.2 Why, according to Joyce, we needn’t believe the fiction

It need not, according to Joyce, because we don’t have to actually believe in categorical imperatives in order for thoughts about them to have the positive effects just discussed. Joyce thinks that a fictionalist, immersed in the moral fiction, will behave much as an ordinary user of moral discourse will behave. That raises the question of why we should not attribute moral beliefs to this so-called fictionalist. Recall that in response to the Horwich/Rosen and Burgess objection, we saw that Nolan et al. could point to a behavioural difference between believers and accepters: believers, if sincere, would affirm that the moral sentences they asserted were true, while accepters would affirm that the sentences they quasi-asserted were untrue. Joyce, however, thinks that there need be no phenomenological difference between a believer and a fictionalist immersed in the fiction, and indeed it is helpful to his case if there is no such difference, because it makes it easier to believe that the moral thoughts of the fictionalist will be motivating:

What goes through [the fictionalist’s] mind may be exactly what goes through the mind of the sincere moral believer – it need not ‘feel’ like make-believe at all (and thus it may have the same influence on her behaviour as belief). The difference between the two need only be a disposition that the fictionalist has (though is not paying attention to): the disposition to deny that anything is really morally wrong, when placed in her most critical context. (Joyce 2005: 306)

Her “most critical context” is the context in which she is most “undistracted, reflective, and critical” (Joyce 2001: 191), and Joyce claims that what she is prepared to assent to/dissent from in her most critical context determines what her
actual beliefs concerning morality are, even though “[s]he might … credit herself with [moral beliefs] if she is not thinking carefully about the matter” (193). So there is, according to Joyce, a behavioural difference between believers and accepters, even though it does not manifest itself in all situations. I find this way of drawing the distinction dubious, however.

Consider the kind of unhealthy thought one might have in a despondent mood. Feeling insecure about my athletic ability and ruminating enviously about the ability of some rival, I get to thinking it unfair that he is so blessed in comparison to me. Now, if the differences in question are innate (we have had access to the same amount and quality of training, nutrition, healthcare, etc.) it is not really the case that any unfairness is involved, for I cannot sensibly claim that I have a right to be as athletic as my rival. Perhaps a counterpart of mine in antiquity would have cursed the gods for his misfortune, but that isn’t an option for me. When I am at my most undistracted, reflective, and critical, I shall (hopefully) dismiss the thought of unfairness. It doesn’t really matter if you disagree with me, and think that it is for some reason unfair that some people are more physically advantaged than others; the point is only that when I give the matter my keenest attention, I conclude that there is no unfairness involved, and that, for Joyce, determines what I believe.

So should we say that I did not really believe that there was an unfair difference between me and my rival when I was despondent? I think we do often explain ourselves in this way. We say something unpleasant in a sour mood, for instance, which we later regret, and we excuse ourselves saying, “I don’t really believe that, I was just angry”. But is it that our opinion changes with our mood, or
that our *real* opinion is determined by what we are prepared to assert in our most critical context?

Joyce does say that our most critical context should be thought of as a context that we have actually occupied and not one that is merely hypothetical. So with regard to some proposition \( p \), it is not the case that I believe \( p \) just because, if I were to give it my deepest scrutiny, I would come to believe it (192). So Joyce could say that when I was despondent I believed that it was unfair that my rival was more blessed than me, but I corrected this belief when I came to my senses. But Joyce thinks that once we have decided what to believe on some matter in our most critical context, this fixes what we believe on that matter (unless, I suppose, we should later revise our belief in a similarly critical context).

However, what if sometime later I feel despondent again and, having entirely forgotten my conclusion that I don’t have a right to be as athletic as my rival, I again get to thinking how unfair my lot is? I should not, following Joyce, be described as really believing that it is unfair, as that is not the belief that I formed when I considered the matter in my most critical context. How then should my thoughts on this occasion be characterised? Surely I’m not pretending. After all, why say I’m not pretending during my first attack of despondency, but that I am pretending during my second, when my thoughts in my most critical context have no more influence on me on the second occasion than they do on the first? The only difference I have mentioned between these two episodes of despondency is that one occurs before and the other occurs after I consider the matter of fairness in my most critical context, and it seems mysterious how this temporal ordering determines the sort of attitude I am taking to the proposition that my lot is unfair. I conclude, then, that as the most reasonable thing to say about my thoughts during
my first despondent episode is that I believe my lot is unfair, that is the most
reasonable thing to say about my thoughts during my second despondent episode.

If that doesn’t convince you, then consider someone whose cognitive abilities
have been seriously impaired. There are severe and irreversible cases of this, such
as people who suffer from Alzheimer’s disease. In advanced stages, these people
will never be able to return to their most critical context. Or there are mental health
conditions that are episodic, such as depression and schizophrenia, where the
sufferer’s thought processes can be severely distorted for a period of time, but
recovery is possible. We clearly don’t want to deny that people suffering from
severe mental health conditions have beliefs (unless we are sceptical about
propositional attitudes across the board), but the beliefs they do have may be
greatly at odds with those they had when in better mental health. Joyce could say
that we should count someone as believing $p$, if they would assent to $p$ in the most
critical context that is available to them at the time in question. Then we could say
that, as someone suffering from an episode of poor mental health is unable, at least
for the time being, to revisit the most critical context they had when they were
healthy, that context is irrelevant to determining what they believe now. Fine, but
then we can equally say that someone in the grip of a moral thought – incensed,
perhaps, at some injustice he takes to have been perpetrated against him – is also
unable to enter his most critical context, if he is an error theorist in that context,
and remain in the grip of the moral thought. Indeed, the moral thought may be what
is precluding him from entering that context. So his most critical context is
unavailable to him while he remains under morality’s spell, and we should count
him as believing that an injustice has been done to him.
Joyce’s attempt to distinguish believers from accepters by considering what they are prepared to assent to in their most critical contexts fails. If we are to be described as employing morality as a fiction, we shall have to employ it self-consciously as a fiction. But as we shall see next, fictions are not motivating when they are self-consciously employed.

