http://researchspace.auckland.ac.nz

University of Auckland Research Repository, ResearchSpace

Copyright Statement

The digital copy of this thesis is protected by the Copyright Act 1994 (New Zealand).

This thesis may be consulted by you, provided you comply with the provisions of the Act and the following conditions of use:

- Any use you make of these documents or images must be for research or private study purposes only, and you may not make them available to any other person.
- Authors control the copyright of their thesis. You will recognise the author's right to be identified as the author of this thesis, and due acknowledgement will be made to the author where appropriate.
- You will obtain the author's permission before publishing any material from their thesis.

General copyright and disclaimer

In addition to the above conditions, authors give their consent for the digital copy of their work to be used subject to the conditions specified on the Library Thesis Consent Form and Deposit Licence.
Morality, Reason, and Motivation

Jeremy Richard Johnson

A thesis submitted in fulfilment of the requirements for the degree of Doctorate in Philosophy in Philosophy, the University of Auckland, 2015.
Abstract:

This thesis concerns debates in philosophy that centre on the relations between vertices of a conceptual triangle. These vertices, to simplify, are (A) morality, (B) reason, and (C) motivation. There are many philosophical debates that concern a relationship between a concept taken from one vertex of the triangle and a concept taken from another, that is, debates about the relationship between (A) and (C), (B) and (C), and (A) and (B). For instance:

Can someone believe that some action is morally required without being motivated to do it?

Philosophers have generally failed to reach any kind of consensus on questions like these. This thesis begins by explaining why this is, and then explains and demonstrates my preferred, Wittgensteinian approach to such questions. This approach involves two guiding thoughts. First, we must be careful to distinguish between radically different types of claim: grammatical, substantive, and empirical. Second, we must be careful to recognize that the meaning of a piece of discourse can be affected by the whole in which it sits, so that it may be impossible to say whether a statement is true or false when it has been taken out of any real context and examined as an abstract philosophical thesis.

As a result, in this thesis I do not on the whole straightforwardly agree or disagree with those who take up a philosophical position on the relations between morality, reason, and motivation. Instead, I show how their philosophical positions are true in a sense, but false in a sense. This is what I say, for example, about motivational internalism. Still, some philosophical claims are misleading or straightforwardly false, because they get
the grammar of our language wrong, or true (and generally quite trivial and unobjectionable), because they get the grammar right. Finally, some are substantive. By showing which claims are true only *in a sense*, which are misleading, which are straightforwardly true or false, and which are substantive, I hope to lead philosophers involved in unfruitful disputes towards a more nuanced understanding of what can meaningfully be said about the relations between morality, reason, and motivation.
For Tomomi
Preface

When I began working on this thesis, my intention was to apply a Wittgensteinian sensibility to the question “why be moral?” However, as I began to study this question, other questions about motivation, reason, and morality were naturally encountered. I noticed that these questions were receiving much more attention in the literature than the question “why be moral?” For this and other reasons, I came to feel that it would be better to focus the thesis on a cluster of questions, all of which could be approached in a Wittgensteinian way, and that doing so would demonstrate the power and fruitfulness of Wittgenstein’s approach. This thesis is the result.

I owe my deepest gratitude to my supervisor, Professor Rosalind Hursthouse. Without her encouragement, enthusiasm, advice, and support, I suspect this work would have never been completed, and definitely would have been much weaker if it had. As always, whatever faults remain are mine.

I am deeply grateful to Dr Glen Pettigrove for reading a draft of Chapter 5 and giving his suggestions.

I also owe a great debt of gratitude to Dr Paul Johnston for his book *Wittgenstein and Moral Philosophy*. I discovered Dr Johnston’s work midway through my research, and it was his book more than any other that helped me find what I wanted to say, and how to say it, in this thesis.
# Table of Contents

Abstract....................................................................................................................... ii
Preface ........................................................................................................................... v
Table of Contents ......................................................................................................... vi

1 Introduction............................................................................................................... 1
   I Morality, Reason, and Motivation ................................................................. 1
   II Grammatical, Empirical, and Substantive Assertions......................... 3
   III The Life Cycle of a Philosophical Debate ........................................ 11
   IV Obstacles and Impasses ............................................................................. 14
   V Preview of Parts and Chapters ................................................................. 20

PART A: Morality and Motivation................................................................. 28

2 Motivational Internalism and Aliefs......................................................... 29
   I Introduction ................................................................................................... 29
   II The Motivational Internalism Debate ................................................... 30
   III Aliefs Defined ............................................................................................ 33
   IV Kriegel’s Application of the Alief-Belief Distinction ..................... 35
   V The Minor Concern .................................................................................... 37
   VI The Major Concern .................................................................................... 43
   VII Moral Cognitivism and Besires ............................................................ 48
   VIII Conclusion ............................................................................................... 53

3 Motivational Internalism and Authenticity........................................ 54
   I Introduction ................................................................................................... 54
   II Authenticity .................................................................................................. 58
   III Externalism .................................................................................................. 67
   IV Vacuity .......................................................................................................... 71
   V Conditional Internalisms ............................................................................. 74
VI  Moral Cognitivism ................................................................. 79
VII Conclusion ........................................................................ 80

4 Three Concerns for Authentic Internalism ...................... 82
I  Introduction ........................................................................... 82
II  First Person Authority, Inauthenticity, and Grounds for Suspicion .... 85
III Inauthenticity versus Akrasia .................................................... 93
IV  Authenticity and Idealism ...................................................... 96
V  Conclusion ........................................................................ 104

PART B: Reason and Motivation ....................................... 107

5 The Nature of Explanatory Reasons ....................... 108
I  Introduction ........................................................................... 108
II  A Variety of Views ................................................................. 109
III The Personal/Impersonal Dilemma ........................................... 112
IV  Permissiveness ................................................................ 115
V  Ellipticality Defended ............................................................. 124
VI  Constitutionalism Undermined ............................................. 132
VII Resolving the Dilemma ..................................................... 138
VIII Conclusion ..................................................................... 145

6 In What Sense Are Reasons Causes? ..................... 146
I  Introduction ........................................................................... 146
II  “Cause” in Ordinary Language ........................................... 151
III The Epistemology of Causation ........................................... 154
IV  The Ontology of Causation .................................................. 159
V  The Deontology of Causation ............................................. 168
VI Conclusion ..................................................................... 173

7 A Defence of McDowell-Style External Reasons ........ 174
I  Introduction ........................................................................... 174
II  Williams’ Internalist Position ............................................. 176
| III | McDowell’s Response ................................................................. 178 |
| IV | The Possible Explanation Objection ........................................ 181 |
| V  | The Impersonality Objection ..................................................... 190 |
| VI | The Objectivity Objection ......................................................... 195 |
| VII | Conclusion ................................................................................. 197 |

PART C: Morality and Reason .............................................. 199

8 Morality, Irrationality, and Unreasonableness .................. 200

I Introduction .......................................................................... 200

II Moral Reasons as Grammatically Basic ............................ 200

III External Moral Rationalism ................................................. 208

IV Internal Moral Rationalism .................................................... 217

V Conclusion: The Middle Way .................................................. 224

9 A Defence against Moral Error Theory ............................. 226

I Introduction .......................................................................... 226

II The Queerness Arguments .................................................... 229

III The Nesting Argument ........................................................ 241

IV The Evolutionary Argument ................................................ 249

V Conclusion ............................................................................. 258

10 Conclusion ............................................................................. 259

Bibliography ............................................................................ 261
1 Introduction

In this chapter, Section I introduces the topic of the thesis. Sections II, III and IV explain the approach to philosophy that guides the thesis. Section V gives a preview of the content of the remaining chapters.

I Morality, Reason, and Motivation

This thesis concerns debates in philosophy that centre on the relations between vertices of a conceptual triangle. These vertices, to simplify, are (A) morality, (B) reason, and (C) motivation. Actually, at each vertex is a cluster of concepts. At vertex (A) we find concepts such as the moral, immoral and amoral, virtue and vice, the ethical and unethical, moral belief or judgment, and moral knowledge. At vertex (B) we find reasons for action, reason (as a faculty), rationality and irrationality, and the reasonable and unreasonable. At vertex (C) are motivation, motivated-ness, disinclination, indifference and desire. There are many philosophical debates that concern a relationship between a concept taken from one vertex of the triangle and a concept taken from another, that is, debates about the relationship between (A) and (C), (B) and (C), and (A) and (B). For instance,

(1) Can someone believe that some action is morally required without being at all motivated to do it? (the relation of an (A) concept and a (C) concept)

(2) Can someone have a reason for action independent of her actual desires? (the relation of a (B) concept and a (C) concept)
(3) Does everyone have reason to act morally? (the relation of an (A) concept and a (B) concept)

(4) Is it always irrational to act immorally? (the relation of an (A) concept and a (B) concept)

These four questions are the main focus of this thesis. However, connected with these sorts of questions are other questions whose answers might help to shape, determine or justify the answers to (1)-(4). For instance:

(5) What are reasons for action?

(6) Are reasons desire-belief pairs?

(7) Are normative reasons for action and motivating reasons for action the same sort of thing?

(8) Are reasons for action the causes of action?

(9) Can moral beliefs or truths be the causes of action?

(10) Are there any moral requirements or moral truths at all, or is the notion of the existence of moral requirements or moral truths fundamentally mistaken or confused?

If one answers question (1) with “no”, one accepts something known as motivational internalism. If one answers question (2) with “no”, one accepts reasons internalism, that all reasons for action are internal reasons. If one answers questions (4) (and perhaps (3)) with “yes”, one accepts moral rationalism, of which there are two types that I will distinguish – internal rationalism and external rationalism (see Chapter 9). I mention this now so that the reader is forewarned. Each part of the thesis deals, amongst other things, with a different internal/external debate: whether there is an internal or external
relation between moral belief and motivation to act accordingly (Part A, Chapters 2-4),
between having a reason to act and having desire/motivation that would be served by so
acting (Part B, Chapters 5-7), and between something’s being immoral and it’s being
irrational (Part C, Chapters 8-9). I hope it will always be clear from context which type
of internalism is being discussed, but I must admit it is a little unfortunate we
philosophers have such a fondness for the internal/external metaphor.

As is well known, despite many years of effort, philosophers have generally failed to
reach any kind of consensus on the above questions. In this introduction I would like to
explore why that might be, to explain how I nonetheless hope in this thesis to make
progress in dealing with these questions, and to make clear what kind of progress I
believe is (and is not) there to be made. The subsequent chapters will serve as examples
of the approach I favour.

II Grammatical, Empirical, and Substantive Assertions

Throughout the chapters, my concern is with concepts (e.g. moral, reason, motivation)
and conceptual relations, but also with types of assertion. There are at least three types
of assertions in which the concepts I am interested in can occur – grammatical,
substantive, and empirical – and these need to be distinguished, since a tendency to
confuse one type of assertion with another is one reason why we fail to reach agreement
on questions (1) to (10). I should emphasise that even though I will try to say something
about what distinguishes the three sorts of claims, my belief is that the best way to get a
feel for the differences between them is through examples. It is difficult to explain or
give an account of the differences without committing oneself to claims that at least
some readers might reasonably object to. However, my claim that the three types of
assertion are very different is meant to be uncontroversial – in discussing them I aim to
do no more than highlight differences that should be apparent to all, not to put forward a
substantive philosophical or metaphysical position about, say, the limits of science, or
the nature of language. Readers who think I go too far in my commentary should return
to the examples I am about to give and confirm that the three types of claim can be
distinguished. I should also add that I am willing to accept that there are ambiguous
cases – the same sentence may for instance be used to make a grammatical claim on one occasion and an empirical claim on another – and I admit that there may be borderline cases – a sentence may for instance be somewhat grammatical and somewhat substantive.

Here are what I take to be clear examples each of the three types of claim. To bring their differences into relief, I have given them, broadly speaking, shared subject matters: morning, life & death, and selling a car.

**Empirical**

**MORNING**
The sun rose at 5.47 this morning.

**LIFE & DEATH**
Camels are specially adapted to survive changes in body temperature that would kill most other animals.

**SELLING A CAR**
Given the current market, if you sold your car you could expect to get around $16,000.

**Grammatical:**

**MORNING**
The morning is earlier in the day than the afternoon.

**LIFE & DEATH**
One can’t be alive and dead at the same time.

**SELLING A CAR**
If you ought to sell your car, then you should sell your car.

**Substantive:**

**MORNING**
The early morning sky in winter is especially beautiful.

**LIFE & DEATH**
We are all lucky to be alive. Life is a miracle.
SELLING A CAR
You ought to sell your car, and use public transport instead, for the sake of the natural environment and the exploited peoples of oil producing nations.

Empirical claims will be familiar to the reader, but one should note that I am willing to call “empirical” not only those claims which might be seen to belong to the existing sciences, but also everyday non-scientific claims that could be checked by science-like observation or testing, at least in principle. Empirical claims are true or false in virtue of what is and is not the case within the world.

A grammatical claim is one that is true, if it is true, solely in virtue of – and here we may explain ourselves in various ways – the meaning of the concepts it contains, or the rules of the relevant language game, or because it is a truism of the language (rather than a truism about, say, the colour of the sky), or because it is an analytic truth, or a necessary truth, or because it can be known to be true a priori, or because it is a tautology. I readily admit that these different characterizations are not identical. For instance, there may well be claims that are necessarily true but that are not known a priori, as Saul Kripke has argued.¹ And perhaps mathematical claims are necessarily true, but not tautologies. But again, I mean to assert nothing controversial here. To stave off unnecessary argument, let me stipulate that the borderline between what counts as a grammatical claim and what does not is smudged. At its narrowest limit, it bounds only the intersection of what each characterization allows. At its widest, it bounds their union.

It may be a bit misleading to speak in general of grammatical claims and assertions (though I will continue to do so), as if they were something that might ordinarily be asserted, doubted, debated, justified, undermined, proven, and disproved. It is true that a particularly lengthy or involved grammatical claim might not be obviously true at a glance, and might thus require justification (as with some logical and mathematical theorems), but in general grammatical “claims” are those claims that we all treat as obviously true and that thereby make possible agreement and disagreement on matters of fact. Unless we all agreed that “the morning is earlier in the day than the afternoon,”

¹ Saul Kripke, Naming and Necessity (Oxford: Blackwell, 1980).
and on other related truisms, we could not meaningfully agree or disagree about the empirical claim that “the sun rose at 5.47 this morning.” (If this were possible, we would be speaking a different language and meaning something different by these words.) Unless we all agreed that “one can’t be alive and dead at the same time,” and on related truisms, we could not agree or disagree about the substantive claim that “we are all lucky to be alive.” So, unless difficult mathematics or the like is our subject, rather than talk about grammatical claims and assertions, it might be better to talk about grammatical truths and grammatical errors. So, for instance, “2 + 7 = 9” is a grammatical truth. “2 + 7 = 12” is a grammatical error. Because it is a grammatical error, we can also say that what “2 + 7 = 12” states makes no sense, or has no sense.\(^2\) The following sentences are likewise false because what they state makes no sense:

\[
\begin{align*}
\text{The afternoon is earlier in the day than the morning.} \\
\text{A person can be alive and dead at the same time.} \\
\text{If I ought to sell my car, then it is not necessarily the case that I should sell my car.}
\end{align*}
\]

Substantive claims are not grammatical claims. To quote from Paul Johnston’s *Wittgenstein and Moral Philosophy*:

>[N]o substantive claim can be demonstrated (or supported) by purely conceptual argument, and … correspondingly making a substantive claim involves taking up one particular position in the face of other equally coherent possibilities.\(^3\)

Nor are they empirical:

>[T]hat moral judgements as such cannot be supported by evidence (that it makes no sense to talk of evidence in this context) underlines the fact that they belong to a different category from empirical judgements and hence have a completely different grammar.\(^4\)

---

\(^2\) Or if one prefers, one could say it states nothing.


\(^4\) Ibid., 201.
As I see it, if one endorses a substantive claim, one takes a stand, directly or indirectly, on some matter of importance. One does not merely state a truism of the language, or describe the state of the world in the detached manner of a scientist. One expresses a perspective on one’s subject matter that cannot be shown to be right simply by investigating the meaning of concepts or gathering empirical data. Claims about what is morally good, what is beautiful, and what is important are substantive claims, as are claims about what one has reason to do. In fact, we may be able to group substantive claims together by saying that they are all claims about what we have reason to do, because if something is important, we have reason to act towards it in some way which respects that importance. Aesthetic substantive claims concern what we have reason to listen to, look at, eat, read, and enjoy. Ethical substantive claims concern what we have reason to do (or avoid doing) in our relations with each other and the natural world.

As indicated earlier, a sentence may be ambiguous in that it is possible to use the sentence to express two or more of the three types of claim – grammatical, empirical, and substantive. Here is an example:

*If an act is immoral, one ought not to do it.*

This could be a grammatical claim or a substantive claim. The ambiguity arises because there are two ways to read the word “ought”. Here is how the speaker of the sentence might explain himself in each case:

*All I mean is, if an act is immoral, one ought not to do it...you know, morally speaking. That’s just what “immoral” means.* (grammatical)

Alternatively, he might say:

*What I mean is, if something is immoral, one has a reason to avoid doing it that trumps other sorts of reasons one has for doing it, such as that it would be pleasant, or financially beneficial.* (substantive)
I also said that there might be borderline cases – cases which we do not want to call straightforwardly grammatical as opposed to empirical, or straightforwardly empirical as opposed to substantive, etc. Here is a familiar sentence that seems to combine features of the grammatical with features of the empirical:

\[\text{Water is } H_2O.\]

Though someone might discover the truth of this by empirical investigation, could she discover its falsity? If she took a sample of liquid she would call “water” and discovered its molecules were not $H_2O$ molecules, we would say she had been mistaken to think that the sample was water, not that she had discovered that some water was not $H_2O$. This might lead us to think that it is simply a matter of definition that water is $H_2O$. But that is not quite right either, since we only discovered that water was $H_2O$ quite recently in history, and we had a functioning concept of water before the discovery. Perhaps we should say the meaning of “water” has changed, such that “water is $H_2O$” was once an empirical proposition, and is now a definition – but only perhaps. I will just say it is a borderline case.

Before moving on, I should respond to three objections to the tripartite division of assertions. One objection is that grammatical claims are really only a variety of empirical claim. A second is that no line can be drawn between the grammatical and the empirical. A third is that if (so-called) substantive claims are not simply a variety of empirical claim or grammatical claim, they are not really claims at all, but something else.

It might seem that grammatical claims are actually empirical. Is it not the job of the lexicographer to find out how words are used and to present her empirical findings in dictionaries? It is indeed the lexicographers job to do this, but that does not make grammatical claims empirical. A sentence expresses a grammatical claim if the speakers of the language treat it as expressing a grammatical claim, and only so long as they do. If a few of them suddenly start to wonder whether or not the grammatical claim is true, they are not really wondering about whether the grammatical claim is true but about
how the words in the claim are used. As such, they are no longer considering a grammatical claim, but an empirical one about how words are used by the speakers of a particular language, their own language.

A different but related objection is found in W. V. Quine. In “Two Dogmas of Empiricism”, the first of the two dogmas Quine questions is

a belief in some fundamental cleavage between truths which are analytic, or grounded in meanings independently of matters of fact, and truths which are synthetic, or grounded in fact.₅

Is Quine right to question this belief? That depends on what he means by a “fundamental cleavage”. If he only wants us to accept that there are intermediate or borderline cases between analytic and synthetic truths, I can agree – I have admitted as much already. However, if he wants us to accept that there really is no difference between what I would call paradigm examples of grammatical truths and paradigm examples of empirical truths, I must disagree. The reader herself can plainly see there is a difference.

The most that Quine’s famous discussion of the first dogma shows is that we may not be able to explain analyticity without using concepts that are synonymous with analyticity – that we are stuck within an intensional circle, and cannot define analyticity extensionally. Again, I am willing to accept this, but I do not see it as a problem.₆ I can even embrace Quine’s holism, according to which there is no part of the web of belief that is completely immune from revision, but I do this by saying that when a formerly held grammatical truth becomes a target for revision, it ceases to function as a grammatical truth, and in so doing it changes its sense. This is a semantically dynamic holism.

₆ See Hans-Johann Glock, Quine and Davidson on Language, Thought and Reality (Cambridge: Cambridge University Press, 2003), 75. One might add the same is true of other concepts, such as “moral” and “rational”, where again it poses no problem.
An entirely different objection is that there is no space for any truth beyond the grammatical and the empirical, and thus the substantive is either one of these or not really a matter of truth at all. The most famous proponent of this view was A. J. Ayer. In explaining his position in *Language, Truth and Logic*, Ayer asserted that “a statement is held to be literally meaningful if and only if it is either analytic or empirically verifiable.”

Call Ayer’s own assertion P. Let us try, in good faith, to understand what Ayer is saying in P. If he is being consistent, then he must acknowledge that P expresses either (a) an empirically verifiable proposition, (b) an analytic proposition, or (c) no (literally meaningful) proposition at all. Let us consider these in turn.

If P is empirical, one wonders how it could be verified. “Meaningful”, and “meaningless”, if they are to be empirical properties, must be interpreted as something like “having a use within language L” and “having no use within language L”.

“Literally meaningful” would mean, roughly, “having a use within language L and treated as either true or false within language L”. To determine whether P is true, we take a linguistic community L, such as our own, or a range of linguistic communities, and then examine whether the members have a use for statements that they treat as true or false but not as empirical or analytic. It seems obvious that our own community does this, and thus that P is, under an empirical interpretation, false for our community at least.

If P is analytic (or as I would say, grammatical), then to assert ~P is to assert something false and senseless, as “2 +3 = 8” is false and senseless. If this is the nature of P, Ayer must think that someone who believes ~P has fallen into confusion about her own concepts or missed something entailed by the nature of these concepts. Since grammatical claims are, in the main, immediately and obviously true or immediately obviously senseless, this would mean P must be one of those truths whose obviousness is hidden in the same way as some truths of mathematics and logic. In that case, Ayer owes us a proof of P. However, no such proof has been given. Until such a proof is

---

given, and I do not think it can be – I do not think believers in substantive truth are confused – we are entitled to go on thinking that P is not a grammatical truth.

Alternatively, Ayer might claim that P does not express a literally meaningful proposition, but it something else – an order, an expression of feeling, or the like – disguised in the form of a proposition. He does after all say that a statement “is held” to be literally meaningful if and only if it is either analytic or empirically verifiable. Who is the agent of this passive? If not everyday speakers of English, it must be Ayer himself and those of a like mind. Thus perhaps what Ayer is really doing is recommending a particular way of talking – a way of talking in which we restrict ourselves to only treating empirical and analytic claims as true or false - that he finds congenial. If so, he is suggesting that we modify our linguistic practices. There is nothing wrong in principle with recommending a change in the way we speak – to recommend the dropping of a word or a concept or way of talking, (say, let’s no longer talk in terms of “God”, or let’s cease to use the concept “feminine”), but we would want to know what reason there was for making the change. Here Ayer has a problem. He cannot say there is good reason to change to this new way of talking, because this would be to make a substantive claim, one that is not empirically verifiable nor analytically true, because it concerns what is of importance. Science does not deal in importance, and neither does grammar. Thus the most Ayer can say is that he wants us to change how we talk (an empirical claim), but not that we ought to change how we talk (a substantive claim).

To conclude, I see no problem with accepting the tripartite division between the empirical, the grammatical, and the substantive.

III The Life Cycle of a Philosophical Debate

I now wish to say a little about the conception of philosophy and philosophical progress that I operate with throughout this thesis. To do this, I will imagine a naturalist’s
encyclopaedia with an entry on the life cycle of the species *controversia philosophicus* (the philosophical debate). Here is what it might say.

The life cycle of *controversia philosophicus* consists of four stages: infancy, adolescence, maturity, and decline. A philosophical debate is born when philosophers are first inspired to ask, say, “Are tings tongy?” – e.g. “Are moral judgments cognitive?” “Are reasons to act causes?” “Are reasons to act based in an agent’s desires?” – and to disagree about the answer. Some believe “all tings are tongy”, or even “all tings are necessarily tongy”. Others believe, or suspect that, “some tings are not tongy”, or “it is at least possible for a ting not to be tongy”. In infancy, these two positions are starkly opposed, and are supported by straightforward arguments, such this:

\[
A \text{ ting is necessarily tonky.} \\
\text{Whatever is tonky is necessarily tongy.} \\
\text{Therefore, all tings are necessarily tongy.}
\]

If the debate is to have any longevity, then the opposition must dispute the argument by disagreeing with one or more of the premises (as well as the conclusion), possibly by pointing to an example of what they take to be an un-tonky or un-tongy ting.

*Controversia philosophicus* enters adolescence as each side proceeds to offer ever more sophisticated and involved arguments for their position, arguments that are sophisticated and involved because they are responses to sophisticated and involved arguments from the opposition. The nature of the thesis being argued for (or against) also becomes more sophisticated and involved. For instance, a participant may attempt to define or refine what “tongy” means, to distinguish tongy-ness from related but different properties, and to undermine alternative understandings of what is involved in being tongy. The final stage of adolescence is reached when, having come to something of a stalemate in the debate, the participants begin to wonder whether the sophisticated versions on both sides are truly opposed or not, and whether philosophers in the different camps mean the same thing by the words they use (“ting”, “tongy”, “tonky” etc.) or are merely arguing past each other. Meaning and interpretation become the focus.

Some philosophical debates remain in permanent late adolescence – a state of permanent impasse and frustration. However, when conditions are right, a debate reaches maturity, and transforms (like a caterpillar emerging from its chrysalis as a butterfly) into something other than a philosophical debate. There
are two types of maturity. Firstly, the debate over whether tings are tongy may disappear and be replaced with a shared understanding of the meanings of a range of related terms and phrases. Substantive and empirical questions are disentangled from grammatical questions, and participants achieve a clear oversight (or Übersicht\(^8\)) of the grammatical domain – the range of language games that are connected to the original dispute, and their form and rules. In the process, the illusion of a substantive question disappears.

Alternatively, though participants come to understand clearly what their disagreements with the opposition are, what each side means by what they say, and where there are differences in preferred vernacular, substantive disagreement remains. Philosophical debate and reflection have achieved all they can and cannot achieve more. If agreement on substantive matters has not emerged as a result of the earlier stage of argument and counter-argument, or the later stage of clarifying meanings, it is not a philosopher’s job anymore, as a philosopher, to bring about agreement. The mature stage of *controversia philosophicus* is one of wisdom, and part of wisdom is knowing one’s limitations, and when to stop.

So, either the question “Are tings tongy?” transforms into functioning as a kind of signpost to a complex landscape of shared language games, and is as such a question which people are wise not to try to take sides on, or the question remains but in a purified substantive form, one that each person must choose how she answers. In the latter case, each person effectively chooses, or reaffirms, a form of life – for instance, a religious form of life, or a humanitarian form of life.\(^9\) Philosophy has nothing to say about which form of life we should choose.\(^10\)

Finally, although maturity can be permanent, sometimes the understanding, sophistication and wisdom achieved in the development of the debate are lost and forgotten, as attention shifts onto other areas of investigation, and as history runs its course. This is the stage of *controversia philosophicus*’s decline. The decline continues until a philosopher is once more inspired to ask the question,

---


\(^10\) At least as I understand it, as a grammatical investigation. There are other less restricted but legitimate conceptions of what philosophy is and what it can accomplish, ones which allow the philosopher to argue for substantive conclusions by invoking other substantive commitments, but they are not the conception of philosophy that I will operate with in this dissertation.
or something like the question, “Are tings tongy?”, whereupon the life-cycle beings again.

As should be obvious from this fictional encyclopaedia entry, my conception of the mature stage of debate is deeply indebted to Wittgenstein. When I explain that stage, I make use of Wittgensteinian concepts: Übersicht, grammar (in Wittgenstein’s sense), language game, and form of life, and a Wittgensteinain vision of the nature, function and limits of philosophy. I believe that philosophical debates cannot reach maturity if the participants do not share in something like this vision. Permanent adolescence is all we can hope for if, for instance, we conceive of philosophy along the model of the natural sciences, and see ourselves as constructing philosophical theories that are straightforwardly and unproblematically true or false and that can be confirmed or falsified by some philosophical equivalent to empirical investigation and testing. This conception will not impede early adolescent development of a debate, but to reach maturity we must start to self-consciously explore issues of meaning and interpretation in a way that scientists, who endeavour to express their claims in semantically unobtrusive ways (i.e. in ways that demand no process of careful interpretation), do not and ought not. The main purpose of this thesis is to take some philosophical debates that seem largely stuck in adolescence, and nurture them into maturity, or at least into the final stages of adolescence.

IV Obstacles and Impasses

When we debate, our primary goal is to reach consensus. Along the way to consensus we may encounter relatively intractable obstacles, or come to a complete impasse. In the sciences, when a consensus is not reached, this is often because there is a problem with conducting the relevant tests or gathering the necessary observational data to decide the issue. This is how it stands, for instance, with the question of whether there is life on

11 “In science, we take care that the statements we make should never depend upon the meaning of our terms…We are always conscious that our terms are a little vague…and we reach precision not by reducing the penumbra of vagueness, but rather by keeping well within it, by carefully phrasing our sentences in such a way that the possible shades of meaning of our terms do not matter.” Karl Popper, The Open Society and Its Enemies, Volume Two: Hegel and Marx (London and New York: Routledge Classics, 2003), 21-22.
other planets, and the question of how our moon formed. In mathematics, we may be unsure about whether a possible theorem is indeed a theorem. If we cannot agree, it is because we have not been clever enough to construct a proof or disproof of the theorem, or a proof that shows that the theorem cannot be proved. These kinds of obstacle – a lack of data, and a lack of cleverness in constructing demonstrations – are not I think especially relevant to the development of philosophical debates to maturity.

A different and more relevant kind of obstacle to consensus is that some of us who are engaged in philosophical debates may be more interested in being proven right, or seeming to win an argument, or defending a position against attack at any cost, than listening to what the opposing side has to say on the matter, genuinely questioning our own views, and getting at the truth. One can only hope that philosophers come at philosophy with the right attitude, in an authentic and open way. Beyond hope and self-analysis, there is little to say or do about this type of obstacle.

There are two other types of obstacle (or one type of impasse, and one type of obstacle) that I think it is a philosopher’s job to recognize and be mindful of. Difficulties of the first sort I call substantive impasses and difficulties of the second sort I call semantic obstacles. We have already met substantive impasses in the discussion of the mature stage of the debate life cycle. Let me just give an example from the coming chapters. The final chapter of this thesis explores moral error theories. A moral error theorist believes that there is something fundamentally wrong with moral discourse, in that this discourse is committed to the existence of entities (moral facts, desire-independent reasons for action, etc.) which do not exist. In Chapter 9 I defend ordinary moral discourse, and moral belief about moral facts, but the chapter does not consist of an argument which shows that moral facts genuinely exist. It is a discussion that leads to a point of impasse. After philosophical clarification, we are left either as believers in morality or non-believers. No rational argument exists which would show that moral error theorists are wrong to disbelieve in moral facts, but also no rational argument exists which would show that those who believe in moral facts are wrong. The most I show in Chapter 9 is that the error theorists are wrong for thinking such an argument exists. The error theorists who do so confuse grammatical questions (for instance, about
what it means to have a reason to act) with substantive questions (for instance, about what reasons we actually have to act). Philosophy reminds us about and clarifies the grammar of our concepts. If substantive disagreement dissolves in the process of getting clear about the grammar well and good, but there is no guarantee that clarifying the grammar of reasons talk and moral talk will have this effect.

As regards semantic obstacles, philosophical consensus cannot be reached if we do not respect the difficulty of interpreting each other, even when we share (broadly speaking) the same language. This is a difficulty which is related to the variety of acceptable ways of using words and phrases, in context.

Philosophers often operate with an unhelpful caricature of how language functions. In this caricature, individual words or phrases have fixed meanings. When these words and phrases are combined in grammatically correct ways to make larger, meaningful structures – longer phrases, sentences, paragraphs, and so on – the meaning of the whole is determined by the meaning of the parts and the grammar of the language. (Here I mean grammar in the everyday sense, not Wittgenstein’s sense.) This sort of view is evident in what is known as Frege’s Principle, or the Principle of Compositionality. The Stanford Encyclopaedia of Philosophy explains the principle:

> Anything that deserves to be called a language must contain meaningful expressions built up from other meaningful expressions. How are their complexity and meaning related? The traditional view is that the relationship is fairly tight: the meaning of a complex expression is fully determined by its structure and the meanings of its constituents—once we fix what the parts mean and how they are put together we have no more leeway regarding the meaning of the whole. This is the principle of compositionality, a fundamental presupposition of most contemporary work in semantics.¹²

---

The relationship between the parts and the whole is sometimes not as simple as this principle suggests, at least where everyday rather than technical language is concerned. Although the meaning of the whole obviously does depend on the meaning of the parts and the grammar, the meaning of the parts is also somewhat determined or shaped by the meaning of the whole – where the whole includes not only all pieces of language that sit around the part in question, but features of the discourse situation as well. So, in particular, we often understand what a speaker’s words mean by understanding the overall gist or point the speaker (or writer) is making in the context of a discussion, and we interpret parts, where there is some ambiguity, to align with this gist or point.\footnote{The interdependence of semantic parts and wholes means interpretation must follow a process known as the hermeneutic circle. See Hans-Georg Gadamer, \textit{Truth and Method}, trans. Joel Weinsheimer and Donald G. Marshall (London and New York: Continuum, 2004).}

It is therefore quite possible that when one speaker asserts, to return to our Encyclopedia entry, “all tings are tongy” it is true in the context of the larger whole of what she is saying, but taken out of this context and placed in another, “all tings are tongy” will be false (or ambiguous, or meaningless). This is because the difference in context affects the meaning of the sentence. When we recognize that this happens, we can say “all tings are tongy” is true, in a sense, but also false, in a sense. Of course, there are limits to how far the meaning of a part can be stretched and shaped by the whole in which it sits. We do not have complete leeway to use words however we like, since we must speak in such a way that we can be understood. Nonetheless, there can be a two-way semantic interdependence between parts and wholes, not just the one-way dependence suggested by the principle of compositionality.

To the degree that we hold onto a caricature of the workings of language, we will find it hard to reach consensus in philosophy, because we will attempt to interpret statements without due sensitivity to the effect of the whole on the meaning of the parts. Semantic obstacles can be overcome by exploring the various ways in which a proposition can be understood and used, and the various senses in which it can be true or false. To overcome these obstacles, we must often find other things to say about a proposition $P$ other than that it is simply true or false. Although sometimes we evaluate in this way:
P is straightforwardly true.

Or this way:

P is straightforwardly false.

We can also evaluate in these ways:

P is true if the speaker means P₁, but false if she means P₂.
P is true, but potentially misleading, in that it seems to imply Q, which is false.
Some think that saying P is misleading, or even false, but it can be acceptable to say P.
P is not the clearest way of expressing the point the speaker/writer wishes to make. It would be better if she said/wrote P'.
P, as it is used in this context, is ambiguous. It might mean P₁ or it might mean P₂.
P is a helpful way of talking about the matter at hand, but R is not.
One can say P, so long as one does not also say S, or one can say ~P, so long as one does not also say T.
One can say P, so long as one also admits U, or one can say ~P, so long as one also admits V.

It is worth having a name for the last two sorts of evaluation, where what is said or not said along with P will create the context in which P can be interpreted as true or false (or misleading or ambiguous). If a proposition can be accepted as true, so long as certain other things are not said (or said), but would be false (or meaningless or ambiguous) otherwise, I say the proposition is conditionally permissible.

Semantic obstacles arise from the fact that there are many things we can say which are conditionally permissible, and thus not straightforwardly true or false when taken in isolation from any larger piece of discourse in which they are used rather than
mentioned. To overcome semantic obstacles we must explore together the conditions of permissibility, ask for clarification from each other, try to speak in such a way that it is clear what we mean, and admit when necessary that there are other ways of talking that are conditionally permissible too.

To clarify what has been said and see its application to our topic, rather than continue to talk about “tings” and whatnot, let me give one example from the coming chapters. In Chapters 2 to 4, I examine the question of motivational internalism: is there a necessary connection between someone’s believing a moral claim, and being motivated by it to act accordingly? In the course of dealing with this question, I am forced to dwell on the issue of what beliefs are, which is to say, what “belief” means, or how “belief” is used. I identify two ways in which “belief” might be used. One is a comparatively liberal sense, where what is believed can be empirical, grammatical, or substantive, and can therefore accommodate practical belief about what one has reason to do. A contrasting use of “belief” is one that only admits the empirical and the grammatical, but classifies what we might want to call beliefs about practical matters as non-beliefs, as really an amalgam of the first sort of belief and something non-cognitive – a feeling, a desire.

In the light of what I have said about conditional permissibility, rather than deciding which way of talking is “correct”, we should ask whether these ways of talking can be equally acceptable, and if so when, and question whether and when one or the other tends to be misleading if not straightforwardly wrong, and whether and when one is clearer or more helpful. I argue in Chapters 2 to 4 that there is nothing impermissible about using “belief” in the liberal way, and this is indeed how it is used when we talk about the everyday beliefs of people from an everyday, engaged, perspective. Perhaps, when we are engaged in science, and especially if we want to construct a testable theory whereby we can predict, explain and control some aspect of human behaviour, it may be useful to use “belief” in the restricted sense. That is, if it is helpful to talk here about belief at all, rather than neurological states or stimulus and response patterns. However, if we are going to choose to talk in this restricted way, we should not pretend that this means that the more liberal sense of “belief” is inferior or confused, in virtue of its being less amenable to science. Certainly we should not do this while there is no pure
science of human behaviour requiring a restricted sense of “belief”. It will probably turn out that the notion of “belief” in both the restricted and liberal senses are no use to the science of human behaviour, but I see no problem with allowing scientists to attempt to make progress by dividing the (so-called) cognitive from the (so-called) affective. The rest of us can continue to talk in terms of practical belief as we always have.

V Preview of Parts and Chapters

A quick sketch of Part A

Part A contains three chapters that explore the issue of how moral belief and motivation, or moral belief and desire, are related. Chapter 2, “Motivational Internalism and Aliefs”, and Chapter 3, “Motivational Internalism and Authenticity”, examine the debate over moral motivational internalism. This is the debate over whether it is possible to believe that one morally ought to do something (or ought not to do something) and yet fail to be motivated to do (or not do) it. Chapter 2 outlines an interesting but ultimately unsuccessful attempt by Uriah Kriegel to resolve the debate using the distinction between belief and something called “alief”. In Chapter 3, I respond with my own solution (of sorts) to the debate, a solution which takes up and complicates the alief-belief distinction and uses it to develop a notion of authentic belief and to defend a conditional form of motivational internalism that I call authentic internalism. In Chapter 4, “Three Concerns for Authentic Internalism”, I deal with three concerns that might be raised against authentic internalism: that it conflicts with our having first-person privilege with respect to our own thoughts, that it leaves no space for the phenomenon of akrasia, and that the thesis cannot be made more definite without losing its necessary character or becoming in other ways unattractive.

Part A in more detail

14 I contrast pure science with interpretive or hermeneutic studies of human behaviour, such as belong to history, economics, and social psychology.
The motivational internalism question (*does moral belief require motivation?*) exists as a philosophical question because of failure to notice the different ways in which we use the noun “belief”, and the related verb “believe” and noun “believer”. From a scientistic point of view, the object of belief is only a true or false empirical claim, a logical claim, or a mathematical claim. Since these claims are meant to be understandable and testable (e.g. falsifiable, disprovable, deducible, demonstrable) independently of a conception of the good life, there is no requirement that a believer do (or feel) anything as a result of believing a claim. For the scienticist (the conscious or unconscious adherent of scientism), the same holds true of moral belief, because such belief is interpreted on the model of empirical, logical and mathematical belief. Therefore, scienticists are typically believers in moral motivational externalism.

As I discuss in Chapter 2, Uriah Kriegel gives a two-part answer to the motivational internalism question based on Tamar Gendler’s distinction between aliefs and beliefs.\(^{15}\) (Gendler distinguishes belief from alief, where aliefs are mental states\(^{16}\) “whose content is representational, affective, and behavioural” and which are “developmentally and conceptually antecedent to other cognitive attitudes that [we] may go on to develop”. Aliefs are typically “affect-laden and action-generating”.\(^{17}\) As concerns motivational internalism, Kriegel argues that there *is* a necessary connection between moral alief and motivation, but *not* between moral belief and motivation. His position is therefore fundamentally externalist.

In using the alief/belief distinction to answer the motivational internalism question, I believe Kriegel’s solution to be ultimately scientistic. At the very least, Kriegel does not consider the possibility that, in order for some types of belief to be genuine beliefs of

---


16 I think it fair to say that beliefs and standing desires are not, in ordinary talk as opposed to philosophical jargon, *states of mind or mental states*. In philosophy it can be handy to have a word or phrase with which to group together things such as beliefs, desires, intentions, feelings, hopes, and so on, and “mental state” can serve that purpose, so long as we are careful to remember that nothing philosophically significant about the nature of beliefs (etc.) is implied by this choice. My acceptance and (infrequent) use in this thesis of the “mental state” idiom should be understood in the light of these comments. See P. M. S. Hacker, “Davidson on the Ontology and Logical Form of Belief,” *Philosophy* 73, no. 283 (1998), 81-84.

their type, the agent who holds them might need to have motivational states and dispositions to feel and act in certain ways. In particular, Kriegel simply assumes that an agent can have genuine moral beliefs without any moral aliefs. This, I think, is to beg the question against the motivational internalist.

I should add that Chapter 2 also contains my thoughts on the moral cognitivism/non-cognitivism debate. Since moral cognitivism, the Humean theory of motivation, and motivational internalism form an inconsistent triad, the thesis of moral cognitivism is relevant to the question of motivational internalism.

Chapter 2 does little more than raise my objection against Kriegel. In Chapter 3, I offer both an alternative answer to the motivational internalism question and a more faithful characterization of the nature of belief that is moral belief. I call my position authentic internalism. Authentic internalism is a form of conditional internalism. That is, it states that there is a necessary connection between an agent’s having a moral belief and her being accordingly motivated, but on condition: the existence of the connection depends on the agent’s being, or the belief’s being, authentic as opposed to inauthentic. Authentic internalism is, if understood correctly, a grammatical truth. Chapter 3 explains what is involved in the idea of authenticity, and why it secures a necessary connection between belief and motivation. Authentic internalism is also contrasted with other forms of conditional internalism. While I do not argue that these other forms are straightforwardly wrong, I argue that authentic internalism is a less misleading way to talk about the connection.

Chapter 4 mops up three niggling concerns the reader of Chapter 3 may be left with. The first is that authentic internalism seems to conflict with our having first-person privilege with respect to our own thoughts. I argue that first-person privilege is limited, and there are cases where there are grounds for saying that someone does not authentically believe what she tells herself she believes and publically claims to believe. However, special grounds are indeed needed. Without such special grounds, if a person believes that she believes some proposition, then we should say she genuinely and authentically does believe it. I make clear what some of these grounds are. The second
concern is that authentic internalism leaves no space for the phenomenon of akrasia. I argue that moral akrasia and moral inauthenticity are distinct but related defective relations to one’s own practical beliefs, and that one can be authentic (or have an authentic moral belief) and yet behave akratically. The third concern can be seen to arise from something in the previous sentence. The condition in my favoured form of conditional internalism is one of authenticity, but is authenticity to be conceived of as a property of agents, or as a property of beliefs, or of both? Also, there are questions about how high the threshold for achieving authenticity is, and a worry that if the bar is set too low, we will lose the necessary connection between moral belief and motivation, but if it is set too high, we lose the sense that talking about authentic agents or beliefs is a useful way to talk about imperfect humans such as we are. The concerned reader wonders whether the thesis of authentic internalism can be made more definite on these points without losing its necessary character or becoming unattractive in other ways. I argue that there are different ways of understanding the thesis of authentic internalism, and that each has at times its advantages and disadvantages, but that we are not forced to choose between them. The solution is to remain somewhat flexible about what is being asserted, but we should be self-conscious and open about doing so.