4.5.3 Why can’t we self-consciously employ the fiction?

Joyce never really explains why thinking of certain actions as being categorically forbidden should motivate us to avoid doing them; however, in “Moral Fictionalism” (Joyce 2005) the suggestion seems to be that it is through affecting us emotionally that the moral fiction would motivate us to act one way rather than another. The thought that an action would make us feel guilty or ashamed is a powerful disincentive, as are emotionally-laden thoughts about how others would judge us. Presumably it is thoughts like these that Joyce has in mind. Joyce defends the view that paradigmatic fictions, such as films and novels, arouse genuine emotions, and that our emotional reactions to them can be beneficial, and motivational: “Reading *Anna Karenina* may encourage a person to abandon a doomed love affair; watching *The Blair Witch Project* may lead one to cancel the planned camping trip in the woods” (Joyce 2005: 303). Nolan et al. make a similar point in an earlier version of their paper:

> If people can be moved to support generous welfare policies by reading *Grapes of Wrath*, or to oppose genetic testing by reading *Brave New World* etc., how much more might they be concerned to “get things right” with respect to the story of goodness, justice and decency, to which they are deeply attached? (Nolan et al. 2002: 42)

---

61 See Joyce (2000), where he argues against Colin Radford’s view that the emotions aroused by fiction, while genuine, are irrational.
I won’t dispute that novels and films can have beneficial effects on us, or that they do so by engaging our emotions. Nevertheless, if we look at these examples in a little more detail, it should be obvious that it is not the fiction that is motivating the reader/viewer, but her beliefs and desires about the real world. If reading *Anna Karenina* causes someone to re-evaluate her romantic attachments, then this is because it makes her realise something that is true about her own life – e.g. that she is in a doomed relationship. Watching *The Blair Witch Project* might make a timid camper reflect that the *real* woods are a frightening place at night, and there is a possibility of getting lost. Reading *Brave New World* might bring home to someone how much she values natural human development over the sort of controlled development described in that novel, and to realise that what she values may be lost if genetic research in the real world continues to advance. In each of these cases, the fiction is prompting the reader/viewer to make a factual inference or to consider what she really values, and it is these things that are motivating her to act.

It is true that having her emotions engaged may be critical to her drawing this inference or considering what she values. Indeed, this seems to be what is going on with someone who comes to support welfare policies after reading *The Grapes of Wrath*. The novel helps them to see how awful dire poverty is, and arouses their sympathy for members of society who are experiencing dire poverty, or would do if the welfare system were weakened. But it is clearly not the fictional suffering that occurs in the novel that is the motivating factor here, but the real suffering of actual people (which the novel helps to make more vivid to the reader). You might object that fictions are motivational by noting that an agent A being motivated to φ is standardly analysed as A having a belief-desire pair:
*Motivation:* \( A \) is motivated to \( \varphi \) if and only if \( A \) desires to \( \psi \), and believes that by \( \varphi \)-ing she will \( \psi \).

Then you might say that the fiction, by altering what the reader values, causes her to desire to \( \psi \), and thus is at least part of what motivates her to \( \varphi \). But that isn’t right. We can distinguish between the causes of the agent’s beliefs and desires from the beliefs and desires themselves. It is her beliefs and desires that motivate the agent, not the causes of her beliefs and desires.

It would be natural to say the fiction was motivating her if some amongst the combination of beliefs and desires that motivated her were *about* the fiction. Perhaps someone could be motivated to improve the plight of the poor after reading *The Grapes of Wrath*, despite utter ignorance of the current condition of the nation’s poor. But in a case like this, the fiction would only be sparking the reader’s interest in a matter she had not considered before. The thought might be, if there are any people living in such conditions then I will do something about it. What this person decides to do will depend on how she takes things to really be and her desires will concern actual people. It is true that someone might have a distorted picture of some social phenomenon because one of their sources of information for it comes from a fictional source, and to that extent the fiction may be motivating them, rather than reality. But that will be a case of them mistaking fiction for reality. After all, it doesn’t make much sense to say that we ought to do such and such because of the way things are *in some story!*
These considerations show us that will not be able to get the positive motivational effects of thinking about our actions under moral concepts if we self-consciously think of moral claims as only fictionally true.

Joyce’s proposal that fictionalism allows us to retain the motivational benefits of morality without incurring the cost of believing falsehoods is unsuccessful. His identification of beliefs with the attitudes that we have when we assent to propositions in our most critical context is not convincing. And it does not seem that we would gain the motivational benefits of moral thought if we were self-conscious fictionalists. For these reasons, Joyce’s fictionalism does not succeed.