A quick sketch of Part B

Each of the three chapters in part B deals with one of three different but related debates that concern the relation between motivation (or desire) and reason. Chapter 5, “The Nature of Explanatory Reasons”, explores the nature of explanatory reasons, motivating reasons, and normative reasons, and the question whether these various types of reasons are the same sort of thing, ontologically speaking, and in particular whether they are psychological states of desire and belief. Chapter 6, “In What Sense are Reasons Causes?”, looks at whether reasons for action are the causes of action, as might seem to be the case if reasons for action are, as Humeans believe, psychological states of desire and belief that can lead to action. Chapter 7, “A Defence of McDowell-style External Reasons”, examines the internal/external reason debate about whether it is possible to have a reason to act in some way without having a desire that would be served by acting in that way.
**Part B in more detail**

Chapter 5 explores the various ways in which the terms “reason” and “explanation” are used to explain, or give the reason why, an agent acted in some way. I argue that there are many different things which it is permissible to call a reason or an explanation, and that in deciding what counts as a reason or explanation we are guided by pragmatic concerns (by whether something functions to explain) rather than by ontological concerns (what type of entity a reason or explanation is). By taking this line, we avoid a dilemma, or trilemma, that arises if we think we can straightforwardly answer the question whether explanatory reasons are the same sort of entity as normative reasons. Let me explain the trilemma. If we say that reasons of these two types do not have the same constitution, then we encounter the problem that we can no longer talk about agents acting for good reasons. If we say that the two types do share a constitution, and think that constitution is one of something like desires, desire-belief pairs, intentions or goals, this seems to force us into accepting an unsatisfactory position regarding when normative reasons, and especially moral reasons, exist, and whether we always have reason to act morally. Alternatively, if we say that explanatory and normative reasons share a constitution, and think their constitution is one of something like facts, propositions, states of affairs, considerations, or the contents of beliefs, we run into difficulty in relation to how explanatory reasons can explain action. The solution to the trilemma is to stop talking as if there were a single sort of thing that reasons are – to remember the function of reasons, and the according variety of reason talk, and to resist trying to give an account of the relation between reasons and motivation that ignores or occludes this variety. In the course of making my argument, I also argue that the term “motivating reason” is currently an unhelpful one for making progress, as it is used quite differently and sometimes ambiguously by philosophers on both sides of the debate.

Chapter 6 looks at whether reasons for action are the causes of action, as many philosophers since the publication of Donald Davidson’s article “Actions, Reasons and
Causes” have believed. Instead of straightforwardly agreeing or disagreeing with Davidsonians, I consider in what sense reasons for action can be said to be the causes of action, and in what sense they cannot. Thus the claim “reasons are causes” is shown to be conditionally permissible. I distinguish between two conceptions of causation, soft causation and hard causation. I distinguish between soft and hard causation by their different epistemologies, ontologies, and deontologies, and I argue that reasons for action should not be construed as hard causes of action, but as soft causes.

Chapter 7 examines the internal/external reason debate about whether it is possible to have a reason to act in some way without having a desire that would be served by acting in that way. (This debate should not be confused with the motivational externalism/externalism debate discussed in Chapters 2–4). The issue was introduced by Bernard Williams in his article “Internal and External Reasons”, where Williams tried to show that external reasons do not exist. In Chapter 7 I disagree with Williams by taking up, developing and defending a response to Williams from John McDowell. Williams thinks he can show that external reasons do not exist because (1) reasons must be able to explain action, but (2) external reasons cannot explain action, since (3) if a reason is external for an agent, by definition she has no desires that would be served by acting on it, and thus cannot be motivated to act on it, and thus cannot act for that reason. Perhaps the key point made in Chapter 7 is that Williams’ argument only works if one understands the term “possible” in the claim “if someone has a reason to perform some act, it must be possible for her to perform that act for that reason” in a particular way. We are all inclined to accept this claim, but care is needed to know what one is endorsing. I argue that the claim can be true if one uses a different interpretation of “possible” than the one Williams would use, that one can accept the claim under the different interpretation while rejecting it under Williams’, and that the claim under this different interpretation is consistent with a belief in external reasons. Once this is apparent, it becomes clear that Williams has (and can have) no argument to show that external reasons do not exist. Instead, the claim that external reasons do not exist is a

---

substantive claim about what we have reason to do, and externalists and internalists must recognize that they this claim signifies a point of substantive impasse.

**A quick sketch of Part C**

The two chapters in part C concern the relation between moral belief on the one hand and rationality and reasons on the other. Chapter 8 argues that it is at least misleading, if not straightforwardly wrong, to label people who ignore moral reasons as irrational, and argues that a better term for them would be unreasonable. Chapter 9 defends moral belief against a number of attacks from moral error theorists, who think there is something fundamental at fault with all moral reasons talk.

**Part C in more detail**

Chapter 8 questions moral rationalism, the view that acting immorally is necessarily acting irrationally. The chapter distinguishes and then argues against two types of moral rationalism – external moral rationalism and internal moral rationalism. External moral rationalism sees practical rationality as lying underneath morality and justifying it from apart and below, whereas internal moral rationalism sees moral concerns as simply part of how rationality is to be characterized, so that practical rationality does not justify but is instead infused with morality. I use two external rationalist arguments, one from Alan Gewirth and other from Julia Markovits, to explain how external rationalist arguments fail, and look at a passage from Philippa Foot to highlight what is wrong with internal rationalism. I argue that moral reasons are grammatically basic, which is to say, it is part of the grammar of moral reasons that they do not rest on extra-moral reasons. This is so even though one can sometimes justify behaving in the same way as a moral person would on the basis of extra-moral reasons. I also argue “unreasonable” is a better word than “irrational” for describing a person who does not feel the force of genuinely moral considerations, unless one goes to the trouble to make clear that one is operating

---

with a moralised conception of rationality, as might be achieved by using Jürgen Habermas’ term “communicative rationality”.22

Chapter 9 considers the moral error theorists’ claim that moral discourse is committed to the existence of reasons of a sort that cannot be made sense of or do not exist. At the heart of the chapter is my counterclaim that error theorists confuse substantive and grammatical claims about reasons. Once these are distinguished, we can see that there is no argument from grammatical considerations to show either that there are genuine moral reasons, or that there are not moral reasons. Whether moral reasons exist or not is a substantive issue that philosophy cannot adjudicate. Thus moralists (people who participate in moral discourse and believe in moral reasons or facts) cannot be accused of holding views that do not make sense, as error theorists have claimed. In addition, Chapter 9 questions J. L. Mackie’s view that moral facts are problematically queer, and examines and refutes Richard Joyce’s evolutionary debunking argument, according to which, since our moral beliefs have evolved and would evolve whether or not they were true, all of our moral beliefs are unjustified if not false. The chapter’s conclusion is similar to Chapter 7’s – those involved in the error theory debate should see that we reach a point of substantive impasse on whether or not reasons which are in a certain sense independent of desire exist or not.

PART A: Morality and Motivation
2  Motivational Internalism and Ailiefs

I  Introduction

This chapter examines Tamar Gendler’s distinction between belief and alief, and Uriah Kriegel’s attempt to use this distinction to argue for a hybrid internalist-externalist position in the debate over the relationship between moral judgment and motivation. I say hybrid internalist-externalist position, but we’ll see that Kriegel’s view is actually more like a species of externalism. I will argue that the alief-belief distinction somewhat oversimplifies the complexities of psychology in relation to the issue of moral motivation, and fails to settle the debate in the externalist’s favour, contra Kriegel. The examination of Kriegel’s failed attempt sets the stage nicely for a related but more satisfying approach, one which hinges on the distinction between two ways of believing rather than the distinction between belief and alief. This positive project will be left for Chapter 3. For now, my primary aim is to undermine Kriegel’s argument.

The structure of this chapter is as follows. Section II introduces the motivational internalism/externalism debate. Sections III and IV introduce Gendler’s distinction between alief and belief, and Kriegel’s application of the distinction to the debate. Sections V and VI outline my concerns about Gendler’s distinction and Kriegel’s application, concerns which entail that Kriegel’s application fails to resolve the motivational internalism/externalism debate. In Section VII, I consider what this means for a closely related debate concerning moral cognitivism/non-cognitivism. Also, since my concerns over Kriegel’s approach stem from my belief in something like besires (inseparable belief-desire complexes), I focus the cognitivism discussion on the question of the nature and existence of besires. Section VIII offers a short conclusion.
II The Motivational Internalism Debate

Moral motivational internalism is the view that there is a necessary connection between someone’s making a sincere moral judgment and her being somewhat motivated to act accordingly. If one is a moral cognitivist, motivational internalism implies that it is not possible to believe that a moral claim is true without having some accompanying or resulting motivation to act accordingly – or, on a weaker reading, that it is not possible to be like this with regard to all of one’s moral beliefs all the time but is possible on special occasions, such as during a period when one is suffering from depression, or is possible with respect to some but not all of one’s moral beliefs.

If instead one is a moral non-cognitivist, such that moral judgments are understood not as beliefs but as prescriptions, expressions of feeling, or the like, motivational internalism will amount to something like the claim that it is impossible to genuinely prescribe an action or express a moral feeling (etc.) without some motivation to act accordingly – or, on the weaker reading, it is impossible to do this generally or always, but perhaps possible in special cases. For some versions of non-cognitivism, this sounds quite plausible.

However, the motivational internalism/externalism debate is typically of more interest (and concern) to moral cognitivists, because the combination of motivational internalism and the Humean theory of motivation is inconsistent with moral cognitivism. Together they form an inconsistent triad. The elements of the inconsistent triad can be explained in different ways. Here is one:

Moral judgments are inherently motivational.  
(moral motivational internalism)

Moral judgments purport to be about objective matters of fact.  
(moral cognitivism)

Mental states, including moral judgments, cannot both have objective purport and be inherently motivational.  
(the Humean theory of motivation)
Here is another:

One cannot make a moral judgment without being motivated accordingly.
(moral motivational internalism)

Moral judgments are beliefs.
(moral cognitivism)

Belief alone is insufficient for motivation.
(the Humean theory of motivation)

By contrast, according to moral motivational externalism it is possible for someone to make a moral judgment, (or assuming moral cognitivism, to believe a moral assertion), without being motivated to act accordingly. In the fight against motivational internalism, the main though not only weapon in the arsenal of the motivational externalist is the actual or at least conceivable existence of people we may call amoralists. In the Collins dictionary, amoralism is defined as “the doctrine or attitude that ignores, rejects, or deems moral values to be irrelevant.”23 Yet in the philosophical literature on motivational internalism, amoralists are not typically characterized by their adherence to a doctrine but by absence of motivation. For instance, Björklund et al. say amoralists are “persons who make moral judgements despite lacking suitable motivation.”24

Not every amoralist that falls under the Collins definition would, if she existed, count as a weapon in the fight against internalism. The conceivable existence of an amoralist who rejects the practice of making moral judgments is of no use to externalists – this type of amoralist does not believe moral claims or express any moral feelings or prescriptions, if she is consistent. However, an amoralist who sometimes makes or accepts moral judgments, but who always deems them irrelevant and so ignores them,

so as to be completely unmotivated by them, certainly would be a counterexample to internalism.

Whether an amoralist who is largely rather than completely or often rather than always unmotivated by her moral judgments is also a counterexample to internalism depends on the version of internalism being discussed. We can simplify things by taking internalism to be the thesis that being completely and always unmotivated by each of one’s moral judgments is impossible.

The question then is this: is such a pure amoralist possible, or is the very thought of such a person some form of confusion. This question requires us to evaluate supposed cases of amoralists: is a supposed amoralist person really completely and always unmotivated to act morally, and if so, does she genuinely make the moral judgment (or believe a moral claim) in question, or not?

Morality is not the only subject matter around which a motivational internalism/externalism debate could be constructed, though it is the one most often discussed. Similar issues can arise in relation to non-moral judgments. To give some specific examples, (1) we may wonder whether it is possible for a person to judge (or believe) that salmon sushi is scrumptiously delicious but to be completely unmotivated to eat it. (2) We may wonder whether someone can genuinely think it imprudent or foolish to walk down the train tracks but feel no qualms about doing so. (3) We may wonder whether someone may use a term like “nigger” or “homo” to refer to another person without having at least some disposition to treat that person with disdain. Also, (4) we may wonder whether a person can remark in all sincerity that a hammer she is hammering with is “too heavy” for the job, and yet be completely unmotivated to switch to a lighter hammer. In this chapter we are concerned with the moral question, but it will help to have these other cases in mind, and we will return to them later.

The side we take in the internalism/externalism debate will largely depend on how we view supposed amoralist cases. Recent work by Uriah Kriegel has made use of a distinction introduced by Tamar Gendler in order to characterize the amoralist, and this characterization seems to offer us a way of settling the debate that is arguably in the externalist’s favour, but in such a way that the intuition that motivates motivational
internalism is respected. This distinction is the one between alief and belief, first presented by Gendler in “Alief and Belief” and “Alief in Action (and Reaction),” and then employed by Kriegel in relation to internalism in “Moral Motivation, Moral Phenomenology, and the Alief/Belief Distinction.” Let us start with Gendler’s distinction.

III Aliefs Defined

Aliefs are mental states that are similar to but distinct from beliefs. The best way to introduce the notion of alief is through a few examples, as Gendler does. Here, slightly modified, is one of Gendler’s examples, which will be familiar to readers of Hume.26

A man suspended in an iron cage over a chasm may experience tremors of fear, sweaty palms, and a need to clutch at the bars around him even though he correctly believes that the cage is perfectly secure. We can even imagine that he designed the cage himself and tested it with weights, but has never stepped in it before, and certainly not at such a height. According to Gendler, his alief is a mental state with three content components: “high up above the ground right now, dangerous scary place to be, tremble.”27 However, his belief is that he is in a perfectly safe place with no need to fear. The fear component of his alief is not in accord with his belief. Gendler calls this a case of norm-discordant alief. Her articles contain many examples of norm-discordant aliefs, but aliefs need not be norm-discordant. If the man in the cage were in a genuinely dangerous position over a chasm, his alief would be norm-concordant.

Here is another example of Gendler’s. Imagine that a chef has reorganized his kitchen recently. He walks towards the drawer which used to hold the knives to get his cleaver and simultaneously remarks to another chef that he is much happier with the way the kitchen is arranged now. According to Gendler, the chef believes the knife is in a new drawer, but alieves that it is in the old drawer – another case of norm-discordant alief.28

28 Ibid., 556.
What specifically is involved in having an alief? According to Gendler,

To have an alief is, to a reasonable approximation, to have an innate or habitual propensity to respond to an apparent stimulus in a particular way. It is to be in a mental state that is … associative, automatic, and arational. As a class, aliefs are states that we share with non-human animals; they are developmentally and conceptually antecedent to other cognitive attitudes that the creature may go on to develop. Typically, they are also affect-laden and action-generating.\textsuperscript{29}

And,

as a reasonable (if cumbersome) approximation – we can say that a subject in a paradigmatic state of alief is in a mental state whose content is representational, affective, and behavioural.\textsuperscript{30}

We have already seen what this means in relation to the man in the cage. He is in a tripartite state with this content: “high up above the ground right now [representational], dangerous scary place to be [affective], tremble [behavioural].”\textsuperscript{31} Let us also see what this means for the case of the chef. Suppose that the stimulus that leads the chef to go get his cleaver is that a visitor to his kitchen has asked to see it. (The stimulus could equally be the fact that he has reached the stage in preparing a meal when a cleaver is needed, or a sudden impulse to sharpen the blade, etc.) When the chef experiences such a cleaver-demanding stimulus, a habitual pattern is triggered, which has representational, affective, and behavioural components. The state is something like this: cleaver in that drawer [representational], keen to show it [affective], go get it from there [behavioural]. The false representational element of the alief can exist and produce action even while the chef is speaking in such a way that it is clear that he is in some sense consciously aware that the cleaver is not to be found where he is going – that is, it is clear that he does not believe it is in the old drawer.

Now, though interesting (and sadly familiar!), the chef case is not a perfectly paradigmatic case of alief, as the affective component is rather muted, if there at all. We

\textsuperscript{29} Ibid., 557.
\textsuperscript{30} Ibid., 559.
\textsuperscript{31} Ibid., 559.
can imagine a case where the chef is not consciously aware at the time of any affect in relation to the cleaver, of any sense of keenness to get it and show it to the visitor. He may simply go to get it, and only feel something (frustration, embarrassment, humour) when he fails to find it there. This would not be a problem with Gendler’s analysis – she can call it a non-paradigmatic case, or stress (as in the quote above) that aliefs are only “typically” affect-laden. The case of the person suspended over a chasm has a far more apparent affective element, and so is a more paradigmatic case of alief.

Gendler’s motivation for introducing the idea of aliefs is that she believes relying on a narrower range of concepts (that is – belief, desire, imagination) in characterizing cases like the chef and the suspended man may lead us “to misattribute mental states…overlook important similarities…neglect certain continuities…and lack explanations of certain evaluations.”32 So, for instance, we may misattribute to the chef the belief that the cleaver is in the old drawer in order to explain why he goes to that drawer to get it, and we may overlook the similarities between the chef and suspended man cases if we think of the first as simply a case of incorrect belief and the latter as, say, a quite different matter of imagining one is in danger, with the image triggering the fear.

Whether Gendler is right or not that the concept of alief is needed to avoid these supposed dangers, one attraction of her work is its cautious and provisional nature. She admits that in articulating the nature of aliefs in contrast to beliefs she may have “misdrawn the boundaries” or classed together “cases that ought to be treated as distinct.” 33 This salutary caution should be borne in mind when I come to criticize the crudeness of the distinction and its employment by Uriah Kriegel.

IV Kriegel’s Application of the Alief-Belief Distinction

Let us turn to how Gendler’s idea has been put to use in relation to motivational internalism. Uriah Kriegel applies the alief/belief distinction to the motivational

32 Ibid., 555.
33 Ibid., 555.
internalism debate and more specifically the case of the amoralist. Here is Kriegel on the amoralist:

A common challenge to internalism is the argument from the conceivability of the amoralist. It is conceivable that an agent should realize that she ought to φ and yet fail to come up with the motivation to actually φ – perhaps she is depressed, perhaps she just doesn’t care enough (Brink, 1989). There are various more or less technical responses to this challenge – most notably, ‘inverted comma’ responses (Hare, 1952) – but the present [alief/belief] picture neutralizes the challenge rather straightforwardly. In conceiving of the amoralist, we conceive of someone who believes that they ought to φ but who fails to alieve that they ought to φ: someone who has reasoned her way to the realization that φ-ing would be the right thing to do, but whose motivational set-up, determined by associative habit, has not yet caught up with the realization. Thus in conceiving of the amoralist we conceive of someone whose moral beliefs are disconnected from motivation, in a way that suggests that moral beliefs are not constitutively connected to motivation; but none of this undermines the notion that moral aliefs are constitutively so connected.  

If Kriegel is correct then the motivational externalists are right (and, it seems, the internalists wrong) with respect to moral belief and its supposed necessary connection with motivation, and the internalists are right (and, perhaps, the externalists wrong) with respect to moral alief and its genuine necessary connection with motivation.

To solve the problem of the inconsistent triad mentioned earlier, Kriegel argues that the term “judgment” is equivocal. To see how this works, suppose one wishes to avoid inconsistency and one accepts the Humean theory of motivation:

Belief alone is insufficient for motivation.

---

Now consider the other two claims, which cannot both be maintained together with the preceding claim:

*One cannot make a moral judgment without being motivated accordingly.*
(moral motivational internalism)

*Moral judgments are beliefs.*
(moral cognitivism)

If the judgments in question are *aliefs*, Kriegel says one should accept that moral motivational internalism is true but moral cognitivism is false. That is, having the alief does guarantee a certain degree of motivation, but an alief-style moral judgment falls on the non-cognitive side of the cognitive/non-cognitive divide, and is not a belief. However, if the judgments in question are *beliefs*, internalism is false and cognitivism is true. That is, having a moral belief does not guarantee any degree of motivation, but moral beliefs are, of course, beliefs, and fall on the cognitive side of the cognitive/non-cognitive divide. Thus inconsistency is avoided.

Kriegel’s application of the alief-belief distinction has an initial plausibility in that it would seem to show that both internalists and externalists are partly correct, which would explain why the debate exists and also why it has remained quite intractable. Without realising it, internalists and externalists have been talking about different mental states.

If Kriegel is right, this is a tidy resolution of a long-standing problem. But can things really be so easy? I will argue that despite its attractions, the alief-solution leaves the question of the relation between belief and motivation just as open as before.

**V The Minor Concern**

Before expressing any criticisms, I want to make clear the aspects of the alief-solution to the motivational internalism problem which are helpful and which point us in, if not
the right direction, at least a direction which if followed will place us somewhere where the right direction is easier to see.

These then are the positive aspects of the solution. It is true that the amoralist’s psychology is missing something which a moral person’s has. Definitely, one thing the amoralist (and akratic person, too) lacks are a large number of (something like) powerful action-feeling-perception dispositional tendencies, and if one wishes to group together many cases of these tendencies under the name “alief” I see no serious problem with it, so long as one is cautious about the grouping, as Gendler is. Furthermore, the action-feeling-perception tendencies that a moral person has are important because they allow us to explain many typical cases of right action. When we explain a good and spontaneous action by citing the virtue of the agent we effectively explain her action by way of an explicit or implicit alief, because the tendency towards having such aliefs in the right situation is part of having the virtue. She, (being brave and caring) rushed into the burning building when she heard a baby’s scream. He (being sensitive and grateful) thanked his hosts for all their efforts. Though virtues are more than aliefs, (since an agent does not have the virtues of bravery and kindness unless she sometimes deliberates so as to choose the right action, and since people who experience norm-discardant moral aliefs are probably not fully virtuous), still, virtuous acts that occur without deliberation do rest on something like cultivated alief.

This brings us to another positive. It is good to have the non/pre/sub-rational aspects of action recognized, since they are a key part of real moral psychology. A moral psychology that believes right action always requires deliberation on the basis of beliefs is poor psychology. In addition, having our attention drawn to the habitual aspects of behaviour helps us to recognize the importance of the kind of Aristotelian training directed at virtue whereby habits of thought, feeling and action are formed, and indeed, in one of her articles Gendler quotes Aristotle to this effect.35

Furthermore, it is good that the non/pre/sub-rational aspects of behaviour are not simply characterized as being or including a “desire” or referred to with terms such as “intention”, or “goal”. These terms are I think less misleading when left for cases where there is deliberation going on and self-awareness about the situation and the possibilities

therein. Though it is not strictly speaking wrong to say that the cage designer desires to get out of the suspended cage, it is not strictly speaking right either. It helps to have some other term for the case so that we do not give misleading descriptions.

Also good is the fact that Gendler does not demote this non-rational aspect to mere “reflex”. That too would be to mislead, for whereas “desire” might overemphasise the conscious, cognitive and rational aspect of some alief cases, “reflex” does the opposite (though it may be appropriate for the suspended man case, where we might talk of a vertigo reflex or some other reflex to feel uncomfortable with heights). When the distracted chef goes to get the cleaver from the wrong drawer, this is quite different from, for instance, the reflex to blink when air is blown in the eyes, or a knee-jerk reaction. The chef had to learn that the cleaver was in the drawer, and developed his cleaver-getting routine by repetition until a reliable habit had been formed, leaving his over-stretched attention for other things. One can call the developed habit a reflex, I suppose, but again it is good to have an alternative term rather than to potentially mislead.

Finally, Gendler’s distinction is backed up by some very interesting empirical research, and we can hope that continuing research in this area will help us to understand better what is going on in moral and immoral behaviour, and especially in various kinds of akrasia.36

Despite the many benefits of marking the alief-belief distinction, concerns and problems arise, especially when we move to apply the distinction to the internalism/externalism debate. One concern that I am not going to spend time on is the concern over whether we need the notion of alief at all. Jack Kwong argues that most if not all of Gendler’s cases can be accounted for simply in terms of certain sorts of beliefs and desires.37 Similarly, in a recent article by Greg Currie and Anna Ichino it is argued, as stated in the article’s title, that “Aliefs Don’t Exist, Though Some of Their Relatives Do.”38 For instance, although Gendler has taken pains in her work to argue against the idea that

---

36 For references to empirical research see Gendler’s footnotes to Section III of “Alief and Belief”, 656-661.
alief cases are actually cases where imagination is doing the work (e.g. the suspended man imagines he might fall, triggering his reaction), there is space to disagree with her here. Now, if either of these objections is correct, and aliefs do not exist to be contrasted with beliefs, Kriegel’s use of the distinction to solve the issue of motivational internalism falls apart. Perhaps it could be salvaged in terms of some other distinction, but perhaps not. Nonetheless, that is all I want to say on the issue. I think it is more interesting and fruitful to simply grant the existence of aliefs as Gendler conceives of them, and then see whether Kriegel’s use of the distinction solves the problem of motivation.

Assuming aliefs exist, I still have two concerns – a minor concern, which connects with a major concern. Let’s start with the minor, and deal with the major in Section VI. I am somewhat concerned about how many different types of case are being grouped together under one heading. In contemporary philosophy, actions are often explained exclusively by way of desire-belief pairs, and this type of explanation can sometimes seem misleading or in some way unsatisfying. The unsatisfying way of explaining the chef’s going to the wrong drawer is to say that he had a desire to get the cleaver and an implicit or unconscious belief that the cleaver was in the drawer. This, as Gendler knows, will not do. However, though Gendler’s approach is an improvement on this, it seems to leave us with only two ways to explain action: desire-belief explanation and alief explanation. I do not want to criticize Gendler too harshly for not covering or separating out all possible cases of action explanation given that covering and separating out all possible cases was probably not her goal, but it is worth looking at what is potentially occluded by speaking in terms of only two forms of explanation just as it was useful of Gendler to look at what is occluded by speaking in terms of one.

Rather than operate with a single distinction, I think it might be more helpful to talk in terms of a continuum of types of behaviour or mental states or forms of intentionality or modes of engagement with the world. That is, why not describe a full range of cases of intentionalistic “behaviour” (in the broadest sense of behaviour, so as to include thought) that lie on a continuum or ladder beginning at or emerging from metabolism and autonomic functioning and innate reflex, where there is presumably no intentionality, and passing up through the kind of inattentive coping or learned reflex or
habit as occurs in typical cases of walking and un-self-conscious use of equipment, then up from inattentive to more attentive coping, where the behaviour is being consciously managed so as to come off, from there to conscious rule-based coping (or at least attempts to cope with entities in the environment by following explicit rules), from there to reflecting on the object of intention in relation to one’s goals and reflecting on one’s goals themselves, then on to detached unmotivated noticing of properties of the objects of intention, to decontextualised theory (i.e. science), and finally ending at reflection on theory and reflection itself, that is, the practice of logic and philosophy? I call this the I-scale, in virtue of the fact that, apart from the lowest level, the scale covers different modes of relating to all manner of intentional objects.39

What is interesting is to try and map Gendler’s alief-belief distinction onto the I-scale, so as to see the variety of what is being grouped together, and what is being separated. Obviously, alief would seem to cover or be related with the lower ends of the scale: learned reflex and habit, inattentive coping, attentive coping, and some parts of rule-based coping. Belief on the other hand seems more closely related to the upper range, but with some overlap: though inattentive coping would not seem to require or involve belief, attentive coping might, and certainly reflecting on one’s goals, noticing properties, science, philosophy and logic do.

Higher positions on the I-scale often involve reflection upon lower positions, and one can say that the higher forms of behaviour develop because at the lower ends our coping is not always successful. It is typically when we fail at one level of engagement that we shift our orientation towards the environment upwards on the scale. If a human being is to interact successfully with her environment, there should be a degree of integration between what happens at higher and lower levels: if things are going well, the higher level behaviours are either not needed, or end up serving the same ends as the lower by guiding the lower, for instance, towards employing better means to one’s ends. If things are going badly, the higher level behaviour should be such as to promote the solving of

39 I myself am simplifying matters by locating these states on a scale, as if there were only a single dimension (say, of abstraction) which separated them, rather than a cluster of dimensions: concrete to abstract, specific to general, engaged to detached, unconscious to conscious, non-linguistic to linguistic, practical to theoretical. If I were to be more precise, I would need to give an account with space for cases where one state could be higher than another with respect to one dimension, but lower with respect to another dimension – say, more detached, but less abstract. However, states that are higher on one dimension are generally higher on the others, so I feel my caricature is not too misleading.
the problem at hand, or such as to allow the agent to identify the more ultimate end at stake and to pursue that end more effectively by letting the less ultimate end go.

In terms of the I-scale, Gendler’s norm-discordant behaviours are cases where our lower I-scale behaviour is not well integrated with our higher I-scale behaviour – where reflex and behavioural routine are not being managed properly by attention and theory and so on, that is (if you like) by belief.

This may not sound like much of a concern – Gendler divides things into two, I prefer a scale. However, there is an interesting area of overlap in the middle of the scale, where the distinction between alief and belief might break down. The most problematic point is whether to characterize what Heideggerian’s would call the 

unready-to-hand or unhandy mode as involving the mental state of alief or belief. Here Heidegger introduces the mode:

Modes of taking care belong to the everydayness of being-in-the-world, modes which let the beings taken care of be encountered in such a way that the worldly quality of innerworldly beings appears. Beings nearest at hand can be met up with in taking care of things as unusable, as improperly adapted for their specific use. Tools turn out to be damaged, their material unsuitable. In any case, a useful thing of some sort is at hand here. But we discover the unusability not by looking and ascertaining properties, but rather by paying attention to the associations in which we use it. When we discover its unusability, the thing becomes conspicuous. Conspicuousness presents the thing at hand in a certain unhandiness.40

So, suppose I am sitting on a chair which I find so uncomfortable that I wriggle about trying to find a better position, and then, failing, I get up from. When I regard the chair as too uncomfortable to sit on, what I am thinking is in a sense a belief, but it is arguably not the kind of belief that one can have absent of any dispositions to solve the problem at hand – it is not the kind of belief one can have without a disposition to feel a certain way and to do something to fix things. That is, one cannot have the belief

without matching alief. The beliefs which can be detached from such dispositions are the beliefs that relate to what Heideggerian’s call the *present-at-hand* mode, the mode of detached observation of objective properties – such as the belief that the chair can support a maximum weight of 100kg.\(^{41}\) One can believe *that* without feeling anything about it or being disposed to do anything as a result. My feeling that the chair is too uncomfortable is not like that, not strictly speaking a case of belief as opposed to alief. Similarly, it would be misleading to call this “the-chair-is-too-uncomfortable” state alief as opposed to belief, simply because it is fine to speak of belief here: I *do* believe that the chair is too uncomfortable. What we have then is a mental state or mode of engagement that cannot be called an alief as opposed to a belief, or a belief as opposed to an alief. We have something else, and there is a danger that we miss this if we think solely in terms of an alief-belief opposition. We can say that it is a case of alief and belief together, but what’s interesting is that we cannot separate the two away from each other without fundamentally changing the mode of orientation. Remove the belief element, and it is no longer seems right to say that I think the chair is too uncomfortable. Remove the alief element, and it also no longer seems right to say that I think the chair is too uncomfortable. This concern, my minor concern that the alief-belief distinction oversimplifies things, leads to my major concern.

### VI The Major Concern

My major concern is that we cannot simply assume that alief and belief can be detached as they are in norm-discordant cases. More specifically, we cannot simply assume that certain sorts of belief (such as moral belief) can exist as genuine beliefs without accompanying aliefs. We may be tempted into thinking so by the fact that Gendler focuses so much on norm-discordant cases. She has a reason for doing so – she is trying to show the need in our conceptual space for an alternative notion to belief, and the norm-discordant cases help her achieve this. However, a focus on norm-discordant

cases may tempt us into assuming that all beliefs can exist equally independently of alief. Indeed, this is the temptation I believe Kriegel has succumbed to.

To see what I mean, we can start by seeing how an internalist might want to respond to Kriegel’s type of account. Rather than choose a moral case however, let us return to those other cases of judgment that I began the chapter with: delicious sushi, imprudent train-track walking, racist and homophobic slurs, and too-heavy hammers. We can examine Kriegel-like accounts of those cases, and see whether these accounts are satisfactory. I will return to the topic of morality afterwards.

Recall that the case of delicious sushi is one where someone believes, or seems to believe, that sushi is delicious, but has, or seems to have, no inclination to eat it. Call her Kate. Kriegel’s solution to the case would presumably be to say that Kate believes that sushi is delicious, but does not alieve that it is delicious. However, does that claim constitute a solution which culinary internalists and externalists would unite over? Doesn’t the claim simply assume that Kate genuinely believes what she says and assume that she is genuinely unmotivated, and then account for this by way of the alief/belief distinction? Suppose we agree that Kate has no motivation to eat sushi. Surely it remains open to the culinary internalist to question whether Kate genuinely believes sushi to be delicious. The internalist could even challenge that claim on the basis of the fact that Kate is unmotivated to eat sushi. While granting that Kate has no sushi-loving alief, the internalist could question the genuineness of her belief.

Even were the internalist to conclude, in the face of Kate’s protestations of sincerity, that Kate was speaking sincerely, the internalist might still deny that Kate genuinely believes that sushi is delicious by saying that Kate must not know how to use the word “delicious”, or by saying that Kate is using it in an odd and unacceptable way. Kate must really mean something like “popular with foodies” or the like. Otherwise, she would (in the right circumstances) be motivated to eat the sushi.

Of course, the culinary externalist will try to respond. Kate does genuinely believe that sushi is delicious and does understand what it means to call something delicious, but – say – her revulsion to the idea of eating raw fish actually prevents any motivation to eat sushi from forming. Perhaps Kate concluded that sushi was delicious when she first
tried it but didn’t know that it was raw fish. If so, then she still knows it is delicious, but she is now revolted by it and has no motivation to eat it. The internalist responds that if her motivation to eat sushi has completely disappeared, then she no longer genuinely believes that sushi is delicious. She believes it is disgusting. The externalist, of course, will disagree.

What I think is clear is that the belief/alief distinction has not helped here. It was meant to resolve the internalist/externalist debate, but has instead simply left us with much the same question as before. Even if we all grant that there is such a thing as an alief that something is delicious, and that Kate lacks it, we disagree over what it takes to have a genuine belief that something is delicious.

Consider another of our examples. Suppose an externalist describes the case of John, whom we are told expresses beliefs of the following sort, that “so-and-so is a nigger” and “so-and-so is a homo,” and suppose further that we are asked to accept that John has no racist or homophobic motivational dispositions. This, says the externalist, shows that an externalist theory of hate-speech terms is correct. An internalist disagrees, arguing that either (a) John actually does have those dispositions, contra the externalist’s claim, (otherwise why use such feeling-laden terms?), or that (b) John is using the terms in a non-standard way – jokingly perhaps, or in an inverted-commas sense, or with little or no understanding of how those terms are typically used, and not expressing the belief he might seem to be. Who is right?

A Kriegel-like solution of the case would be to say that John believes that so-and-so is a nigger, and so-and-so a homo, but John does not alieve that so-and-so is a nigger or so-and-so is a homo. That is, though he has the stated beliefs, he does not have an innate or habitual tendency to respond to black people and homosexuals in the associative, automatic, arational way that a racist or homophobe does. For instance, his face does not tend to screw up in disgust when he meets them. His fists do not clench angrily at the sight of them, he does not feel the need to avoid them, he does not find himself assuming they have character flaws, nor does he enjoy it when they are made the butt of jokes. He has the belief, but not the related aliefs.
But just as with the sushi case, this does not actually settle the issue. The internalist might allow the externalist to stipulate that in the imagined case John has no racist or homophobic aliefs. However, once that is accepted as stipulated, the question remains open whether John is genuinely using the terms “nigger” and “homo” in their proper or standard sense to communicate his beliefs. The internalist may say that any beliefs John expresses in which “nigger” and “homo” feature will not be genuine uses of the terms with their true meaning, but non-standard uses – humourous uses, inverted-commas uses, linguistically incompetent uses, or standard uses in a different dialect of English to our own, and that the evidence for this is the lack of racist or homophobic alief in John. The ball moves into the externalist’s court, and the debate continues. Similar issues arise in the case of the imprudent train-track walker, and the dissatisfied hammerer, but I think the point is clear.

Let us return to the primary topic: morality and the possibility of amoralists. I would argue that this is a case just like the others we have been considering, a case where it is at least debatable whether genuine belief can exist without at least some alief, a case where the absence of dispositions of feeling and action can be argued to show that some other interpretation of the amoralist’s utterances or mental life is required. Suppose our amoralist declares something like the following. “I know my pick pocketing of unwitting tourists is immoral, but so what – I don’t care a jot about morality.” And suppose he then heads off and happily pickpockets another tourist. The externalist wants us to take what the amoralist says at face value. However, there are other ways to characterize what is going on here.

Let us call what the amoralist says his utterance. Let us also, to make what follows simpler, allow that moral judgments (and delicious judgments etc.) can be called beliefs, without taking any substantive stand on the truth of moral cognitivism. Below I have included a table of different possible ways to characterize utterances, only one of which is straightforward, sincere expression of belief. This list is not meant to be exhaustive, but to suggest some of the different ways in which we can utter “X is A” (“sushi is delicious”, “so-and-so is a nigger”, “pick pocketing is wrong”) which are different to straightforward expression of belief. The ones in italics could be used by the internalist
to recharacterize the utterance of a putative amoralist who says, for example, that “pick pocketing is wrong.”

<table>
<thead>
<tr>
<th>Uttering: X is A</th>
<th>Further from the standard case (and definitely not belief)</th>
<th>Closer to the standard case (and possibly (not) belief)</th>
<th>Standard case (of belief)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insane raving$^{42}$</td>
<td>Flippant declaration</td>
<td>Sincere expression of belief</td>
<td></td>
</tr>
<tr>
<td>Merely reciting /singing / parroting</td>
<td>Declaring X is A with only partial understanding of what is being declared e.g. of the meaning of A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lying</td>
<td>Wishful / neurotic thinking</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baby talk</td>
<td>Suspicion (i.e. The utterer suspects that…is the case, but is not really sure.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Merely being silly / joking</td>
<td>Inverted commas declaration (using a term borrowed from an alien view of life)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Some forms of self-deception</td>
<td>Some other forms of self-deception</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fictional discourse</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In general, the absence of the amoralist’s motivation to act according to the moral judgment that he seems to be expressing can be regarded by the internalist as evidence that he is not expressing a sincere judgment, but doing something else. The externalist will see things otherwise – perhaps the beliefs are being expressed in a somewhat non-standard fashion, but they are beliefs nonetheless.$^{43}$ I am not arguing that the internalist

---

$^{42}$ The items in these columns are arranged in no particular order relative to each other, or to the other columns.

$^{43}$ Of course, all this is assuming that the amoralist genuinely has no moral motivations. An alternative account that the internalist can give is that the amoralist does genuinely believe what she is saying, but is simply insufficiently aware of her own motivations.
or externalist would be right in saying any of these things. I am simply arguing that the debate is not over by any means.

So, what can we conclude? Why is the alief/belief distinction not helping us to resolve the internalist/externalist debates? Gendler’s division of the origin of action into two distinct sources – alief and belief – potentially occludes the continuum or ladder of cases from autonomic functioning through skilful coping, to theory-based action, to detached theoretical thinking – that is, the full intentionality scale (or I-scale). It more particularly occludes the potentially important interplay between the positions on the scale.

Since acting typically involves an interplay between the different levels, and aliefs and beliefs are associated with lower and higher levels of the scale respectively, we may ask this key question:

\[
\text{What kind of interplay (if any) must exist for some thought (or mental state) to genuinely be an alief or belief of a certain sort?}^{44}
\]

This is the question Kriegel has missed. To assert, as Kriegel does, that the alief/belief distinction “accommodates the conceivability datum [of amoralist cases] straightforwardly” is simply to assume that supposed amoralist cases are genuine, and thus that little or no interplay with lower levels is required for a belief to be a belief, and this is what the internalist can resist.

\section*{VII \hspace{1em} Moral Cognitivism and Besires}

Throughout the chapter so far, I have avoided taking a stand on the issue of moral cognitivism, but now is the time to say something on that issue. In the course of expressing my concerns around aliefs I found it useful to introduce the idea of an I-scale, a scale which includes states which are clearly beliefs, states which are clearly not, and intermediate or mixed states. I also introduced a key but overlooked question

\begin{flushright}
\footnote{When I say “of a certain sort,” I mean, for instance, a belief that something is immoral. This would be one sort of belief quite different to the sort involved in the belief that something is white, or that some number is bigger than another number.}
\end{flushright}
about the nature of genuine belief. Both the I-scale and the key question bear on the issue of moral cognitivism, and it is worth making this bearing clear before finishing the chapter.

What is moral cognitivism as opposed to moral non-cognitivism? Simon Blackburn describes a way of dividing up elements of our psychology, according to whether they fall under the heading of Dionysus or Apollo:\(^\text{45}\)

<table>
<thead>
<tr>
<th>Dionysus</th>
<th>Apollo</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Emotions</em></td>
<td><em>Cognition</em></td>
</tr>
<tr>
<td>Upsets</td>
<td>Representations</td>
</tr>
<tr>
<td>Passions</td>
<td>Knowledge</td>
</tr>
<tr>
<td>Arousals</td>
<td>Truth</td>
</tr>
<tr>
<td>Excitement</td>
<td>Reasons</td>
</tr>
<tr>
<td><em>Desires</em></td>
<td><em>Knowledge</em></td>
</tr>
<tr>
<td>Impulses</td>
<td>Knowledge</td>
</tr>
<tr>
<td>Whims</td>
<td>Truth</td>
</tr>
<tr>
<td>Lusts</td>
<td>Reasons</td>
</tr>
<tr>
<td>Stances</td>
<td></td>
</tr>
<tr>
<td>Dispositions</td>
<td></td>
</tr>
<tr>
<td>Sentiments</td>
<td></td>
</tr>
<tr>
<td>Postures</td>
<td></td>
</tr>
</tbody>
</table>

We could define moral cognitivists as those who think moral judgment falls under the heading of Apollo, and non-cognitivists as those who disagree, either because they think moral judgments express an item or items on the Dionysian list, or because they think moral judgments express something else, such as prescriptions, which are not on the Apollonian list and are more closely associated with items in the Dionysian list.

The problem, as I have tried to make clear, is that this division of the terrain obscures the possibility of a certain kind of belief or knowledge which the believer cannot have without appropriate dispositions and desires. “The chair is too uncomfortable” could be one example. “Pick pocketing is immoral” could be another. It is also unhelpful, I think, to use the term “cognition” for the Apollonian items, but not the Dionysian items. Cognition is surely involved in most if not all of the items in the table.

The cognitivism debate is better seen as a debate over whether it is appropriate, or helpful, or misleading, or dangerous, to call moral judgments *beliefs* (and likewise, whether to say that these judgments can be *true* or *false*, allowing for the possibility of

---

moral knowledge, error and ignorance). In facing this question, we cannot automatically assume that this amounts to deciding whether moral judgments fall on the left or right of the Dionysian-Apollonian divide. The helpfulness of thinking in terms of this very divide is what is in question.

What unifies cases like the uncomfortable chair case, the pick-pocketing case, the hate-speech case, and the imprudent train-track case, is that the judgments involved pertain to action and reasons for action. They are inherently practical judgments. The person who judges a chair to be too uncomfortable has a reason to stand, to sit elsewhere, to buy another chair, or to complain. The person who thinks her neighbour is a “homo” or a “nigger” has, in his own eyes, a reason to treat that person with disdain, to avoid that person, to worry about house prices, or to move house.

What we want to know is whether there is good reason not to speak of practical beliefs (and practical knowledge). Some will say there is not, so long as we are clear that not every belief is practical. Practical beliefs can be contrasted with the kind of beliefs that have no special, necessary connection with action or reasons for action. Let us, just to have a name, call these neutral beliefs. Neutral beliefs fall under the Apollonian heading. They include most obviously the beliefs of the empirical sciences. By contrast, practical beliefs resist categorization under Blackburn’s division. If moral beliefs exist, they are practical beliefs, not neutral beliefs. Neutral beliefs that appear to be moral beliefs are really something else: anthropological or ethological beliefs rather than moral beliefs. “Pork is not kosher” is, for a non-Jew, an anthropological neutral belief pertaining Jewish practices. In the mouth of a practicing Jew, it is a practical belief.