However we would continue to gain these benefits if we continued to believe that certain acts were morally required or forbidden. In the next section I defend conservationism, the position that we both can and should continue to have moral beliefs, even if we are convinced of error theory.

4.6 Conservationism

We can be motivated by fictions only if we do not conceive of them as fictions. If moral motivation is desirable – at least for creatures like us – then we shall get what we desire only if we have moral beliefs. But can we go on having moral beliefs if we accept the error theory? And if we can, is it desirable to go on in this way? The conservationist is an error theorist who answers ‘yes’ to both these questions.

The first thing that I want to do is give some additional reasons for thinking that moral discourse and practice is worth preserving. Then I will address some objections to conservationism.
Joyce defends morality on the grounds that thinking of certain actions as categorically forbidden or required prevents us from behaving rashly in the face of temptation. As I indicated above, I think this is very plausible. But there are other advantages that might be cited. I will name four more.

The second advantage comes to mind when we recall that a common answer to the question, “why be moral?” is that without morality society would not function: we would not be able to trust others to keep their promises, or to keep their hands off our possessions, or to behave in a civilized fashion. It is a mistake to think that we succeed in being moral if we accept certain restrictions on our behaviour because we perceive that it is in our own self-interest to do so (that is only to succeed in acting out of self-interest). But it seems sensible to want there to be devices in place to make social life possible and prosperous. If thinking and acting in accordance with what we take to be moral is such a device, and if it seems likely to succeed better than non-moralized practical deliberation, then that is a reason to retain a commitment to morality. This, it seems, was Mackie’s main motivation for retaining moral discourse and practice:

Men sometimes display active malevolence to one another, but even apart from that they are almost always concerned more with their selfish ends than with helping one another. The function of morality is primarily to counteract this limitation of men’s sympathies. (Mackie 1977: 108)

The third advantage is related but independent. We can begin to see it when we ask why we are moved by moral considerations. Some have thought it is because we fear punishment – from God, from our rulers, or from our neighbours – if we act immorally. But even if that is part of the answer, it can’t be the whole answer. Atheists can be motivated by moral concerns, and people have often been moved by
considerations that they consider moral, even though they risk punishment from their rulers and neighbours, and from religious authorities, for so acting. A second explanation is that we gain benefits by being moral – e.g. the benefits of living in a prosperous society, that have just been mentioned. But people can be moved to act morally (as they see it) even though they predict that they will be materially disadvantaged by doing so, or even at great risk to their own lives. Whilst fear of punishment and hope of advantage can be powerful motivations towards acting in accordance with morality, neither need be present. What other motivation can there be? John Stuart Mill supplies the answer:

The internal sanction of duty, whatever our standard of duty may be, is one and the same – a feeling in our own mind; a pain, more or less intense, attendant on violation of duty, which in properly cultivated moral natures rises, in the more serious cases, into shrinking from it as an impossibility. (Mill 1863: 161)

We are motivated by our conscience to act in the ways that (as we see it) we are morally required to act. We cannot live easily with ourselves if we act against our conscience, and we feel proud of ourselves if we act in accordance with it, despite strong temptation to do otherwise. Giving up morality means abandoning one’s conscience, or – if we cannot – viewing it as an irrational irritant. I remarked in the previous chapter that having a conscience seems to us a burden worth bearing, and one that we would wish others to bear. We don’t want to be like those who lack a conscience (psychopaths and infants were the examples I gave). But we might be moved to see our conscience not merely as a burden (as it may sometimes seem), and not merely as an asset (as it may well be), but as something the having of which is essential to our self-conception. We are concerned to behave well, not just
because of how others will think about us and react to us, but because of how we will think about ourselves if we behave badly.

That might sound off-puttingly self-centred – I want to retain moral thought because I want to think of myself as good – but there is more to it than that. I can want to retain moral thought because I want to think of you as good and for you to think of yourself as good. Or it might be thought that, once the scales have fallen from our eyes, we will no longer value being seen as morally good, or seeing others in that way. Perhaps people drawn to eliminativism feel that way, but it doesn’t seem inevitable.

The fourth advantage is that conservationism is psychologically less demanding than fictionalism. I argued above that this could not be cited as an advantage that fictionalism has over eliminativism, as they are not means to the same end. But fictionalism and conservationism are means to the same end – the end of retaining the benefits of moral discourse despite an acceptance of error theory – so the relative ease of adopting conservationism over eliminativism (if it is easier) does speak in favour of it.

The fifth advantage has been discussed earlier in this chapter. I claimed that fictionalism might be motivated on the grounds that moral utterances have non-moral implications – about the speaker’s motivation and the speaker’s beliefs about reasons – that might be more conveniently expressed in moral terms than in non-moral terms. This advantage comes with conservationism as well.

4.6.1 Objections to conservationism
An obvious objection is that the conservationist displays a distinct lack of philosophical integrity by defending an outlandish metaethical theory, but refusing to accept the consequences of it.

In defence the conservationist can say that she accepts the consequences of error theory in the appropriate contexts, such as when she is discussing or writing about metaethical issues. The norms that we accept as part of joining the philosophical conversation on metaethics include the norm of consistency, so the error theorist is obliged to hold, in the philosophy room, that no sentence attributing a moral property to an object is true. This is obvious – if she isn’t consistent in this way, then she doesn’t defend a coherent thesis. But the norms that we accept when we join in conversations about first-order moral issues – whether with philosophers or non-philosophers – do not require us to express only those views that are consistent with our metaethical views, as many participants in the first-order debates will not have settled views on metaethical matters.