Practical beliefs are obviously quite similar, if not identical, to what have been called besires. A besire is usually defined as a mental state with both directions of fit, world-to-mind and mind-to-world, where the content of the besire that has world-to-mind

\[46\] Practical knowledge should not be confused with know-how. The former is knowledge that something is the case related to action and reasons for action. The latter is knowledge of how to do something.

\[47\] The practical is typically contrasted with the theoretical, and often usefully so. However, since engagement in theory is often directly driven by practical concerns, (e.g. as it was in the Manhattan project), I chose to talk of neutral beliefs instead.

direction of fit is not the same as the content that has mind-to-world direction of fit.\textsuperscript{49} If I desire that sushi is delicious, I will try to alter the world so \textit{that I can eat some sushi, or my friends can eat sushi} (etc.), but I will also alter my judgment \textit{that sushi is delicious} if the world teaches me that sushi tastes much worse than I had anticipated or remembered. I do not much like the term \textit{besire}, preferring \textit{practical belief}, and I do not like to put too much weight on the \textit{direction of fit} metaphor, but for now let us allow that practical beliefs and besires are the same thing.

Is there any good objection to practical beliefs/besires? The objection comes from Humeans. As Michael Smith puts it,

\begin{quote}
what Humeans must deny and do deny is simply that agents who are in belief-like states and desire-like states are ever in a \textit{single, unitary, kind of state}. This is the cash value of the Humean doctrine that belief and desire are distinct existences. And their argument for this claim is really quite simple. It is that it is always at least possible for agents who are in some particular belief-like state not to be in some particular desire-like state; that the two can always be pulled apart, at least modally. This, according to Humeans, is \textit{why} they are distinct existences.\textsuperscript{50}
\end{quote}

So, suppose Jane judges a chair to be too uncomfortable. The Humean says that Jane is not in a single, unitary mental state, but in two states: a belief-like state (say, she believes that the chair is very hard) and a desire-like state (say, she wants a chair that is less hard). The Humean argues that this is clearly so because it is possible for another agent, Kevin, to be in the same belief-like state as Jane without being in the same desire-like state. That is, Kevin can believe the chair is very hard, but not want a chair that is less hard – he is glad of or indifferent to its hardness.

However, the argument need not trouble those who like to talk of practical beliefs (or besires). The most that the Humean has shown is that Jane and Kevin can share the same \textit{neutral} belief without sharing the same desires. This is perfectly true, by definition. What the Humean has not shown is that Jane and Kevin can share the same

\textsuperscript{49} Smith, \textit{The Moral Problem}, 118.
\textsuperscript{50} Ibid., 119.
practical belief without sharing the same desires. Indeed, says the practical belief supporter, Jane believes that the chair is too uncomfortable, while Kevin plainly does not. Thus the Humean argument can be resisted.

It seems that the debate reduces to a mere verbal dispute over the use of “belief”. One side draws a distinction between practical beliefs and neutral beliefs, allowing both sorts to count as beliefs. The other side says that practical beliefs are not really beliefs as such but immiscible pairings of beliefs and desires. Are these not simply two different ways of talking, both acceptable in their own way? Is there any reason to prefer one way of talking to the other?

As I see it, it might be advantageous to follow the Humean and deal in terms of (neutral) belief and desire when one is theorising about the mental life and behaviour of people, viewed with the scientist’s impartial gaze and with an eye on prediction, theoretical confirmation and falsification, and technological control. 51 To take a rather odd but current example, an anthropologist or psychologist who wished to construct a testable theory that could allow her to predict what someone will choose to sit on ought not to construct her theory in the vocabulary of chairs bring “too uncomfortable”, “good enough”, or “just right”. Instead, she would want to deal in measurable properties pertaining to comfort, such as elastic stiffness, plasticity, strain, strength, and toughness. She would try to account for the dispositions of her subjects in terms of their neutral beliefs about these neutral properties, and preferences in terms of these neutral properties, and she would feel she understood an individual’s sitting behaviour when she could reliably predict which chair a subject would choose under suitably constrained experimental conditions, on the basis of which neutral properties and neutral beliefs were present.

However, in everyday life we almost never go in for or require this kind of understanding. What we want is to understand each other in the everyday way – to make sense of each other and make sense of a situation, and thereby make sense of what we and others do and ought to do. If a guest tells me her chair is too uncomfortable, I offer her a cushion, or offer to trade seats. I do not need to think in terms of neutral desire-independent properties, and I am not after a theory on the basis of which I could predict

51 See Chapter 6.
which chairs she will prefer. I only need the kind of understanding that enables us to get on together, and get on with things together – an everyday, practical understanding. When we are in the midst of this everyday non-theoretical interaction, there is nothing wrong with seeing the world and each other in terms of practical belief, and calling it belief. Neither should we see these sorts of belief as inferior to their neutral cousins. We grow up learning the concept of belief by hearing and engaging in talk that involves both practical and neutral exemplars. To object to talk of practical belief and knowledge, and specifically moral belief and knowledge is, I think, merely a prejudice of the theoretician.

VIII Conclusion

In conclusion, the alief-belief distinction is an interesting and useful addition to our vocabulary in the philosophy of action and the philosophy of motivation, but it cannot be used to resolve the motivational internalism debate in any straightforward way, because it ignores the possibility of practical as opposed to neutral belief.
3 Motivational Internalism and Authenticity

Saying the words that come from knowledge is no sign of fully having it. For people affected in these ways even recite demonstrations and verses of Empedocles. Further, those who have just learned something do not yet know it, though they string the words together; for it must grow into them, and this needs time.\footnote{Terence Irwin, trans., \textit{Aristotle: Nicomachean Ethics} (Indianapolis: Hackett Publishing, 1999), 1147a 19-23.}

— Aristotle

I Introduction

This chapter continues our focus on the debate between motivational internalists and externalists. The internalism in question comes in an older, simpler form and in forms that have been modified in response to dialectical pressure. The simpler, much criticized version goes like this:

Simple internalism: necessarily, if a person judges that she morally ought to φ, then she is (at least somewhat) motivated to φ.\footnote{Björklund, et al., “Recent Work on Motivational Internalism,” \textit{Analysis} 72, no. 1 (2011): 125.}

One reason some have for wanting to object to this claim is that it is, if combined with a Humean Theory of Motivation, inconsistent with the moral cognitivism they support. Their concern is that, if motivation is internal to moral judgment, and beliefs do not motivate alone (the Humean Theory), then moral judgments cannot be a species of belief. Thus moral cognitivists usually try to undermine internalism.
Objectors to simple internalism maintain that there are or could be people who at least sometimes make moral judgments without being at all motivated to act accordingly, at least for a time if not consistently. These unmotivated people are the clinically depressed, the mentally or physically exhausted, the psychologically abnormal, the emotionally disturbed and (more controversially) amoralists – those who are generally (contentedly, even proudly) unmotivated by their own moral judgments.

Defenders of internalism have by and large accepted that simple internalism needs modification, and this has resulted in the presentation of conditional forms of internalism to allow for cases where people are relatively or completely unmotivated by their moral judgments. The general form, with an as yet unspecified condition C, is this:

Conditional internalism: necessarily, if a person judges that she morally ought to \( \varphi \), then she is (at least somewhat) motivated to \( \varphi \) if she is C.\(^{54}\)

Björklund et al. give a helpful taxonomy of the types of conditional internalism, identifying three basic types depending on the nature of condition C:

- \( C = \) psychologically normal
- \( C = \) practically rational
- \( C = \) morally perceptive\(^{55}\)

Some philosophers argue that the internalist connection between judgment and motivation is only secured if the person is free from the kind of mental conditions, such as depression, that undermine deliberation and action guidance. In other words, the connection is only secured if the person is *psychologically normal*.\(^{56}\) Other

---

\(^{54}\) Ibid., 126.

\(^{55}\) Ibid., 127-8.

philosophers believe the condition securing the connection is that the person must have
full rational control over her actions – she must be practically rational. Others hold
that it is moral perceptivity or discernment that is required, such that a person cannot
discern what she ought to do without being motivated accordingly. That is, the
connection between judgment and motivation is secured if she is morally perceptive.
Such a view means that her moral perceptions rest on or are constituted by some
mixture of beliefs and desires, e.g. the besires discussed in Chapter 2.

Let us call the relevant forms of internalism with these conditions normal internalism,
rational internalism, and perceptive internalism, respectively. These forms of
internalism may be able to meet the challenge of accommodating many or all cases
where the depressed, exhausted, and disturbed are unmotivated to act according to their
moral judgements. However, the actual or possible existence of a psychologically
normal amoralist would falsify normal internalism, the existence of a practically rational
amoralist would falsify rational internalism, and the existence of a morally perceptive
amoralist would falsify perceptive internalism. Additionally, if there were or could be a
psychologically normal, practically rational, morally perceptive amoralist, this would
count as decisive proof of the falsity of all three forms of conditional internalism, as
well as simple internalism.

In this chapter, I will defend a different form of conditional internalism which I call
authentic internalism, according to which authentic moral belief guarantees motivation,
and no authentic amoralist could exist. I hope to show that authentic internalism casts

57Christine Korsgaard, “Skepticism about Practical Reason,” in Creating the Kingdom of Ends,
(Cambridge: Cambridge University Press, 1996), 311-34; Smith, The Moral Problem; R. Jay Wallace,
58 John McDowell, “Are Moral Requirements Hypothetical Imperatives?” The Aristotelian Society,
McNaughton, Moral Vision (Oxford: Blackwell, 1988), Chap. 8; William Tolhurst, “Moral Experience and the
59 It is important that the amoralist here is understood as someone who makes moral judgments without
suitable motivation. Later we will see that there is a different concept of amoralism, whereby the agent is
not motivated by moral judgments but neither does she make them. The existence of the first type of
amoralist would contradict internalism, but the existence of the latter would not.
light on the nature of morality and amoralism, and helps to set limits on acceptable interpretations of the other, potentially correct, forms of conditional internalism.

The notion of authenticity involved in authentic internalism is roughly speaking the familiar one from the so-called “continental” and “existentialist” traditions (see e.g. Kierkegaard, Heidegger, and Sartre). This notion will be explained later, but as a quick sketch authenticity involves an agent taking responsibility for and ownership of her beliefs, her judgments, her decisions, and her life, as truly her own. Authentic internalism could take two forms, depending on whether authenticity is ascribed to intentional stances such as judgments/judging (version A below) or to agents (version B):

**Authentic internalism A:** necessarily, if a person judges that she morally ought to φ, then she is (at least somewhat) motivated to φ if her judging so is authentic.

**Authentic internalism B:** necessarily, if a person judges that she morally ought to φ, then she is (at least somewhat) motivated to φ if she is authentic.

If an agent is authentic, then amongst other things, her judgments will be authentic – at least, in general. Whether being an authentic agent requires that one never make an inauthentic judgment depends on how high one wishes to set the bar of agent authenticity. The issue is similar to deciding whether an individual act of immaturity is enough to mean that an agent is not mature, or an individual act of folly is enough to mean she is not wise, or an individual act of vice is enough to mean she is not virtuous. I prefer not to weigh in on such questions, but for the purposes of this chapter, I will stipulate that an authentic agent is authentic through and through, such that she cannot make a single inauthentic judgment. Likewise, I will stipulate that the making of authentic judgments is only possible for authentic agents. Given these stipulations, authentic internalism A and authentic internalism B stand or fall together.

Here is how I mean to proceed. In Section II, I will explain what is involved in an

---

60 For a more nuanced approach to the requirements of authenticity, see Chapter 4 Section IV.
agent’s being authentic, and present the argument for authentic internalism. In Section III, I will defend the argument against the objection that authentic internalism is consistent with externalism and thus misnamed, and in Section IV against the objection that it is vacuous. In Section V, I will consider whether authentic internalism is different to and superior to other forms of conditional internalism. In Section VI, I will comment on what the truth of authentic internalism means for the moral cognitivism/non-cognitivism debate. Section VII contains concluding remarks.

II Authenticity

To understand what is involved in being authentic or inauthentic, we can begin with quotations from the two philosophers most often associated with these concepts: Sartre and Heidegger. Sartre mentions two connected conditions of authenticity, one of lucid consciousness, the other of taking responsibility:

If it is agreed that man may be defined as a being having freedom within the limits of a situation, then it is easy to see that the exercise of this freedom may be considered as authentic or inauthentic according to the choices made in the situation. Authenticity, it is almost needless to say, consists in having a true and lucid consciousness of the situation, in assuming the responsibilities and risks that it involves, in accepting it in pride or humiliation, sometimes in horror and hate.61

Accordingly, to be inauthentic is to have a cloudy or false consciousness of one’s situation, and at the same time to fail to take up one’s responsibilities with respect to the situation. Heidegger’s terminology and style are more forbidding than Sartre’s, but he adds to our understanding of the nature of inauthenticity (and thereby authenticity) by relating it to a clouding of the act of choice itself, thanks to our ability to reduce our identity (as Da-sein, a human being) to that of the one, the they (Das Man).

The they [Das Man] even conceals the way it silently disburdened Da-sein of the

---

explicit choice of these possibilities. It remains indefinite who is “really” choosing. So Da-sein is taken along by the no one, without choice, and thus gets caught up in inauthenticity. This process can be reversed only in such a way that Da-sein explicitly brings itself back to itself from its lostness in the they.62

This means a person achieves authenticity when she becomes fully conscious of her choices as choices, and as her choices, rather than as externally determined requirements that she must follow because that is simply what one does, or one cannot help doing, or because that’s what they say to do.

In this chapter I am making use of the distinction between authenticity and inauthenticity, (the Heideggerian Eigentlichkeit and Uneigentlichkeit), to put forward a modified version of the internalist point of view. However, I do not wish to give a careful exegesis of the Heideggerian (or Sartrean) distinction, but only to take inspiration from Heidegger and Sartre. I am not especially worried if my explanation of the nature of authenticity simplifies, adds to, or conflicts with Heidegger’s or Sartre’s.

The notion of authenticity, as I see it, is intimately connected with the notion of maturity; indeed it is a form of maturity, and if the reader feels the word “authenticity” is too infected by unattractive links to that which is bad in continental philosophy, (e.g. to Heidegger’s notion of resoluteness applied, notoriously, in support of the Nazis), then in all that follows I would ask her to read “mature” for “authentic” and “immature” for “inauthentic”. Nonetheless, I think there is a conceptual richness in Heidegger’s (and Sartre’s) explication of authenticity that might be overlooked by speaking of maturity, and I will try to bring out this richness as we proceed. In any case, given the close link between authenticity and maturity, a good way to get an initial grasp on the nature of authenticity in relation to morality is by characterizing, very roughly, the difference between a child’s and an adult’s moral life.

In this vein, let us compare the bad child with the bad man or woman, and the good child with the good man or woman. As children, we imbibe our culture’s way of doing

62 Martin Heidegger, Being and Time, §54.
and seeing things just as we learn a language and the use of simple tools, largely by unselfconscious copying, but also by following instruction. We are not, at least at first, aware of the possibilities that lie outside our local culture’s form of life, which when we are very young is experienced in the context of the family. When very young we are not in a position to rebel against our family’s or culture’s way of doing things.

At most, a bad child is one who regularly falls into being naughty because of his impulsiveness, inattention, ignorance and forgetfulness. This is importantly different from rebellion and rejection of morality. Likewise, when a good boy does what a good boy should (when he behaves as one does and should, as Das Man does, in Heidegger’s terminology), it is not because he has committed himself to living according to the rules he is taught, just as it is not the case that he has chosen the way he pronounces his words or holds a fork. He has simply learnt (with the help of a model, and maybe with some instruction and correction) to behave as one should, and that makes him a good boy.

This is the undifferentiated stage of existence we all pass through, when we haven’t developed very much awareness of other options in how to live, nor much ability to judge them and on that basis consciously choose or reject them. Young children are not developed enough to be perturbed by the difficulty of such choices, and so neither are they able to hide from their responsibilities and pretend they have no choice (as adults are prone to), and thus to culpably allow a culture or a role to tell them what to do. That is, children do not exhibit Sartrean bad faith. In this way, children are not capable of either authenticity or inauthenticity.63

As we mature, and especially as we become capable of contemplating our own limitations, our mortality (Being-toward-death) and the meaning of our lives, it becomes possible to live authentically or inauthentically. It becomes possible to follow a social

---

63 Children are not capable of inauthenticity in the pejorative sense. “Inauthenticity”, like “immaturity”, has a neutral and pejorative sense. There is a difference between the neutral judgment that a child’s understanding of politics is immature, and the critical judgment that an adult’s is. The same applies to inauthenticity. Children are inauthentic in the neutral sense that they do not (and cannot) take a hold of their lives as their own, but are not inauthentic in the pejorative sense, which would be to flee from, wilfully ignore or repress any awareness of our existentially weighty responsibilities.
rule and inhabit a role authentically or inauthentically, and possible to break a social rule or reject a role authentically or inauthentically. The difference is that, whether accepting or rejecting a rule or role, the inauthentic person veils from herself the nature of the human condition as it applies to her: the matrix of freedom and limitation within which she must make her choices, the certainty of her own death, and the fact that in whatever she does she takes a stand on who she is and what her life is about, with all the individuating angst that involves. The inauthentic person hides the choices she makes from herself, disguising them as things she must do, things one can’t help doing, or things one simply does, purposefully stopping short of questioning her relation to that must, that can’t and that one. By contrast, the authentic person lives with angst and consciously and responsibly takes a stand.

Talk of the authentic person as “taking a stand” and “choosing” or “deciding” to live a moral life (or not to live one) may lead some readers to imagine an oversimplified line of moral (or amoral) development, one which I would like to undercut. The oversimplified (indeed false) story is one where a child learns which things are called good or bad, moral or immoral, kind or cruel, honest or dishonest, and so on, as she learns the facts of some school subject. Then, suddenly, one day, it occurs to her that these moral judgments are meant to count as reasons for her to do (or not do) things, as guides for how she should and should not live her life. On that day of realization, she either takes a stand – she accepts this role for moral judgment, or she rejects it, and thereby becomes authentic – or she runs from the decision, pretending it does not need to be made, and becomes inauthentic.

This story is misleading for at least two reasons. Firstly, one’s stand for morality (or against it, or in avoidance) is something that emerges over the course of many years, and is then usually upheld for the rest of a life. For some people, one or two days may stand out as particularly important in that emergence – there may be Damascus road transformations – but for most of us this is not how it happens, and even Damascus road conversions to a moral (or amoral) life only make sense against a background of attitudes, feelings and beliefs that are not transformed. In the Biblical story, Paul

64 The irony of inauthenticity is that even in avoiding taking a stand one effectively takes a stand.
changed his opinion on the identity of Jesus thanks to his Damascus road experience, but he retained the religious fervor and respect for Jewish law he had as Saul.

Secondly, children pick up a good portion of our moral vocabulary along with the associated patterns of feeling and behaviour, not as neutral descriptors that have no relation to feeling or behaviour. Consider, for instance, a child’s understanding and use of the word “cheat”, and how it shapes her interactions with others. Only later in life can she attain a relatively detached or contrary perspective on such concepts, deciding for instance that cheating is fine for her because there’s no reason for her not to cheat. So let us be careful of caricatures of the development to authenticity and maturity, especially in relation to morality. For the most part, and for most of us, taking a stand on morality means growing into a deeper, more sophisticated and perfected respect for the moral norms that we learnt as children.

That said, authenticity does not entail obedience to the prevailing rules of society, nor perhaps to anything we would recognize as morality. Authentic taking a stand and inauthentic avoiding of taking a stand (and thereby effectively taking a stand) can both manifest themselves in following or breaking social rules, including the rules of law and particular moral codes. Likewise, a person can authentically or inauthentically endorse a moral judgment, or authentically or inauthentically reject a moral judgment.

With these preliminary explanations and clarifications out of the way, we can turn to the argument for authentic internalism. Let us call someone who endorses at least some moral judgments a moralist, and someone who rejects all moral judgments an amoralist – here I leave behind the sense of “amoralism” discussed earlier which concerns the lack of relation between judgment and motivation and define “amoralism” in terms of judgment alone. There are four types of adult agent: the authentic moralist, the authentic amoralist (which for the sake of argument we will allow the possibility of), the inauthentic moralist, and the inauthentic amoralist. Here is the basic argument for authentic internalism:

---

65 This latter point is debatable. One may argue that true authenticity is not achieved until, say, one has come to respect others, or overcome egoism, or found a place within a community. Within this thesis though, I allow for the possibility of authentic amoralists.
1. People are authentic moralists, authentic amoralists, inauthentic amoralists, or inauthentic moralists.

2. Authentic moralists make moral judgments and are always (at least somewhat) motivated to act accordingly, and necessarily so.

3. Authentic amoralists do not make moral judgments, and necessarily so, but at most quote them or flatten them into something other than moral judgments.

4. Therefore, necessarily, if a person judges that she morally ought to φ, then she is (at least somewhat) motivated to φ if she is authentic.

Also, though not required for the argument, we can add the following:

5. Inauthentic moralists make moral judgments but may at times be completely unmotivated to act accordingly.

6. Inauthentic amoralists may express moral judgments, and may at times be completely unmotivated to act accordingly.

Since no argument is required to justify premise 1, what remains is to justify premises 2 and 3. Let us begin with 3. Though one cannot be a good man or woman without being authentic (though one can be an inauthentic upstanding citizen), it would seem possible that one can be authentic without being good. An authentic bad man would be one who looks at society’s rules, its endorsement of respect for others, of kindness, honesty, and communal spirit, and decides to have nothing to do with them. His conception of the good life, or at least the good life for him, and the type of life he wants to live, is an amoral conception. Unlike the authentic moral man or woman, he sees the good life as not even partly constituted by what we might call the moral virtues (and values), as opposed to the less peculiarly moral virtues, such as bravery. Instead, the good life for him is the life of pleasure, power, or scientific discovery etc. conceived as independent of kindness, justice, etc. He may see the need for certain virtues, such as diligence and bravery, but those virtues will be conceived of as independent of the moral virtues of justice, kindness, and so on. In order to be brave, it is not necessary in his opinion to be brave on behalf of a community, or brave on behalf of the vulnerable. Bravery could be
shown in resisting the demands of the community at some risk to himself. Likewise, diligence could, in his eyes, be diligence in carrying out a complicated criminal heist, diligence in carrying out human experimentation, or some other immoral enterprise.