Perhaps the objection is not that there is a norm dictating consistency between first- and second-order ethical views, but just that there is a norm dictating consistency across all of one’s opinions. One reason for this norm might be that it is disrespectful to one’s interlocutors to rely on evidence or principles in one debate (first-order ethics) that you swiftly disown when the subject changes (to metaethics). How are they supposed to keep track of what you really think? But the conservationist can answer that when she is doing metaethics, she really thinks that nothing has moral properties, but when she is discussing ethics in other contexts, she really thinks that some actions are right and others wrong.

Consistency in opinion may have *prima facie* value, but that can be overridden in the circumstances that one holds a philosophical opinion that conflicts with how
we are accustomed to think in our everyday lives, but it is desirable to go on thinking in the way we have become accustomed to in everyday contexts. We can achieve consistency by either rejecting the philosophical opinion, or changing the way we have become accustomed to think. But to do either of these is to lose something of value. (Of course, you may not think the philosophical opinion in question is worth anything if you don’t share it, but it is valuable to the person who holds it.)

It is not as if it is only moral error theorists who are in this position. There are many philosophers who have held positions that seem to defy commonsense, such as that time does not exist, or that the past does not exist, or that nothing persists unperceived, or that there are no composite objects, or that beliefs, desires, and other propositional attitudes do not exist. I find it difficult to believe that one can consistently maintain any of these opinions when engaged in non-philosophical conversations that concern time, or composite objects, or beliefs. But I can’t see why a philosopher who, for instance, denies the reality of composite objects, should be pressed to be consistent in this view when she is, say, ordering furniture for her office. Or why she should adopt a fictionalist attitude, only pretending that she wants a desk. Neither courses of action seems necessary, because her metaphysical position just seems irrelevant in the context. It is reasonable to expect her to have a consistent view of the reality of composite objects when she is doing metaphysics, but not throughout her day. I would suggest that the same goes for the moral error theorist.

The second objection follows on from the first. It might be accepted that conservationism is appropriate if being conservative about our beliefs makes no practical difference, but the nature of morality is such that it inevitably will. Moral
judgements are, at least in large part, judgements about what we ought to do and refrain from doing. If none of them are true, the second objection goes, then we need some other way of thinking about how to behave. As Joyce emphasizes, the problem with moral judgements is their claim to be categorical: to apply to all agents, regardless of their desires or interests. But agents vary in their desires and interests in such a way that we cannot believe that there are any categorical normative facts. In light of this, shouldn’t we relativize our normative judgements to individual agents?

The first thing to say in response to this objection is that we do, of course, relativize many of our normative judgements to individual agents, because what anyone should be doing at any particular time usually depends on contingent facts about their wants and needs (not necessarily selfish wants or needs). However, it makes sense, from a social point of view, to make some demands that apply to everyone, such as demands that individuals not be aggressive or dishonest. And it is these sorts of demands that get moralized: we think that everyone ought not to be aggressive or dishonest (at least in most circumstances), whether it suits them to refrain from aggression or dishonesty or not. It seems clear why we, as a society, make these demands universal. It is very important to the integrity of a society that its members are generally peaceable and trustworthy, and although societies obviously can survive some degree of free-riding, all free-riding is damaging, and there is no principled reason for allowing some citizens to free-ride and not others. If error theory is true, we have no moral right to demand that anyone refrain from aggressive or dishonest behaviour, but there is a pragmatic justification for making such demands of everyone, and for making them using moral terms (as that they are more likely to be motivationally efficacious when expressed in those terms).
And the second thing to say is that although we should not expect there to be full convergence in our judgements about practical reasons, even given flawless deliberation, I don’t think it is at all naive to expect there to be a great deal of convergence under those conditions. Joyce’s argument against categorical imperatives only requires the possibility that, for any proposed categorical imperative, there could be an individual with a good reason to flout it. But many – hopefully most – of us want to live by certain principles that we wish ourselves to be held to as much as we wish others to be held to, those principles being those that make social life possible and pleasant. This is why I think Michael Smith’s optimism that there will be complete convergence in what we would desire to desire, if we were fully rational, is not entirely misplaced. He’s too optimistic, because there will always be some individuals who care nothing about exploiting others, and could not be brought to care. I suggest that we tone our optimism down slightly, and only claim that there would be such a convergence amongst those of us who are concerned to live by principles that make social life possible and pleasant. But isn’t that most of us?

The last objection that I will consider is that conservationism leaves those who adopt it with contradictory beliefs, such as the belief that that our actions both have and entirely lack moral properties. Believing in contradictions seems a paradigm of irrationality. Can we avoid this consequence? Yes, because it seems possible that a person could believe (i.e. be disposed to sincerely assert) $p$ and believe (i.e. be disposed to sincerely assert) $\neg p$, but never be disposed to sincerely assert $p \& \neg p$. This would be so if the individual was disposed to assert $p$ under some conditions, and $\neg p$ under different conditions. In the conditions under which she was disposed to assert $p$ she would not be disposed to assert $\neg p$, and in the conditions under
which she was disposed to assert \( \neg p \) she would not be disposed to assert \( p \). Such a person has beliefs that are contradictory, but he does not believe a contradiction.