The authentic amoralist does not characterize and evaluate actions as a moralist does. One way to say this is that he does not make moral judgments (i.e. premise 3), but this needs to be understood with care. Moral judgments play no part in his thinking. Now, he can use de-moralised versions of virtue and vice terms in his thinking (easier for terms like “brave” and “diligent” than “just” and “compassionate”, and presumably impossible for ‘moral’), and he can report and think about another’s thinking by using the terms they would use. However, when he looks at others’ use of these notions (“just”, “compassionate”, “generous”, “moral”), he will feel that they are misconceiving the world. If kindness is supposed to mean something defined in terms of behaving in such-and-such a way and at the same time as something valuable irrespective of who one is, including valuable to him, he will say there is no such thing as kindness – it does not exist. That is, if a connection with value and genuine reasons for action is built into the notions of the virtues and of morality, he will be an error theorist of virtue and morality, and thereby make no moral judgments.\(^{66}\) So, in his eyes, either others are mistaken about what really matters in life and what the good life really is, and this is reflected in their concepts, or others are alien – what matters to them need bear no similarity to what matters to him. Either way, he endorses no moral judgments of the sort they would endorse.

As an analogy, consider the way an authentic non-racist stands towards a term like “nigger”, which parallels the authentic amoralist’s stance towards “just” and “immoral”.\(^{67}\) We would expect that a non-racist normally has no use for the term “nigger”, given that part of the term’s standard meaning or use is the expression of an attitude towards black people she does not share. There are other terms, such as “black”, or “African American”, which can be used to refer to the people that “nigger” is used to

\(^{66}\) He may also believe that moralists do not know what bravery or diligence really are (which is to say, non-moral qualities), and that the true natures and names of kindness and justice, the kind that actually exist, are weakness and manipulation.

\(^{67}\) For an argument against an analogy between racial slurs and moral vocabulary, see Mark Schroeder, “Hybrid Expressivism: Virtues and Vices,” *Ethics* 119, no. 2 (2009): 257-309.
refer to, and which do not have or suggest the attitude as part of their meaning. Though the non-racist does not form (excuse the word!) *nigger-judgments* in her thinking, she can of course do two things. She can use a disinfected form of the word – the word can, for instance, be re-appropriated by the very people it was meant to demean to refer to themselves in a neutral or positive way. This seems to have happened to some degree with “nigger” in black youth culture and, long ago, with “Christian” in the early church community. Secondly, the non-racist can report on others’ racist beliefs and attitudes: “My neighbour hates *niggers.*” Inverted-commas or italics are the typical way to punctuate such reportage, and an unusual pattern of intonation marks it out when speaking, so as to avoid the obvious possibility of being misinterpreted as endorsing a slur.

It should thus be clear why the existence or conceivability of authentic amoralists is not of any use to the externalist who is arguing against motivational internalism of any sort, be it simple or conditional. The authentic amoralist does not, strictly speaking, make or believe moral judgments. Therefore, the fact that she is unmotivated to act morally is of no importance in deciding whether motivational internalism is true or not.

Let us now consider premise 2. The authentic moralist is a person who has achieved a kind of maturity such that one can say of her that she is committed to a particular form of life, the good life, and that the good life is conceived of by her as necessarily a *moral* life. If she does not see this as a necessary connection – believing that morality is merely contingently related to the good life rather than (partially) constitutive of it, the moral judgments she makes are not genuine or authentic moral judgments, and thus she is not the type of person covered in premise 2. The authentic moralist is committed to living morally, and this will manifest itself both in her judgments (about what kind of life is worth living etc.) and in her motivation to act accordingly. If she is not motivated to act morally, she has not matured enough to be called authentic. Of course, she may fail to act morally on some occasions, when she is very tired or stressed or faced with strong temptation, but that does not mean she was not or is not motivated to do the right thing. Rather, her motivation is at times overcome by something else. This is shown by the fact that she will regret what she has done and take steps where possible to prevent it.
happening again or reduce the chance that it will. Acting immorally is, for her, acting out of character. If her character is not such that we can say she acted out of character when she acted immorally, she cannot be called an authentic moralist. Thus it is a necessary truth that she is motivated to act according to her moral judgments – i.e. premise 2. (The possible vacuity of premise 2 will be examined in Section III.)

Although not strictly necessary for my argument, in which 5 and 6 play no role in justifying the conclusion, it is worth looking at the inauthentic cases as it may help to clarify further what authenticity is by contrast. Inauthentic men and women can be inauthentic amoralists and the inauthentic moralists. What distinguishes the two is (roughly) whether they say they care about morality – the inauthentic amoralist says she doesn’t, and the inauthentic moralist says she does. However, what is more interesting is what they have in common. Both types use moral language as chatter, as idle talk. Here is Heidegger on the nature of idle talk (or Gerede):

[S]ince this discoursing has lost the primary relation of being to the being talked about, or else never achieved it, it does not communicate in the mode of a primordial appropriation of this being, but communicates by gossiping and passing the word along. What is spoken about as such spreads in wider circles and takes on an authoritative character. Things are so because one says so. Idle talk is constituted in this gossiping and passing the word along, a process by which its initial lack of grounds to stand on increases to complete groundlessness. And this is not limited to vocal gossip, but spreads to what is written, as “scribbling.” In this latter case, gossiping is based not so much on hearsay. It feeds on sporadic superficial reading: The average understanding of the reader will never be able to decide what has been drawn from primordial sources with a struggle, and how much is just gossip. Moreover, the average understanding will not even want such a distinction, will not have need of it, since, after all, it understands everything.68

---

68 Heidegger, Being and Time, 158. Heidegger’s comments on idle talk, and inauthenticity more generally, unfortunately somewhat confuse the undifferentiated “positive” phenomena (which we had better not call “idle talk” or “inauthenticity”) that underlie any way of being and the genuinely inauthentic, which is only a possibility for Da-sein. See Hubert Dreyfus, Being-in-the-World, 154.
The inauthentic amoralist’s and inauthentic moralist’s “knowledge” that lying is wrong is of the same sort as someone’s “knowledge” that Paris is the capital of France (though he has never been there), that a carbon atom has 6 protons (though he has never done any serious science), and that *Great Expectations* is great literature (though he has only skimmed the back cover). With the inauthentic types, moral truth is often leveled down and trivialized to stand alongside all other *information*, so that it has no connection with any genuine stance on what constitutes the good life. This type of amoralist or moralist has no problem with answering pub-quiz questions about morality…

**Which of these is a vice?**

A) Gluttony  
B) Glastonbury  
C) Glittery

**Who was the more moral man?**

A) Goebbels  
B) Gandhi

…because he has picked up both a language and a good general knowledge of *all* kinds of matters – he “understands everything”. What he lacks, though, is an authentic relationship with what he is saying, whereby what he is saying *matters* to him. It is very easy for the inauthentic amoralist to say “I know people ought not to steal, but I don’t care about doing what people ought to do”, because his connection with the word “ought” is inauthentic. Similarly, it is very easy for the inauthentic moralist to say “I know people ought not to steal, and I try to do what I ought to do”, but to (a) steal in any case, without genuine guilt thanks to her tendency to find excuses for her actions, or to (b) avoid stealing, but only for extra-moral reasons such as that thievery might be punished. In some rather anemic sense of “believe”, of course, inauthentic moralists believe what they are saying, but in another (authentic) sense of “believe”, they do not.

**III Externalism**

The first objection to respond to concerns whether or not *authentic internalism* is actually a form of *motivational externalism*, and is therefore misnamed. Recall that authentic internalism is the thesis that:

*Necessarily, if a person judges that she morally ought to φ, then she is (at least somewhat) motivated to φ if she is authentic.*
A proponent of externalism will remark that this thesis is consistent with externalism. Externalism allows that some people could be motivated by their moral judgments (or moral beliefs, accepting cognitivism), but states that at least some people will not be motivated by their moral judgments or that at least such unmotivated people are conceivable if not actually existent. Authentic internalism seems to say the same thing: the people in the first group, i.e. those who are motivated, are called authentic judges/believers, and the people in the second group, those who are unmotivated, are called inauthentic judges/believers. If authentic internalists admit that at least some inauthentic judges/believers exist, or even merely could exist, then authentic internalism amounts to nothing more than a form of externalism. Since this is so, says the objection, the view should not be called authentic internalism but authentic *externalism*.

First, one should note that similar arguments might be directed against other conditional forms of internalism, such as normal internalism, rational internalism, and perceptive internalism, introduced in Section I. Normal internalism admits that psychologically abnormal individuals may be unmotivated by their moral judgments. Likewise, rational internalism admits that practically irrational individuals may be unmotivated, and perceptive internalism admits that the morally unperceptive may be unmotivated. Thus all forms of conditional internalism could be taken to be disguised forms of externalism.

I think the charge of disguised externalism is most easily seen to be mistaken when directed at rational internalism, so I will explain how the charge fails against rational internalism before showing how a similar response works to defend authentic internalism.

The point, function, or heart of a judgment about the best course of action is, surely, that it is to be used to guide action, and for a practically rational person it does so.\(^\text{69}\)

*Someone who makes a judgment about the best thing for him to do, but who does not*

---

\(^{69}\) I accept that moral judgment has an essence or function. See Gunnar Björnsonn "How emotivism survives" on the optational and etiological function of moral judgment. Unlike Björnsonn, I prefer not to speak of the function of moral judgments as a *causal* function. To understand my reluctance, see Chapter 6 on the meanings of “cause”.
use it to guide his action, misses the very essence or function of practical judgment. In one sense, the practically irrational person’s judgment that X is the best thing to do is the same as that of a rational person who judges that X is the best thing to do – that is, they share the same view about what it is best to do. However, in a very important sense the practically irrational person’s judgment is not a full or proper judgment – it is a shadow of true judgment, earning the name of judgment only because of its similarity to true judgment. One way to say this would be to exploit the ambiguity of the word “judgment”, which can refer to either the act or object of judgment, and say that the irrational person’s object of judgment is the same as that of the rational person (i.e. what she judges), but her act of judgment is not (i.e. how she judges), as it lacks the link with motivation. However, this kind of talk should only be seen as a rough way to tease apart the sense in which the irrational agent does, and the sense in which she does not, judge “as the rational agent does.”

The rational internalist can resist the charge of disguised externalism by saying that she understands externalism to be the claim that someone can be practically rational and yet unmotivated by her moral judgments. This is the thesis she objects to, since only the judgments of the practically rational are true (i.e. genuine, full, proper) judgments. The quasi-judgments of the practically irrational are irrelevant to her thesis.

The application to authentic internalism should be readily apparent, and needs little comment. The authentic internalist can say that her understanding of externalism, the position she is rejecting, is as the thesis that one can authentically make a moral judgment and yet be unmotivated to act accordingly. For, she argues, only the judgments of the authentic are proper, which is to say authentic judgments. The quasi-judgments of the inauthentic are irrelevant to her thesis.

To make this a little more persuasive, I will point out an analogy between moral judgment’s relation to authenticity and the making of promises’ relation to sincerity. Suppose that teenage brothers Andy and Billy make a promise to their mother to clean their bedrooms while she is out at work. While Andy makes the promise with every intention of carrying it out, Billy makes it with no intention of carrying it out, but as a
way to get his mother off his back. We can all agree that whatever their intentions, Andy and Billy both made a promise. We do not say that Billy, who had no intention of doing what he said, did not actually promise to clean his room. If that were so, we could not say that he broke his promise when he failed to clean it. The uttering of the right words in the right context is sufficient for the illocutionary act of promising to come off. However, this does not mean that Andy and Bill’s promises are equally good instances of promising according to its true essence or true function. Only sincere promises are true promises, in that sense. One aspect of this is that there could be no institution of promising if all promising were insincere. Likewise, one could not teach a child what promising was by using only insincere examples. Insincere promising is parasitic on sincere promising, with the latter being – again, in one sense but not another – what true promising is. In the same way, inauthentic moral judgments – those with no connection to motivation – miss the essence or function of moral judgment, and thereby fail, in a sense, to be cases of true judgment, and for that reason and in that sense, I call my position authentic internalism rather than externalism.

It is perhaps necessary to make one more comment, lest the analogy between authenticity and sincerity be taken too far. When an inauthentic person expresses a moral judgment, she might nonetheless be sincere in her way, especially if her inauthenticity is that of a child, or like that of a child. Inauthenticity does not imply straightforward insincerity. We can imagine an inauthentic person who has not yet gained a proper understanding of what she is saying, and what it takes to truly mean it, yet still – given what she does know – is expressing something she feels is true, and not – as Billy was – lying, or aiming to mislead. It is once again helpful to think of children in this regard. The child who expresses support for the Labour party that her father also supports is not (except in special cases) being insincere. She may express her convictions with all the sincerity she can muster. Nonetheless, if she is too young to understand what political commitments involve, she cannot be taken to have expressed an authentic political commitment. She is sincere but inauthentic (or, more charitably, pre-authentic).

70 See Blackburn, Ruling Passions, 61-2, on the parasitic as regards internalism, loving/hating and forbidding.
IV Vacuity

Let us now turn to the second major objection. Authentic internalism, like other forms of conditional internalism, is open to a charge of vacuity. To see this, we can examine the charge laid against Michael Smith’s conditional internalism, a rational internalism, a charge which with very little change can be redirected against authentic internalism.

According to Smith,

If an agent judges that it is right for her to φ in circumstances C, then either she is motivated to φ in C or she is practically irrational.71

Adina Roskies comments that Smith’s position is unsatisfactory without a further account of what it is to be practically rational. For if, as is often held, to be practically rational is merely to desire to act in accordance with what one judges right or best, then [the claim above] is trivially true.72

Similarly, James Lenman criticizes Smith’s claim that an agent’s belief about what he has reason to do causes him to desire to act accordingly if he is rational. Lenman argues that this is only true on a reading of ‘rational’ whereby the latter claim simply says, in effect, that my beliefs about normative reasons cause the appropriate desires insofar as I am someone whose beliefs about normative reasons cause the appropriate desires. That is a conceptual truth but it is not a very interesting one.73

71 Smith, The Moral Problem, 61.
Turning to authentic internalism, the charge is that my thesis of authentic internalism essentially amounts to the uninteresting truth that:

*If an agent believes that it is right for her to φ in circumstances C, then either she is motivated to φ in C or she is inauthentic, which is to say she is someone who can be unmotivated to φ despite believing that it is right to φ in C.*

Surely, says the criticism, this is “a conceptual truth but it is not a very interesting one” (in Lenman’s words), or “trivially true” and “philosophically anaemic” (in Roskies’).74

In response, I agree that it would be uninteresting to assert that authentic agents are motivated by their moral judgments if an authentic agent was that defined as someone who is motivated by her moral judgments. I do grant that it is part of the definition of authenticity that an authentic agent is so motivated, and that is why the thesis of authentic internalism begins with the word “necessarily”. However, though a connection of judgment with motivation is part of what I take authenticity to be, it is not the whole. Furthermore, we can say something to explain why motivation and judgment are linked for the authentic agent by looking at what else is involved in authenticity. This is, I hope, one thing that makes authentic internalism philosophically “interesting”.

A full characterization of authenticity involves many, if not all, the following ideas:

1. An authentic agent is deeply aware of her *facticity*, that she is forever and unavoidably limited by such things as her social context, upbringing, and particular array of talents and deficiencies of talent.
2. An authentic agent is deeply aware of the costs of her decisions and commitments, how choosing one thing makes other things difficult or impossible.
3. Notwithstanding point 2, an authentic agent is deeply aware of her ability to step back from her cultural inheritance and question, always from a committed perspective rather than pure detachment, whether individual parts of that

---

74 Roskies, “Are Ethical Judgements,” 54.
inheritance have the value commonly ascribed to them or not.

4. An authentic agent is deeply aware that 1 to 4 are not just her condition, but the human condition. Others are in the same boat.

And perhaps…

5. An authentic agent is deeply aware of her inevitable death, and that this has implications for her projects.75

As regards morality…

6. An authentic agent is aware that moral commitments involve choices about how to live, choices that involve costs (see 2) as well as benefits, and that these choices are ultimately made by individuals who bear responsibility for their choices.

7. An authentic agent understands moral grammar and (something like) authenticity – she knows that to express a moral judgment is to, if authentic, express a commitment of a certain sort.

8. The authentic agent is resolute and relatively self-controlled. When she decides on doing something she is motivated to do it, and remains motivated to do it until it is done, or until such time as she sees there as being good reason not to do it, though she may sometimes fail to act on a decision in trying circumstances. Likewise, when she thinks that something has value, she is motivated to pursue or preserve it, though other concerns or lack of opportunity may prevent this motivation from being acted on.76

On the basis of these points, we can say that an authentic agent is one who will not

---

75 Heidegger put great stress on the link between authenticity and Being-toward-Death. (See Being and Time, Division II, 1). I believe Heidegger overemphasised the link, and the reader is free to ignore claim 5 if he or she likes.

76 It is a useful exercise, which I leave to the reader, to consider how poorly the typical child does in these respects, and how much better but imperfectly the typical teenager does, and this will give some idea of what the inauthentic adult’s position is.
express a moral judgment that she is not committed to, which is to say she will not express a moral judgment that she is not motivated to act in accordance with. This is not to say that her motivation always results in her doing the moral thing, any more than that a sincere promise is always followed by the fulfilling of that promise.

It is the interconnection between points 1 to 8 and the thesis of authentic internalism that saves authentic internalism from being vacuous and philosophically uninteresting. If there is a form of vacuity here, it is much like the vacuity of a mathematical demonstration of equivalence in algebra, such as the proof that $e^{i\pi} = -1$. Though the final line of the demonstration (i.e. $\ldots = -1$) says nothing more than the first line ($e^{i\pi} = \ldots$), such that the whole demonstration could be written in reverse order and still constitute a demonstration, the ability to construct the demonstration evidences a grasp of the mathematical concepts involved, and can be informative in that sense.

V Conditional Internalisms

We now turn to the difficult question of the relation between the form of conditional internalism I have defended and the other forms of conditional internalism. There would seem to be two possibilities – either the thesis of authentic internalism is equivalent to one or more of these other forms, or it is not. We can ask the following questions about our concepts:

- Can an agent be practically rational and morally inauthentic?
- Can an agent be morally authentic and practically irrational?
- Can an agent be psychologically normal and morally inauthentic?
- Can an agent be morally authentic and psychologically abnormal?
- Can an agent be morally perceptive and morally inauthentic?
- Can an agent be morally authentic and morally imperceptive?

However, to set about answering these questions is to assume that we have a sufficiently precise understanding of the terms involved to ground an answer. Each of the ways of describing an agent above makes use of a term or terms that already had uses in
ordinary language before philosophers latched onto them and attempted to use them, with explicit or implicit “refinements”, in their discussions. It is, I think, too much to expect that out of such a history of usage we will be able to extract the correct definitions of the terms, such that the above questions could be definitively settled on the basis of pre-existing truths about their meanings. Instead, if one were to answer the questions, one would effectively be stipulating how one intended to use the terms. Such a stipulation could not drift too far from the history of usage, lest it be misleading, and that this was not the case might need some defending, but it would effectively be a stipulation nonetheless.

That is, I am willing to admit that the terms involved have sufficient elasticity of meaning to allow a philosopher to stretch them about so that the answers to the questions above were all “no”, and at the same time I am willing to admit that a different philosopher could stretch the terms in different ways and yield one or more “yes” answers. Neither philosopher would have told us the correct meaning of the terms, but neither philosopher would have told us something incorrect.

Therefore, I think it is fruitless to concern ourselves with whether, with respect to the extension of terms, my form of conditional internalism maps out the same territory as the others. However this does not mean that the intensions of the terms are the same, or that the connotations as opposed to denotations are the same, and this is where we can begin to see the value and distinctiveness of authentic internalism as opposed to other forms of conditional internalism.

What unites all forms of conditional internalism is that they see breaks in the internalist belief-motivation connection as indicative of some kind of defect. Internalisms therefore open up a space for the criticism of agents, and this is what I see to be the usefulness of the conditional internalist theses: they highlight what kind of criticism we think is appropriate to direct against the morally unmotivated moralist (i.e. the person who expresses, but is not motivated in accordance with, a moral judgment).

Perhaps, as I said before, one cannot be inauthentic without being practically irrational,
psychologically abnormal, and morally imperceptive. Still, the fact that these concepts differ in their connotations means that criticism expressed in these terms will lead audiences to picture situations in different ways, with different aspects foregrounded and others backgrounded, and to behave in different ways accordingly.

As an analogy, suppose I wished to make a request to the Queen of England. I might address her as *your majesty*, *Elizabeth*, or *Lizzie*, and even though these terms refer to the same person (they have, at the time of writing, the same extension), they connote different things about the way I see our relationship and my attitude to the monarchy. Depending on how I addressed her, the Queen would see me and my request in a certain light and be more or less moved to grant my request. The choice of one form of address rather than another would be judicious or injudicious depending on the effect I desired.

In the same way, suppose I am confronting an unmotivated moralist. It can make a difference whether I tell her (and other observers) that she is (A) practically irrational, (B) morally imperceptive, (C) psychologically abnormal, or (D) morally inauthentic, even if she cannot be one of these things without being one or all of the others.

It is too difficult to map out exactly how these terms differ in sense and connotation, but we can see that they do not equally make it likely that the criticized agent will:

a) Check the logical connections between, or empirical evidence for, her moral beliefs and other beliefs.
b) Wait for her lack of motivation to change by itself.
c) Seek out pharmacological treatment, or some form of psychotherapy.
d) Happily or unhappily accept her condition as unchangeable.
e) Look again at the case to which her judgment applies, hoping or willing herself to see in it what a moral person would see.
f) Feel guilty, berate herself, and demand more from herself.
g) Try to (re)articulate her core commitments.
h) Step back from and question what she feels about aspects of her upbringing.
and cultural environment – what she is willing to go along with and what she resists.

i) Commit to some new way of living.

The first advantage authentic internalism has over the other forms is that it is the most explicit about what we might call the *existential* nature of genuine moral belief, and it is in this direction, or in this light, that I think the criticized agent should direct her critical attention. I use *existential* in the manner of the continental or existentialist philosophers (and not to refer to the empirical, logical or mathematical notion of existence).

Existential concerns pertain to the type of existence appropriate to an ontological being such as you or I, a being whose existence is an issue for it – a Da-sein. If her position is seen in an existential light, she will be more likely to engage in practices (f) through (i), which are what I think are most appropriate and useful in her case, rather than (a) through (e), which are less appropriate or useful, though not always or totally useless.