How can this happen? Someone convinced of error theory when engaged in metaethical reflection sees the weight of evidence to favour the proposition that no positive moral judgements are true. But that same person may feel quite differently when reflecting on people’s actual behaviour. When thinking about metaethics, one thinks quite abstractly about the nature of moral judgements, the meaning of moral terms, and the ontology that would have to be in place to make those judgements true, and give those terms an extension. In the error theorist’s case, this leads her to believe that moral discourse is systematically defective. But when considering actual events, that same error theorist is subject to different considerations and different types of evidence. For example, thinking about cases where one individual, for self-interested reasons, deliberately causes suffering to another individual tends to elicit, in most people, feelings of blame. Thinking about cases where one’s own actions have caused others to suffer, or for others to mistrust one, or to have various other negative feelings about one’s character, tend to elicit feelings of guilt. Feelings of blame and guilt can be taken as evidence that certain moral propositions are true. After all, those feelings only seem appropriate if certain moral norms have been violated. We can find further support for our moral opinions by finding them to be shared by others, especially when they are shared by those we take to be judicious.

I don’t think it is at all implausible to think that we have this tendency to go back and forth in our opinions, having a fairly settled opinion in one context that is reliably overridden in another. Scanlon provides a nice illustration of this:
I may know … that despite Jones’s pretensions to be a loyal friend, he is in fact merely an artful deceiver. Yet when I am with him I may find the appearance of warmth and friendship so affecting that I find myself thinking, although I know better, that he can be relied on after all. (Scanlon 1998: 35)

In this sort of case, it would of course be preferable not to be fooled by the appearances. But to allow ourselves to be taken in by moral appearances – that some actions are non-optionally required of us, and that certain emotions are apt responses to compliance or non-compliance with those requirements – is, I think, the best response to the error theory. I contend, then, that conservationism has all the advantages of fictionalism, and that objections to it can be met.
5: Summary and Conclusion

On the most plausible view of the way we use moral language, indicative moral sentences are used to assert that objects have moral properties; that those moral properties are objective; that they are the source of reasons for us to behave in some ways rather than others; that the moral judgements we make have implications about what we are motivated to do; and that moral properties are determined by non-moral properties. As we saw in chapter one, this combination of features requires us to say either that moral properties are non-natural, or that our moral obligations can be determined merely by reflecting on our status as rational agents. The view that there are non-natural properties, which we would need some mysterious faculty of moral perception or intuition to detect, is *prima facie* implausible. And it seems that we have good reason to doubt that such properties exist, as we can account for the various moral phenomena – such as our propensity to make and be influenced in various ways by moral judgements – without positing such properties. The view that moral obligations arise out of our status as rational agents is also doubtful, as there seems no reason to think that there would be a complete convergence in moral opinions if we were all to achieve maximal rationality (as this view requires).

In chapter 2 I considered a number of objections to error theory. Some of these traded in the idea that error theory is in some way incoherent. The self-undermining objection was meant to show that error theorists are committed to the truth of some of the first-order moral judgements that they claim to reject entirely. But that objection only goes through on the presupposition that there are such things as moral values (so it is ultimately question begging). It also implies an implausible scepticism about metaethics as a subject matter. The first partners in
guilt objection that I considered was that moral error theorists, for consistency, ought to accept a corresponding epistemic error theory, as epistemic properties and reasons are no more or less queer than their moral counterparts. But I showed how we can escape this conclusion by showing that there are no irreducible epistemic reasons: epistemic imperatives apply universally, but our reasons to abide with them are contingent on our non-epistemic goals. The second partners in guilt argument was that the reasons we have to abide by the hypothetical imperatives that apply to us are not wholly contingent on our goals or desire, but also on the fact that it is rational to comply with them. However, I found no reason to think that the fact that an action would be rational is, in itself, a reason to do it.

The other objections that I discussed in chapter 2 traded in our confidence in either the actuality or possibility of moral facts. Against the claim that error theory defies commonsense, I argued that appeals to commonsense are only reasonable ways to respond to sceptical theses if there is independent evidence for the phenomenon in question, which is not undermined by the sceptics argument. I showed that opponents of error theory do not have such evidence. And against the claim that the mere possibility of moral facts at world non-morally identical to our own (in conjunction with the supervenience of moral and non-moral facts) is sufficient to prove the actual existence of moral facts, I showed how each of the arguments that have been offered in support of error theory imply that there are no moral facts at worlds that are non-morally like our own, and thus that the objection begs the question.

In chapter 3 I began to consider what we should do with moral discourse if we come to believe error theory. The first option that I considered in detail was eliminativism – the idea that we should just purge the discourse. I found the
motivations for this option unpersuasive. Hinckfuss argued that morality is harmful to most members of society, because, amongst other things, it fosters elitism, exacerbates conflicts, and makes painful moral emotions possible. I found his claims about the role moral discourse and practice play in shaping the structure of societies unpersuasive, and I countered his examples of the harm that moral practice can do with examples of the ways that people have campaigned for more egalitarian societies under a moral banner. And I argued that the cost of ridding ourselves of moral emotions (if that could even be done) was that this would lead others to think that we no longer cared as much about the harm we might cause others. The other motivation I considered for eliminativism was the thought that it is always better to act on the basis of true rather than untrue judgements, and that if we allow ourselves to swallow the myths necessary to sustain moral beliefs, that will leave us ‘epistemological wrecks’. Against this, I showed how, at least with regard to certain matters, false beliefs can have positive effects on our mental health, and can make us more effective agents. I also found no reason for thinking that having moral beliefs is likely to make us ‘epistemological wrecks’. Clearly many extremely intelligent people have strong moral beliefs.

The second option that I considered in chapter 3 was that we could choose to become moral relativists. I argued that it would be very difficult to be an honest moral relativist, because that would involve only expressing moral opinions about people who fundamentally share our values, and there is a great diversity of values, even within a single society. If, on the other hand, we decided to be dishonest relativists, with the aim of persuading people that they had moral reasons they do not in fact have, we would be more successful if we actually believed these judgements.
In the fourth chapter I considered moral fictionalism. One of the possible motivations I considered for fictionalism was that moral language allowed us to express certain non-moral judgements, which would either be impossible to express in non-moral vocabulary, or which it would be more economical to express in moral vocabulary. I found no reason for thinking that there are any non-moral thoughts that can only be expressed with moral vocabulary, but I tried to show that statements about what it is right or wrong to do imply things about what we have reasons to do, or what we are motivated to do, and that using the moral terms allows for more economical expression.

The other motivation I considered was that thinking in moral terms can have positive motivational effects, helping us to avoid imprudent actions. Joyce claims that we could reap these benefits, even if we regard morality as a fiction, because in most contexts, the fictionalist’s thoughts about moral matters will be phenomenologically indistinguishable from the genuine moral believer. In Joyce’s view, the fictionalist’s beliefs about moral matters are fixed in her most critical context. However, I showed why this is not a convincing criterion for belief, by giving examples where it seems clear that a person has a belief, in a less critical context, that differs to her judgement on the same matter in her most critical context. I then went on to show why a ‘self-conscious fictionalist’ would not reap the motivational benefits that Joyce highlights. This was because cases where we seem to be motivated by fictions are really only cases where fictions prompt thoughts about reality, and it is the thoughts about reality that motivate us.

Lastly in chapter 4, I considered conservationism – the idea that we should just allow ourselves to continue to have moral beliefs, even if we are persuaded by error theory. I considered three objections to this position, such as that adopting it
would reveal a lack of intellectual integrity; that it would leave us with unreliable normative beliefs; and that it would leave us with contradictory beliefs. I showed how each of these objections can be dealt with.

Moral error theory strikes many as a metaethical theory of last resort. It can seem massively dispiriting – absurd even – to think that our most central moral beliefs are simply untrue. However, the attitude that we take to morality when we consider it at the metaethical level need not be the attitude that we take to it in our everyday lives. If there is value in our ordinary moral discourse and practice – and I agree with many philosophers that there is great value in these things – then that is something that we can preserve, whatever the outcome of our metaethical research.
Appendix

Non-cognitivists have made several attempts to solve the Frege-Geach problem, which I discussed in chapter 1(1.2.1). The used the following modus ponens argument to illustrate the problem:

(i) If it is wrong to tell lies, it is wrong to get your little brother to tell lies.
(ii) It is wrong to tell lies.
thus
(iii) It is wrong to get your little brother to tell lies.

If we just focus on the problem of accounting for the validity of modus ponens arguments containing moral premises, then the problem might be seen as that of accounting for the meaning of conditionals which embed moral statements. One early attempt at this was made by Simon Blackburn (1984). Blackburn thought that the way out of the problem was to expand our notion of the way we use the logical connectives, so that we can understand them as operating not only on truth-evaluable propositions, but also on things like imperatives and expressions of attitudes of approval or disapproval. He argued that we had to do this in the case of conjunctions, because ‘and’ could be used to link two commands, e.g. “fetch me a loaf and slice it thickly”. Similarly, he thought that conditionals could be seen as operating on attitudes. The key notion invoked in this explanation was that of a higher-order attitude. If ‘Lying is wrong’ expressed an attitude of disapproval towards lying (equivalent to ‘Boo lying!’) then “If lying is wrong then getting little brother to lie is wrong” expressed the higher-order attitude of approval of making disapproval of getting your brother to lie follow on disapproval of lying. On this
understanding, anyone who accepted the premises of the above argument, but

denied the conclusion would:

   By accepting (i): express approval of making *disapproval of getting little brother to lie* follow upon *disapproval of lying*.

   By accepting (ii): express disapproval of lying.

   By denying (iii): fail to express disapproval of getting little brother to lie.

Anyone in this position would fail to hold a combination of attitudes that he himself approved of. Blackburn intended his readers to see someone who accepted premises (i) and (ii), but failed to accept (iii) as making a *logical* mistake, but various critics have pointed out that the failure to have a combination of attitudes that one approves is not a failing in logic – but at most a moral or pragmatic failing. This point is forcefully put by Mark van Roojen, who asks us to consider the following argument (van Roojen 1996):

   (1) It would be wrong for me to believe ill of my friends (Premise).
   (2) My mother and father are my friends (Premise).
   (3) It would be believing ill of my friend to think that he had been duplicitous with another of my friends (Premise).
   (4) If the coded valentine is not a joke, my father is being unfaithful to my mother, and hence duplicitous (Premise).
   (5) The coded valentine is not a joke (Premise).
   (6) It is wrong for me to think that my father is being duplicitous with my mother (Conclusion from 1, 2 and 3).
   (7) My father is being duplicitous with my mother (Conclusion from 4 and 5).