In contrast, even if rational internalism is true, when applied as criticism it overemphasizes (a) as opposed to (f)-(i). Likewise, normal internalism overemphasizes (b)-(d), and perceptive internalism overemphasizes (e).

Secondly, authentic internalism emphasizes the full nature of our duty – to manage and align our thoughts, feelings, words and actions, our very lives. This is a duty to integrate our orientation to the world all along the intentionality scale of Chapter 2, to integrate *alief* and *belief*, feeling and action.

The third advantage authentic internalism has is that, if one accepts authentic internalism as characterizing the connection between motivation and moral judgment well, it can also serve as a helpful guide to permissible interpretations of the other forms of conditional internalism. That is, what it is to be practically rational, morally perceptive, or psychologically normal must be characterized in such a way that one could not also be inauthentic, if one is to have a conditional internalism that is true.

For instance, an internalist might make the mistake of explaining what it is to be
psychologically normal as follows:

Definition: To be *psychologically normal* is to be free of any brain damage or deformity that could impair normal brain functioning and mental activity.

If this is what one meant by psychologically normal, it would be possible to be normal but inauthentic, as inauthenticity is not a cause or symptom of a medical condition, or not always so. A normal internalism based on this definition would be false: brain normality is no guarantee that one will be motivated by one’s moral judgments.

Likewise, consider the following definition of moral perceptivevity.

Definition: To be *morally perceptive* is to be able to correctly answer all (or most) questions related to moral judgment, such as those about what one morally ought to do or not do in a specified situation.

This definition will not do because an unmotivated inauthentic agent and an unmotivated authentic amoralist can have a very good grasp of what morality requires (or is said to require). This understanding does not need to generate motivation, just as a perceptive anthropologist who is not a true member of the society she investigates need not be motivated by the norms she sees them as following.

Similarly, we should be wary of understanding practical rationality in these terms:

Definition: To be *practically rational* is to act so as to maximize one’s expected gain as determined by one’s goals, whatever they may be, given the information one has.

An inauthentic agent can pursue her goals in a highly intelligent and consistent way (at least, in as consistent way as one can given that one’s goals often conflict), but – because she has never taken a view of her life as a whole – her goals will not be integrated by a coherent view of what life is about, but be more like a list of separate
elements with an associated level of attractiveness. For such an agent, though she may give lip service to morality, living morally may not figure in her list of goals at all (since moral truths are treated as mere information), or figure but with a very small associated attractiveness and thus little motivation to act accordingly. This would make rational internalism false.

In this way then, the thesis of authentic internalism can help us to understand what is right about the other forms of conditionalism, and how they can go wrong.

VI Moral Cognitivism

Unconditional motivational internalism, Humean theory of motivation, and moral cognitivism arguably form an inconsistent triad.\(^\text{77}\) That is, it is often supposed that it is not possible for (I), (H) and (C) to all be true:

\begin{align*}
\text{(I)} & \quad \text{If a person makes a moral judgment, then she is necessarily motivated to act accordingly.} \\
\text{(H)} & \quad \text{Beliefs without desires are motivationally inert.} \\
\text{(C)} & \quad \text{Moral judgments are constituted by beliefs.}
\end{align*}

For those who accept the Humean theory of motivation (H), the truth of motivational internalism therefore implies the falsity of moral cognitivism. Many internalists have been led to accept moral non-cognitivism by this line of thought.\(^\text{78}\)

We can now ask whether the same pressure to accept non-cognitivism is exerted by authentic internalism as is exerted by unconditional internalism. Thesis (I) above must be modified so as to read:

\footnotesize
\begin{itemize}
\item [\(^\text{77}\)] See Chapter 2, Section IV. I say “arguably” because there is a difference between the claim that the existence of a judgment guarantees the existence of motivation (as in (I)), and the claim that it is the judgment itself that motivates (as is suggested by talk of beliefs being “motivationally inert” in (H)).
\item [\(^\text{78}\)] See, e.g., Blackburn, Ruling Passions.
\end{itemize}
(I)’ If a person makes a moral judgment, *then so long as she is authentic*, she is necessarily motivated to act accordingly.

It is not easy to see whether (I)’, (H) and (C) are mutually inconsistent. At first sight, the authenticity clause in (I)’ weakens the thesis so as to provide enough wiggle room for all three claims to coexist. However, the way I have introduced moral authenticity is to argue that it is of the *point, function or heart* of moral judgment that it be accompanied by motivation and reliably lead to action. This starts to sound like something that might not be true of “pure” belief (belief independent of desire), or the purely cognitive, especially if one’s paradigm of belief and cognitivity is empirical belief. This narrow conception of belief (which I called neutral belief, in Chapter 2) will entail that authentic internalism is just as inconsistent as unconditional internalism with the (H) and (C) pair. However, a more generous conception of belief (one that admits practical belief, and the one which I prefer and believe is more faithful to everyday usage) is not, and this more generous conception makes the Humean theory of motivation deeply problematic, in that this generous conception is at odds with the strict separation of mental states of belief from mental states of desire and intention.79 I do not have space now to give this issue the full attention it deserves, but we can say that the answer to the question of what implications authentic internalism has for the moral non-cognitivism/cognitivism debate depends on what one takes the nature of belief and cognitivity to be.

VII Conclusion

I have, in my explication and defence of authentic internalism, been following the Austinian procedure: “There’s the bit where you say it and the bit where you take it back.”80 After articulating the authentic internalism thesis, in its defence I took it back somewhat by admitting (1) that there *is* a sense in which externalism is true, due to the phenomena of akrasia and inauthenticity, (2) there *is* a sense in which the authentic

---

79 See Chapter 2 Section VII comments on the *desire* theory.
internalist thesis is vacuous, (3) there is a sense in which it asserts nothing different from what other forms of conditional internalism assert, and (4) the thesis may not rule out moral cognitivism. Given these concessions, I prefer to regard what I have offered as a perspective on the connection between moral judgment and motivation rather than a theory or thesis. It is a perspective which emphasises rather than asserts – it emphasises a deep relation between moral judgment, motivation, and the kind of seriousness and individuation inherent in making life commitments, but is not so bold as to assert anything that could be undermined by psychological or sociological research.
4 Three Concerns for Authentic Internalism

I Introduction

In the previous chapter, I explained and defended a form of conditional motivational internalism called authentic internalism. Rather than asserting that authentic internalism was the only correct theory or position with respect to the externalism/internalism debate, or the only correct form of conditional internalism, I offered authentic internalism as a guide to a helpful perspective on the debate, and “authentic” as a helpful term. Thinking in terms of authenticity as opposed to inauthenticity is a good way of approaching the issue of why some people seem not to be motivated to do what they admit is the moral thing to do, or not to be motivated to refrain from doing what they admit is an immoral thing to do. Looking at the issue in this way may be more helpful (even if not more correct) than looking at the issue in terms of rationality, psychological normality, or moral perceptivity. In addition, to connect Chapter 3 to Chapter 2, I suggested that authentic moral belief was belief that was sufficiently integrated with other modes of behaviour or orientation to the world, including those modes Tamar Gendler calls alief (or what I saw as the lower points on the intentionality scale).\(^{81}\) Authentic moral belief was belief that was sufficiently integrated with dispositions to see situations in a certain way, to feel certain feelings, and to act accordingly.

Now, an agent’s actions and (to a lesser extent) her feelings are publicly observable. If inauthentic belief is belief that is insufficiently integrated with feeling and action, it would seem to be possible, at least in principle, for an observer to see that an agent’s supposed moral belief is disconnected from her feelings and actions, and to correctly

\(^{81}\) See Chapter 2 Section V and footnote 39.
identify her supposed belief as inauthentic. The observer could do this even against the protestations of the agent herself. The agent might believe that her moral belief was the genuine article, and yet be wrong, and an observer could know this.

This could be seen as concerning. It is generally thought to be uncontroversial that an agent’s knowledge of her own beliefs, feelings, desires, sensations, and plans is superior to or privileged over the knowledge that others have of these matters. Some would go so far as to say that an agent simply cannot be wrong about these matters (or some subset of them). If I believe I am in pain, then I must truly be in pain. If I believe that I believe in God, then it must be true that I believe in God. These people will have difficulty accepting the possibility of inauthentic belief, or may worry that accepting this possibility is the first step on a slippery slope to a general scepticism wherein agents have no special authority over the contents of their own minds.

More modestly, accepting the possibility of inauthentic belief seems to put pressure on, or be in tension with, the possibility of akrasia. The concern is that what we formerly would have wanted to call “akratic”, must now be called “inauthentic” rather than “akratic”. A person who we might formerly have called akratic, because of a mismatch between what she says she believes it is right to do and what she does, would merit the title of inauthentic moralist rather than akratic because, the argument goes, the fact that she does not act in accordance with what she says she believes shows that her belief is inauthentic, rather than genuine and authentic. A true akratic would be a person who genuinely believes that something is morally wrong and yet does it anyway, or who genuinely believes that something is morally obligatory or the best thing to do, morally speaking, and yet does not do it. This could seem on my analysis to be impossible. The reader may therefore wonder whether I endorse the Socratic position in which akrasia is impossible. I do not endorse this position, but this concern needs to be answered.

In addition, there does not seem to be anything in principle to prevent the observer and agent from being the same person. That is, an agent might be able to identify one of her own beliefs as inauthentic.

For a collection of quotes from philosophers such as Descartes and Hume to this effect, see William Alston, “Varieties of Privileged Access,” American Philosophical Quarterly 8, no.3 (1971): 223-241.

Some have (mistakenly, I think) seen Gilbert Ryle’s The Concept of Mind as presenting the view that we know our own minds from observing our own behaviour in fundamentally the same way as other observers know our minds. See Crispin Wright, Barry C. Smith, and Cynthia Macdonald, eds. Knowing Our Own Minds (Oxford: Clarendon Press, 2000), 208.

Socrates defends this view in Plato’s Protagoras.
A third concern develops from the concern over akrasia. In Chapter 3, I explained that there were two forms that authentic internalism could take, depending on whether authenticity was attributed to persons, or to judgments:

*Authentic internalism A:* necessarily, if a person judges that she morally ought to φ, then she is (at least somewhat) motivated to φ if her judging so is authentic.

*Authentic internalism B:* necessarily, if a person judges that she morally ought to φ, then she is (at least somewhat) motivated to φ if she is authentic.

In order to streamline Chapter 3’s argument for authentic internalism, I stipulated two things, namely that an authentic agent is authentic *through and through*, such that she cannot make a single inauthentic judgment and that the making of authentic judgments is only possible for authentic agents. By stipulating in this way, I made sure that authentic internalism A and authentic internalism B stood or fell together. Though this made the argument more straightforward, it leaves us with a concern. The stipulations set an awfully high bar for achieving authenticity, and the concern is that none of us meet it. If it is true that we all, at some time, act contrary to our moral beliefs, does this mean that we are all inauthentic? Given these stipulations, isn’t one act of (so-called) akrasia actually proof of (1) the absence of moral motivation with respect to the case at hand, (2) proof of judgment inauthenticity with respect to the case at hand, (3) proof of general agent inauthenticity, and finally therefore (4) proof of the inauthenticity of all the agent’s moral judgments across all cases? If it is proof, then unless we are perfect saints, we are all inauthentic, and every one of our judgments is inauthentic, and we can probably never hope to be authentic or to make even one authentic judgment. But this hyper-idealistic view is at odds with my claim that thinking in terms of the authentic/inauthentic division is the most helpful way of thinking about the issue of moral motivation. One possible solution is to speak of degrees of authenticity. That is, some of us, or some of our judgments, are more authentic than others. However, the worry is that once we make this kind of move, we will lose the necessary connection between authenticity and motivation given in both A and B versions of authentic internalism. If this necessary connection is lost, we end up with a form of motivational externalism.
My goal in this chapter is to respond to these three areas of concern. Section II will argue that there can be good grounds for charging a person, or her belief, with moral inauthenticity, even when she holds that she, or her belief, is genuine and authentic, but this does not mean there is no such thing as first-person authority. Section III will explain how authentic internalism leaves space for a genuine distinction between akrasia and inauthenticity. Section IV considers whether the stipulations regarding authentic internalism A and authentic internalism B need to be replaced with something less idealistic, and if so, with what, and whether this affects truth of the authentic internalism thesis, and argues that we should remain flexible about what precisely the internalist thesis states. I conclude in Section V by summarising what the whole argument from Chapters 2 to 4 means for the motivational internalism debate.

## II First Person Authority, Inauthenticity, and Grounds for Suspicion

Concerns arise when we consider what are, and whether there can be, grounds for charging someone with being morally inauthentic or having a morally inauthentic belief. In particular, we may be troubled if these grounds are taken to be something that an outside observer can determine as well as (or even better than) the person who expresses the moral belief. 86

This is troubling to someone who holds there to be a relatively simple relationship between second order first-person belief and first-order first-person belief, such that one cannot believe that one believes a proposition P without actually believing P. We can call this relationship *containment*. Sydney Shoemaker claims that there is this sort of “intimate” containment relationship between second and first-order belief, one that is

86 A similar concern arises over whether an outside observer can know that a person is motivated or unmotivated to act in some way or other. These concerns are obviously related: since I have defined authenticity such that there is a necessary connection with motivation, if we have grounds for claiming that someone is unmotivated to act according to her moral belief, then these are also grounds for saying she (or her belief) is inauthentic.
much more intimate than what we find in the case of other standing second-order beliefs – those that ascribe beliefs to other persons, and those that self-ascribe beliefs from the “theoretical stance,” i.e., on third-person evidence. One way of expressing this claim is by saying that in such cases the standing second-order belief contains the self-ascribed first-order belief as a part.87

Containment guarantees the existence of the first-order belief, basically by way of a conceptual truth: to believe that one believes that P is, inter alia, to believe that P. If the first-order belief that is guaranteed to be contained in the second-order belief is always a fully-fledged, genuine, authentic belief rather than an anaemic, pseudo-, inauthentic cousin, then containment spells trouble for the idea of inauthentic belief and for authentic internalism.

The containment worry is not the only way of arguing against the possibility of inauthentic belief and thereby undermining authentic internalism. Another is to hold that introspection is an infallible method of determining what one believes, feels, and wants, or at least, introspection is far less fallible than attributing beliefs on the basis of behaviour. We may think introspection is reliable with respect to belief but not other mental states or dispositions such as motivation/motivated-ness. Or we may think it is also reliable with respect to motivation/motivated-ness.88 If the latter, then what someone is motivated to do, or unmotivated to do, is something that the agent herself can know or determine simply by introspection but observers are less well-placed to determine. Whether for belief alone, or belief and motivation, the thought is that publically observable behaviour (actions, speech, gestures, postures, involuntary reactions, and facial expressions) may help outsiders to guess at someone’s mental states (i.e. beliefs, or feelings, or wants, or motivated-ness), but only the agent herself can know the truth for sure, and she does not need to worry about what her behaviour is like to know that. She need only ask herself if she is in the mental state in question or not, look inside herself (as it were), and she will have her answer. As with containment,

88 See footnote 86.
if she believes following introspection that she (genuinely) believes something, then she is correct that she (genuinely) believes it. This would obviously be at odds with conceiving of an inauthentic agent as someone who might not realise she is inauthentic because of self-deception, as I would conceive of her.

Let me respond to the containment worry first, and then the introspection worry. I agree with Shoemaker that in the typical case, if I believe that I believe P, then I believe P. I agree also that this relationship is less secure in cases where we self-ascribe beliefs from “the theoretical stance”. Shoemaker borrows this idea from Richard Moran.\(^89\) Moran’s example is of a woman who, under the guidance of her therapist, comes to accept that she must believe deep down that she has been betrayed by her brother, even though she cannot identify that belief within herself by introspection. Instead, she comes to accept that she must have this belief by looking at her behaviour as an outsider would, and looking for the best interpretation of this behaviour. In doing this, she takes up a theoretical stance towards her own mental life.

One (rather cryptic) way to say what is going on here is this: the woman believes that she believes her brother has betrayed her, but she does not believe that he has betrayed her. If so, this is a counterexample to containment. Perhaps it would be better to say that the woman only half believes the proposition (that her brother has betrayed her), and half disbelieves it, or that she believes it in some sense but disbelieves it in another, or that she believes it subconsciously but not consciously. Whatever way we say it, we have a case where introspection is not especially reliable, a case where what the woman says of her state of mind may be a less revealing and reliable guide that what she does.

Like Moran’s example, authentic internalism, by introducing the category of inauthentic moral belief, fits within the particularly modern tradition of calling into question the genuineness or true nature of people’s professed beliefs. Inauthenticity is a characteristically modern charge, of an age post Freud, Marx, and Nietzsche (as well as

Heidegger, Sartre and Kierkegaard). Paul Ricoeur called Freud, Marx, and Nietzsche
the “three masters of suspicion”.

If we go back to the intention they [Freud, Marx and Nietzsche] had in common,
we find in it the decision to look upon the whole of consciousness primarily as
“false” consciousness. They thereby take up again, each in a different manner, the
problem of the Cartesian doubt, to carry it to the very heart of the Cartesian
stronghold. The philosopher trained in the school of Descartes knows that things
are doubtful, that they are not such as they appear; but he does not doubt that
consciousness is such as it appears to itself; in consciousness, meaning and
consciousness of meaning coincide. Since Marx, Nietzsche, and Freud, this too
has become doubtful. After the doubt about things, we have started to doubt
consciousness.

The inauthentic agent engages in self-deception, repression, projection, excuse making,
self-distracton, all to avoid taking responsibility for her choices. The fact that she
thinks her moral beliefs are genuine should be understood against this background.
Containment does not apply in such cases – for here we can doubt consciousness.

The inauthentic agent often enjoys, in her immaturity, rebelling against what society
demands, perhaps as a way of getting back at an abusive or oppressive parental figure.
The sweetness of this rebellion would be lessened if she decided that society’s take on
morality was simply wrong. What is sweeter is to say to herself that she knows what she
is doing is wrong, but that she just does not care. I say this to encourage us to take up a
hermeneutics of suspicion with regard to examples such as Nick Zangwill’s mercenary:

[A] mercenary I once met on vacation exuded moral indifference. He was in
control, reflective and articulate. Everything he said convinced me that he was

---

91 Ibid., 33. On the same page Ricoeur adds an important caution: “These three masters of suspicion are not to be misunderstood, however, as three masters of scepticism...All three clear the horizon for a more authentic word, for a new reign of Truth, not only by means of a “destructive” critique, but by the invention of an art of interpreting.”
perfectly aware that his vocation was genuinely morally wrong, not merely what
people conventionally call ‘wrong’. He fully understood the wrongness of his vocation. But he was not very concerned about that. He was more concerned with his immediate interests and concerns, that is, colloquially, looking after number one. There was no moral cognitive lack. He made that quite clear. Indeed he insisted on it. The mercenary was unusually indifferent to the demands of morality; but he shared moral beliefs with the rest of us, and with his former self. He insisted on that.92

The mercenary, I would argue, was inauthentic. He wanted to say that “he was perfectly aware that his vocation was genuinely morally wrong, not merely what people conventionally call ‘wrong’” but he plainly does not take the distinction between a genuine moral belief and mere convention seriously. He does not take the distinction seriously because he is avoiding the issue. If he were authentic, he would either be troubled by his profession, or he would stop talking about morality at all – regarding morality as a foreign and useless concept. This was the point made in Chapter 3.

I may already have said enough to make the introspection worry disappear along with the containment worry, but will now say a little more on this worry specifically. I maintain that both authenticity and motivation (and the lack thereof) are revealed in a pattern of agent activity that is public. It may be dangerous to draw conclusions about such matters from a small sample of an agent’s behaviours, including from a small sample of things they say about themselves and the world, but a perceptive and patient person can understand some aspects of an agent’s psychology even better than the agent if the agent is inauthentic and thus engaged in a form of self-deception or repression, or if she is simply very young and thus lacking in self-knowledge and emotional maturity. This is one reason why therapists and counsellors exist. A therapist can lead an inauthentic person to recognize something about herself that she had been avoiding, and one way to do this is by pointing out to her the mismatch between what she says she believes and wants and what she actually does.

In contrast, the view of belief which says that an agent does not need to take into account her own behaviour in determining whether she genuinely believes a proposition or not, is one that meshes well with a Humean account of moral psychology, one which strictly separates the conative from the cognitive.\(^93\) Humeans will feel no need to divide beliefs or believers into authentic and inauthentic, because Humeans have an explanation ready-to-hand to account for any case where someone does not act as one might think she would, given her professed moral belief. The explanation is that the agent does not desire (or is not motivated) to act according to her belief, which we may as well call a genuine belief, since there is no explanatory need to call it otherwise. (I am not saying most Humeans endorse this view. I am simply saying that if one wants to think that beliefs can be fully determined by introspection, it helps to be a Humean.)

However, the Humean view maintains its veneer of plausibility only if one holds before one’s mind neutral belief examples and forgets examples of practical belief.\(^94\) Practical beliefs, recall, are those that essentially pertain to actions and reasons for action. The belief “stealing is immoral” is an example. Neutral beliefs are those that have no special or essential connection with action or reasons for action. The belief “Zirconium atoms contain 40 protons” is an example. When neutral beliefs are what is in question, it does not seem so problematic to think that an agent can know whether or not she genuinely believes the belief or not simply by asking herself what she believes, because for one thing these are not usually cases where there is any motive for self-deception, and for another little hangs on whether the belief is genuine or not, as genuineness/sincerity is not put to any test. The typical school exam tests neutral belief/knowledge. What matters in the exam is that the student’s answer matches the correct answer, not that the student’s belief in that answer is sincere and authentic.

When it comes to matters of practical belief, we generally accept what an agent tells us are her practical beliefs, and only question what she says when it fails to mesh sufficiently with her actions, or other things she says. In effect, we behave as if the agent had the right to establish what it is she believes merely by thinking about it and

\(^{93}\) See comments on Blackburn and Humeanism in Chapter 2, Section VII.