The argument is valid, but it appears that Blackburn should hold the two conclusions inconsistent. After all, following his analysis in *Spreading the Word*, someone uttering (6) would be expressing disapproval of believing that his father was being unfaithful to his mother, while someone uttering (7) would be
expressing just this belief. But it is clearly not logically inconsistent to have a belief that one disapproves of. Moreover, if (6) and (7) were logically inconsistent, then this would have to be due to some inconsistency in the premises from which they follow. But (1) – (5) are clearly not inconsistent. The problem, as van Roojen diagnoses it, is that although the higher-order account of inconsistency seeks to mirror realist accounts of inconsistency (i.e. to find the same pairs of statements contradictory and to find the same arguments valid or invalid), by broadening the notion of logical inconsistency, it inevitably finds it where realist accounts would find none (van Roojen 1996, p. 330).

Blackburn revised his account in ‘Attitudes and Contents’ (Blackburn 1993b). Instead of seeing conditionals as expressing higher-order attitudes, he interprets them as expressing complex dispositional-states. In stating that “If lying is wrong, then getting little brother to lie is wrong”, I commit myself to either not accepting <lying is wrong> or to accepting <getting little brother to lie is wrong>. The idea then is that my statement expresses this disposition. Taking the argument (i), (ii)/(iii) again, what would it be, on the revised proposal, to accept the premises, but deny the conclusion? Well, if we look at each of the ways the conditional can ‘branch’, we find that, whichever branch we take, someone who denies the conclusion winds up with an inconsistent set of commitments:

1st branch

(i’) Not accepting <lying is wrong>.

(ii’) Accepting <lying is wrong>.

(iii’) Not accepting <getting little brother to lie is wrong>.
2nd branch

(i′′) Accepting <getting little brother to lie is wrong>.
(ii′′) Accepting <lying is wrong>.
(iii′′) Not accepting <getting little brother to lie is wrong>.

Whichever branch of the conditional we take, we end up with inconsistent commitments. This looks good as it stands, but of course, for the proposal to succeed it must not merely account for the validity of modus ponens arguments, but for all types of argument. Bob Hale finds problems with applying the account to modus tollens (Hale 2002). One can validly argue that lying is not wrong from the premises that If lying is wrong then getting your little brother to lie is wrong and Getting your little brother to lie is not wrong. One who denies the conclusion of this argument does not accept that lying is not wrong. But it seems that only one of the branches of the conditional gives us an inconsistent set of commitments:

Branch 1

(α) Not accepting <lying is wrong>.
(β) Accepting <getting your little brother to lie is not wrong>.
(γ) Not accepting <lying is not wrong>.

Branch 2

(α′) Accepting <getting your little brother to lie is wrong>.
(β′) Accepting <getting your little brother to lie is not wrong>.
(γ′) Not accepting <lying is not wrong>.

(α′) and (β′) are inconsistent, but (α) and (γ) are not, because it is not inconsistent to be undecided over whether or not something is wrong. Hale considers whether
this problem could be avoided by reinterpreting the commitment expressed by the conditional. What if accepting the conditional in question did not amount to committing yourself to

either not accepting \(<p>\) or to accepting that \(<q>\)

but rather to

either accepting that \(<\text{not } p>\) or to accepting that \(<q>\)?

Under this interpretation of the conditional, we do seem to get the required inconsistency, as we can see if we examine the different ways the conditional branches:

**Branch 1**

(a*) Accepting \(<\text{lying is not wrong}>\).

(β*) Accepting \(<\text{getting your little brother to lie is not wrong}>\).

(γ*) Not accepting \(<\text{lying is not wrong}>\).

**Branch 2**

(a**) Accepting \(<\text{getting your little brother to lie is wrong}>\).

(β**) Accepting \(<\text{getting your little brother to lie is not wrong}>\).

(γ**) Not accepting \(<\text{lying is not wrong}>\).

(a*) is inconsistent with (γ*), (a**) is inconsistent with (β**). But this cannot be the correct interpretation of the conditional, because it falls pray to a decisive reductio, as Hale shows. Any interpretation must respect the tautology \(p \supset p\). But
on the interpretation we are currently considering, anyone uttering ‘$p \supset p$’ commits himself to either accepting that $\neg p$ or to accepting that $p$. But in cases where our evidence is inconclusive as to the truth of $p$, it seems that we should not commit ourselves to either accepting that $\neg p$ or to accepting $p$. We cannot abandon $p \supset p$, so we cannot accept the interpretation under consideration.

That there is no inconsistency between not accepting <lying is wrong> and not accepting <lying is not wrong> is a point that finds echoes in another aspect of the embedding problem, which is the noncognitivist’s problem with negation. Solutions to the embedding problem have focussed on given noncognitivist accounts of conditionals and disjunctions. But the problem is deeper than that. Any explanation of the validity of a modus ponens argument must explain why it is inconsistent to accept $p$, $p \supset q$ and $\neg q$. So noncognitivists owe us an account of negation. Unfortunately, this is none too easy to do, as Nicholas Unwin has demonstrated (Unwin 1999) (although here I draw on Mark Schroeder’s account of the problem (Schroeder 2008)).

A cognitivist can explain the inconsistency in asserting both “the King is dead” and its negation “the King is not dead” by pointing to the inconsistency in the content of these two assertions. For “the King is dead” is true if and only if the King is dead, while “the King is not dead” is true if and only if the King is not dead — and the King cannot be both dead and not dead. But expressivists do not want to explain the inconsistency of a pair of moral sentences like ‘murdering is wrong’ and ‘murdering is not wrong’ in terms of an inconsistency in the propositions that they express, for expressivists deny that they express propositions. Instead, expressivists think that in uttering ‘murdering is wrong’ I express a mental state, such as disapproval of murdering. So, is it possible to account for the inconsistency
between ‘murdering is wrong’ and ‘murdering is not wrong’ by pointing to an inconsistency between the mental states that utterances of these sentences would express?

Well, it seems that some types of mental state can have inconsistent tokens. Someone who believes both that the King is dead and that the King is not dead seems to have inconsistent beliefs. And this inconsistency seems different from the type of inconsistency involved in both believing that the King is dead and believing that you don’t believe that the King is dead. This latter type of inconsistency seems like the type of inconsistency involved in Moore’s paradoxes, such as “It is raining outside, but I don’t believe that it is raining outside.” As both these claims might be true, the type of inconsistency exemplified by someone making this utterance is not logical inconsistency. Incoherencies between second- and first-order beliefs do not, then, necessarily generate logical inconsistencies.\(^\text{62}\) But the belief that the King is dead and the belief that the King is not dead are both first-order beliefs. We might try supplanting the cognitivist account of the inconsistency involved in asserting both “\(p\)” and “\(\neg p\)” (in terms of the inconsistency of their contents) with an account that explained the inconsistency of these assertions in terms of an inconsistency in the mental states that they express. If this explanation can be made to work for beliefs, then perhaps we can just apply it to attitudes of disapproval. If “Murdering is wrong” expresses disapproval of murdering, then perhaps “Murdering is not wrong” expresses disapproval of something that is inconsistent with murdering.

Not all mental states share this property with beliefs. For example, it is not inconsistent to both wonder whether the King is dead and to wonder whether he is not dead. So we cannot simply assume that “Murdering is wrong” and “Murdering

\(^{62}\text{See van Roojen (1996), p. 332-333.}\)
is not wrong” expresses inconsistent attitudes. However, there may be reason for optimism for the noncognitivist in that intentions seem to have this property. Intending to get married is inconsistent with intending not to get married, for instance. And intentions might be thought of as desire-like states, which are the type of states that noncognitivists think are expressed by moral utterances. However, it turns out that “Murdering is wrong” and “Murdering is not wrong” cannot be inconsistent because they express the same type of attitude to different contents. For suppose ‘Murdering is wrong’ expresses disapproval of murder – how should it be negated? With beliefs, there seem to be three ways of negating the claim that “I believe that murder is wrong”:

(1) I do not believe that murdering is wrong.
(2) I believe that not murdering is wrong.
(3) I believe that murdering is not wrong.

But when it comes to the attitude of disapproval, it seems that there are only two ways to negate the claim that “I disapprove of murdering”:

(1’) I do not disapprove of murdering.
(2’) I disapprove of not murdering.

(1’) and (2’) appear to correspond to (1) and (2) respectively. But what we were after was an explanation of why ‘murdering is wrong’ is inconsistent with ‘murdering is not wrong’, i.e. something corresponding to (3), and that is what we do not have.
Schroeder (2008) provides a neat diagnosis of the problem. If my beliefs are inconsistent, it is because I have the same attitude (belief) to two different contents ($p$ and $\neg p$). But it seems that we cannot mirror this explanation with attitudes of disapproval. If an utterance of “Murdering is wrong” expresses an attitude of disapproval towards the content *murdering*, then to make the current proposal work, “Murdering is not wrong” should express an attitude of disapproval towards some content that is inconsistent with *murdering*. Call this content (whatever it might be), ‘$x$’. Now consider the pair of sentences “Not murdering is wrong” and “Not murdering is not wrong”. If “Not murdering is wrong” expresses an attitude of disapproval towards the content *not murdering*, then “Not murdering is not wrong” should express an attitude of disapproval towards some content that is inconsistent with *not murdering*. Call this content, ‘$y$’. The problem is that if $x$ is inconsistent with *murdering* and $y$ is inconsistent with *not murdering*, then $x$ must be inconsistent with $y$. But it is not inconsistent to state both that “Murdering is not wrong” and “Not Murdering is not wrong” – some actions are permissible without being obligatory.

What an expressivist can try to say is that, in uttering a sentence like “Murdering is not wrong”, a speaker expresses an attitude of *tolerance* towards acts of murder. But, as Schroeder (2008) points out, the above considerations seem to confirm that tolerance cannot be defined from disapproval and negation, they must be different types of attitude. Whereas we had been trying to say that “Murdering is wrong” and “Murdering is not wrong” expressed the same type of attitude to different contents, we are now forced to say that the express different types of attitude to the same content (i.e. *murdering*). But if that is the case, then
what remains to be explained is why it should be inconsistent to have these two
different types of attitude to the same content.
References