\(^{94}\) See Chapter 2, Section VII.
declaring it. However, the right is rescinded if she shows by her behaviour that she is deceitful, self-deceptive, or mentally deranged. One might call this right a mental right – a right to determine what one’s own mental state is. An agent would have this right with respect to what she believes, what she is feeling, what she wants, what her intentions are, the reason why she performed, is performing, or is going to perform some action, and what she means by what she says. Here is a general characterization of a mental right:

**Definition:** An agent S has the mental right of X iff S is entitled to the benefit of assumption with respect to claims she makes about X-type matters in her mental life (unless C).

\[ X = \text{e.g. intention, belief, desire, meaning, …} \]

\[ C = \text{conditions that, if sufficiently satisfied as a group, would remove or shift the benefit of assumption.} \]

So, for instance, a mental right of belief is an entitlement to the benefit of assumption an agent has when she claims to believe, or not believe, a proposition. If we respect this right, we accept her at her word unless we know of sufficient grounds for doubting her. Similarly, a mental right of desire relates to the benefit of assumption an agent has when she tells us what she wants or does not want, and the other rights relate to her claims to mean or not mean something by what she says, and to act for one reason rather than another.

Using the idea of mental rights is one way to put to rest worries about the danger to first-person privilege posed by authentic internalism. We can say that authentic internalism is perfectly consistent with a belief that the right way to treat each other is as if there were a mental right of belief. This would mean we were assured something like first-person privilege, while still leaving room for the possibility that that privilege might be lost. I will have more to say about mental rights in Chapter 6 Section V.

Like me, Eric Schwitzgebel sees having a belief as having a range of dispositions, including the disposition to self-ascribe the belief. On the reliability of self ascription, Schwitzgebel has this to say:
It is […] an empirical question how reliably the disposition to self-ascribe belief that \( P \) tends to align with the remaining dispositions constitutive of belief that \( P \). Here’s my guess: The relationship will be good to the extent that the belief (or other attitude) meets three conditions: (1) Possession of the belief is normatively neutral, in the sense that having the belief reflects, in the subject’s own view, neither poorly nor well upon her, except insofar as believing truths and not believing falsehoods reflects well on her. (2) The belief is straightforwardly connected to observable behaviour. (3) The dispositions most relevant to the ascription of the belief that \( P \) are inward and outward judgments or statements that \( P \) is the case.\(^{95}\)

If Schwitzgebel is right, and I think he largely is, there is particularly likely to be a mal-alignment between self-ascription of moral belief and dispositions related to actually having the moral belief, because (1) moral beliefs are not normatively neutral, and having a particular moral belief often does reflect poorly or well on the agent who has it, (2) moral beliefs are not straightforwardly connected to observable behaviour (due to, for one thing, the phenomenon of \( akrasia \), and for another, the difficulty of determining how a particular moral belief applies or not to a particular case – e.g. “if I did \( \text{that} \), would \( \text{that} \) be stealing?”), and (3) the dispositions most relevant to the ascription of the belief are not, arguably, the agent’s inward or outward judgments or statements that \( P \) is the case, at least, these are no more relevant than her behavioural dispositions to act according to the belief.

I conclude, therefore, that introspection is no more than a somewhat reliable guide to whether one genuinely, which is to say \textit{authentically}, believes a moral claim, and I might add, introspection is no more than a somewhat reliable guide to whether one is motivated to act by one’s belief. True, a wise agent will generally trust that she knows, unproblematically, what she really believes and what she is motivated to do. However, she should be open to questioning the genuineness of her belief and her motivation if

the pattern of her outward behaviour does not look like that of a person who believes as she thinks she does or like that of a person who is motivated as she thinks she is.

III Inauthenticity versus Akrasia

What is sometimes called akrasia is at other times called weakness of the will or incontinence. Whatever one chooses to call it, it can be defined in the way Donald Davidson does:

An agent’s will is weak if he acts, and acts intentionally, counter to his own best judgement; in such cases we sometimes say he lacks the willpower to do what he knows, or at any rate believes, would, everything considered, be better.96

Though Davidson’s definition is broad enough to include non-moral cases of akrasia, our concern is with moral cases, that is, with cases where the judgment concerns what one ought to do, morally speaking. Moral akrasia bears some similarity to moral inauthenticity, and so the two may be confused, or it may appear that inauthenticity swallows up any conceptual space for akrasia. These are the concerns I wish to address in this section. I will begin by explaining the core similarity between akrasia and inauthenticity, and then make clear the important differences. In this way we will see that akrasia and inauthenticity are related but distinct, and that there is space for both concepts.

The core similarity between being morally akratic and morally inauthentic (and, one might add, being morally hypocritical, a third notion) is that we would not think to label a person with either epithet unless she had expressed her endorsement of a moral judgment and yet acted contrary to that judgment. The same type of evidence might be used to charge someone with akrasia as with inauthenticity or hypocrisy. If Martha said “meat is murder”, but one night we caught her tucking into a Big Mac burger, we have some grounds for thinking either a) she has changed her mind about the morality of

eating meat, b) she is weak-willed, to some extent, c) she is inauthentic, to some extent, or d) she is a blatant liar or hypocrite. To know which of these is true, though, we would need to know more about Martha.

Displays of feeling are also, loosely speaking, a form of action. A display of feeling (or a display of no feeling) that does not fit with an expressed moral judgment might be a sign of akrasia, inauthenticity, or hypocrisy. If Martin said “All races should be treated the same”, but we caught him giggling when an Arab was unfairly targeted by airport security, or if Martin did not seem particularly bothered by a news story in which an Arab was beaten up in the street, we would start to question his sincerity or the authenticity of his belief just as we might question Martha’s.

Despite these similarities, there are differences. The first difference concerns the weight of evidence needed to make the charge stick. Basically, if we can see a person’s failure to act morally on some occasion or occasions against the background of a broader tendency to act morally on other occasions and with respect to other moral judgments, then ceteris paribus we lean towards calling her behaviour akratic (or incontinent) rather than calling her inauthentic, or worse, a hypocrite. However, to the degree that such a background of moral behaviour is absent, then ceteris paribus we lean toward calling her inauthentic (or, in more extreme cases, hypocritical) rather than akratic or incontinent. Weightier evidence is needed to make a charge of inauthenticity stick.

We can draw an analogy (and I stress it is only an analogy) between being morally authentic and understanding a subject matter. If I understand, say, the calculus of differentiation, I should be able to perform very well on tests where I am required to differentiate a range of functions, though I may make the odd mistake. A person who did not understand the calculus should perform much more poorly on this kind of test. If a student scored 95% on a calculus test, and we were asked to judge whether she understood calculus solely on that basis, we would say that she understood calculus, but had made a few mistakes. The 5% worth of mistakes would be seen against the background of the 95% worth of correct answers. However, if a student scored only

---

97 We are less likely to think he has changed his mind here, of course.
10%, we would say that she did not yet really understand calculus, though she might have some inkling of the subject. We would say that the 10% worth of correct answers were probably flukes, because they would be seen against the background of the 90% worth of wrong answers, and these 90% were not really mistakes (at least, not in the same way as for the first person), but much more like failed stabs in the dark. Genuine understanding of a subject is like authentic judgment in its relation to behaviour, and partial understanding like inauthentic judgment.

It is not only a matter of degree that distinguishes akrasia from inauthenticity. An authentic moralist is someone who, amongst other things, is committed to integrating her moral beliefs, her attitudes, her feelings, her aliefs, and her behaviour. This commitment to an integrative project will manifest itself in a number of ways. Instances of immoral behaviour, or inappropriate feelings or reactions, will trouble her. She will tend to feel guilty or unhappy when (or soon after) she acts akratically, she will take steps to reduce the likelihood of behaving akratically again, and we can expect some improvement over time as a result, perhaps not with respect to every type of failure – she may struggle and fail to improve in some areas – but overall. By contrast, the inauthentic person is not committed to an integrative project, and so will not manifest the same sorts of dispositions to the same degree.

Clearly, a decision whether to call someone’s moral belief (or the person herself) authentic often requires sound, mature judgment. No formula exists for such judgments. In difficult cases, I may call behaviour akratic that you would call inauthentic or hypocritical. In deciding what to call a person or a judgment, we would amongst other things be taking a stand on how we think it fit to treat the person. Consider first the akratic agent. The thing that is preventing the akratic agent from living according to morality is that she lacks the way of responding to the world of the virtuous agent, a way that is built on good habits. One solution is to encourage her to develop the right habits, and make it easier for her to do so. In doing this we will be giving her what she wants, since she is committed to the aforementioned integrative project. Sometimes, making it easier may involve expressing our anger toward her or disappointment in her.
Expressing contempt or shunning her, though, would be inappropriate and counter-productive.

Meanwhile, an inauthentic agent is one whose dim awareness of the reality of his existential situation makes him uncomfortable, and so he pretends to himself that he does not really have a choice in how he lives, or that he is only doing what everyone else is doing, or that we can’t seriously expect him to live a nobler kind of life than he does. He is not genuinely apologetic for what he does wrong, and not genuinely motivated to change, yet nonetheless he thinks he has a grasp of what morality is (and in a trivial fashion, he does). Though I do not want to say how he should be treated, it does not seem impossible that it is right to regard him with some measure of contempt, and to deny him sympathy and even the help that, in any case, he does not genuinely want.

IV Authenticity and Idealism

Now that I have shown that authentic internalism does not preclude the possibility of an akrasia distinct from inauthenticity, let us turn to the difficult question of how precisely to word and understand the thesis of authentic internalism. The two versions of authentic internalism detailed in the introduction, one of judgments (A) and the other of agents (B), follow the same schema:

Necessarily, if a person judges that she morally ought to φ, then she is (at least somewhat) motivated to φ if X is authentic.

For version A, X = the agent’s judgment/judging that she morally ought to φ. For version B, X = the agent who makes the judgment herself. In this section I consider five different ways of understanding the claim of authentic internalism, and of choosing between, combining, or understanding versions A and B. These I call aspirational internalism, judgment internalism, deep internalism, scale internalism, and flexible internalism. Flexible internalism is the version or view I favour, but as this is a fuzzy, pluralistic combination of the first four, they will need to be explained first.
Aspirational internalism

Aspirational internalism is the version of internalism one gets if one accepts the stipulations of Chapter 3. That is, according to asprirational internalism,

(i) If a person judges that she morally ought to φ (for some action φ), then she is (at least somewhat) motivated to φ if her judgment is authentic.
(ii) If a person judges that she morally ought to φ, then she is (at least somewhat) motivated to φ if she is authentic.
(iii) One cannot be morally authentic as a person and make an inauthentic moral judgment.
(iv) One cannot make an authentic moral judgment and be morally inauthentic as a person.

I call this version “aspirational” because the authenticity in question is one that few of us, or perhaps none of us, have achieved, and perhaps few or none can achieve, but we can regard it as an ideal we can evaluate ourselves against. The closer we come to the ideal the better, even if we never achieve it. Achieving the ideal would mean being motivated to act according to one’s moral judgments at all relevant times and for each moral judgment. On a particularly idealistic view, this would also mean always acting according to one’s moral judgments – being free from akrasia.

As indicated in the introduction, some may feel that aspirational internalism sets the bar too high. It may be counterproductive to label people or judgments as inauthentic according to this standard. If we set a standard that cannot be reached, people are more likely to give up trying to reach it, and will pursue more realistic moral targets instead, or no moral target at all. People are more likely to feel demotivated if even their best efforts to be a good person are seen merely as less egregious failures to achieve what they ought to achieve and be who they ought to be. It may also, for this reason, seem to be an unfair standard.
An analogous difficulty is found in Christian theology, and has had repercussions in the lives of Christians throughout history. The Christian ideal – the supposed perfection of Christ – is unattainable to those who supposedly inherit Adam’s fallen nature. Though the Christian is forgiven by God for being a sinner, she will continue to sin while on earth. Sanctification, the process of being reformed into the image of Christ, is only completed in the afterlife. As a result, a Christian must continually live with the knowledge that she is failing to live as she ought to live, as Christ would live. She must continuously regret her fallen condition and feel pressure to daily, hourly, or each minute even, ask for forgiveness from God. In addition, since converting to Christianity is supposed to manifest itself in an improved life, when she continues to struggle with sin, as she inevitably will, she may wonder about the genuineness of her salvation. Whether living under this psychological burden helps the Christian to live a better life, or is actually destructive of the self-confidence and motivation required to live a good life, is debatable. At least for some Christians, I suspect the pressure is too high, leading them to indulge in a kind of double-think. On Sunday, they feel something like guilt and a need for God’s forgiveness. For the rest of the week, they think and behave almost as if Sunday never happened.

For these reasons, I have some caution about endorsing aspirational internalism. For a certain kind of person, such as one inclined to neuroticism or someone in a depressive funk, it is an unhelpful view to take. Yet for others it can stand as a pristine and always motivating ideal, and have a positive effect.

** Judgment internalism **

Judgment internalism is less idealistic. According to judgment internalism, authentic internalism A is true: necessarily, if a person judges that she morally ought to φ, then she is (at least somewhat) motivated to φ if her judging so is authentic. However, judgment internalism treats authentic internalism B as false. That is, it is not the case that, if a person judges that she morally ought to φ, necessarily she is (at least somewhat) motivated to φ if she is authentic. One can be an authentic agent and yet still, on occasion, judge inauthentically and be accordingly unmotivated.
What some will see as the advantage of judgment internalism over aspirational internalism, others will see as a disadvantage: it sets the bar of agent authenticity comparatively low. This means an agent can regard herself as authentic as a person (“Basically, I’m OK, and my heart is in the right place”) even though she knows that, in some areas, she is prone to a bit of self-deception or lazy and avoidant thinking. As outside observers we may think this self-assessment is good, or bad. Which it depends on the person (her personality, her age, her intelligence) and the seriousness of the self-deception.

A second concern is that judgment internalism treats authenticity inconsistently. If we are willing to grant that an agent is authentic even if she is sometimes unmotivated to act as she believes she ought, morally speaking, why can’t we grant that a judgment is authentic even if, at times, the agent is sometimes (if only for a short time) unmotivated to act according to it? The problem is that if this is granted we are no longer endorsing an internalist thesis.

I am not willing to back down and endorse externalism, because I have already made the only concession to externalism I wish to make, namely that in some sense an inauthentic agent does believe what she says – the inauthentic sense. True, it may seem arbitrary to permit one kind of hedging with respect to the connection between authenticity and motivation, but not the other. I would counter that sometimes it can make sense and be wise to treat the inauthenticity of agents differently to that of judgments. The problem with judgment internalism is that it always treats them differently, irrespective of the person and the case.

**Deep internalism**

Deep internalism, like aspirational internalism, accepts authentic internalism in both A and B forms, yet is not so idealistic. It achieves this by making it relatively easy for an agent to qualify as motivated to act according to her judgment. In particular, the state of being motivated is understood as a state of being disposed, *in congenial enough*
circumstances, to act according to one’s judgment. The criteria for the existence of this dispositional state are largely behavioural. They are not something the agent can reliably determine by introspection. Being motivated may have no particular phenomenological feel. On an occasion where she knows she ought to do one thing and not another, she may not feel any keenness to do the right thing, she may in fact feel keen to do the wrong thing, and she may feel untroubled by all this, yet she may still qualify as motivated to act morally, because the circumstances are taken to be uncongenial; she has just woken up, she’s in a bad mood, she is tired, doing the right thing would involve a lot of pain or risk, the temptation to do the wrong thing is on this occasion bright and shiny, and so on. The motivation is deep in this sense – it requires congenial circumstances to bring it to phenomenological awareness and to manifest in action, while in uncongenial circumstances it will be present but completely hidden or phenomenologically muted.

Deep internalism makes it relatively easy to qualify as authentic, and to qualify as motivated. The obvious concern is that it makes things too easy. The other concern is that, according to some, the phenomenology of motivation is key to genuine motivation: if one doesn’t feel motivated to do something, one isn’t motivated to do it. If this is right, then deep internalism is not an account of the connection between authentic belief and motivation, but something else.

Scale internalism

Scale internalism treats authenticity (either of judgment, of agent, or of both) as a matter of degree, rather than as an all or nothing property or relation. It also treats motivation/motivated-ness as a matter of degree. Scale internalism asserts roughly the following:

---

98 The scale internalism that I discuss in this chapter must be distinguished from a different sort discussed, and rejected, by Nick Zangwill. In Zangwill’s version, “The degree of a person’s moral belief that he ought to do something proportionately determines the strength of his desire to do it.” See Zangwill, “The Indifference Argument,” 95. In Zangwill’s version it is degrees of belief, rather than degrees of authenticity, that are meant to determine degrees of desire or motivated-ness. Zangwill gives examples (such as the mercenary example I quoted in Section II) to show that degree of belief and degree of desire can vary independently of each other, and thus that this sort of scale internalism is false. I agree that degree of belief and degree of desire can vary independently of each other if the belief is somewhat inauthentic. I
Version A: Necessarily, to the degree that a person’s judgment that she morally ought to φ is authentic, she will be motivated to φ.

Version B: Necessarily, to the degree that a person who judges that she morally ought to φ is authentic, she will be motivated to φ.

An initial difficulty, which can be easily remedied, is that someone may be highly motivated to φ, yet be quite inauthentic in judging that she should φ, if she is motivated by extra-moral reasons to φ. I may, for instance, be highly motivated to be a vegetarian for what I think is a moral reason, but is really just a desire to impress others, or to claim moral superiority over them. The solution is to reword the theses. This can be done in one of two ways, yielding versions C and D, or E and F:

Version C: Necessarily, to the degree that a person’s judgment that she morally ought to φ is authentic, she will be motivated to φ by that judgment.

Version D: Necessarily, to the degree that a person who judges that she morally ought to φ is authentic, she will be motivated to φ by that judgment.

Alternatively,

Version E: Necessarily, if a person judges that she morally ought to φ, then the degree to which she is motivated to φ cannot be less than the degree to which the aforementioned judgment is authentic.

Version F: Necessarily, if a person judges that she morally ought to φ, then the degree to which she is motivated to φ cannot be less than the degree to which the aforementioned person is authentic.

can admit this, and yet maintain that the same is not true of the relationship between degree of authenticity and degree of desire or motivation, because one way to explain what is involved in authenticity is say that authenticity involves, inter alia, a scalar relationship between authenticity of judgment/belief and motivation/desire to act accordingly (as in version A) or something similar but a little more sophisticated (as in versions C and E).
Thus one can be motivated to do the thing one ought (or motivated to avoid doing the thing one ought not) by more than moral judgments, but one cannot be relatively authentic and not motivated at all, or highly authentic but only mildly motivated.

The advantage of scale internalism is that we are not forced into dividing agents and judgments into only two groups, the authentic and the inauthentic. We can speak with a little more subtlety, by speaking of agents as totally authentic, very authentic, relatively authentic, relatively inauthentic, very inauthentic and (if this is possible) totally inauthentic, and likewise totally motivated, very motivated, relatively motivated, relatively unmotivated, very unmotivated and totally unmotivated, to act according to the relevant moral judgment.

The difficulty with scale internalism is that, to be used, it requires a greater degree of psychological perceptiveness than we have in many cases. Its greater subtlety means greater complexity and a consequent increase in the burden of finding grounds for our assessments. If we find it difficult to decide whether someone is authentic or not, how much harder it is to decide whether she is very as opposed to relatively authentic, or relatively inauthentic as opposed to very inauthentic. If we could measure authenticity by some accepted method and give it a value we could resolve such questions, but we do not have any easy or acceptable way to assign authenticity a value. Nor do we have a widely accepted way to assign motivated-ness a value. The danger then of scale internalism is that, if it is employed for all cases, it makes it very difficult to know what to say, or to say anything. And yet sometimes the distinction between different degrees of authenticity or motivated-ness is very useful, and sometimes we have grounds for drawing such distinctions.

**Flexible internalism**

Flexible internalism, my preferred alternative, states that one does not need as a philosopher to choose between the four other forms of authentic internalism. It is a pluralist position. All four forms have their uses and are in that sense “correct”. On
some occasions, it will not matter which version one thinks in terms of – the point is that there is a distinction being marked between authenticity and inauthenticity, and that distinction relates to motivated-ness as opposed to unmotivated-ness. If one needs to be more specific, to explain why one is choosing to call someone inauthentic for instance, one can specify which form of internalism is in play, and if need be, why. However, using one form on one occasion does not bind one to using that version on all occasions. The only requirement is that for a particular case where I want to talk internalist talk I either (a) stick within the penumbra of vagueness of the terms “authentic”, “motivated”, “believe” and so on, so that it does not matter how I understand the internalist thesis in detail, (b) I trust that my audience will be able to infer, given the situation, which version exactly is in play, or (c) I explicitly say something to make clear, on this specific occasion, what I mean and which version is in play.

This may sound awfully complicated or unfamiliar, but it is not. We do exactly the same thing with other very useful adjectives such as “hot” and “impossible” and verbs like “can” and “understand”. Most of the time when we use the word “hot” we trust that our audience will automatically infer, from the context, enough of what is meant to understand us. If I say on Monday “be careful, the coffee is hot”, on Tuesday “the sun is hot, the moon is cold”, on Wednesday “what a hot day we’re having”, and on Thursday “the hot tap needs fixing” I do not mean exactly the same thing by “hot” in each case, and my audience knows this. On each occasion, if it was needed, I could give a rough explanation of the sense of “hot” in play, but you would be a fool to demand that once an explanation had been made, I ought to stick to using the word in that way only. So it is, I believe, with “authentic”.99

The reason why I endorse taking a flexible approach to the authentic internalism thesis should now be plain. It allows us to avoid the problems inherent in the four other forms, yet at the same time benefit from their advantages, when they have advantages.

99 And I might add, given that many do not feel much need for the word “authentic”, the same applies to “virtuous” and “mature”, “honest” and “generous”, “rational” and “moral”.

103
V Conclusion

It is now time to summarise the thrust of the arguments from Chapters 2, 3 and 4, and to suggest in the light of the arguments what I think should happen to the motivational internalism/externalism debate.

In Chapter 2, I introduced the internalism/externalism question, and considered an attempt to answer the question via the distinction between alief and belief. I argued that this answer oversimplified the relationship between belief and other dispositions lower down on the intentionality scale. I suggested that there might be beliefs which, by their nature, involved such lower dispositions. In particular, I claimed this is true of practical beliefs/judgments. I also tried to undermine the usefulness of the cognitivist/non-cognitivist distinction. This distinction forces talk of practical belief to be rephrased as talk of desire and desire-independent neutral belief, but I argued that there is no need to feel forced in this direction, which is at any rate at odds with how we actually talk.

In Chapter 3, I introduced authentic internalism as a way to highlight what I see as the determinant of whether moral beliefs are accompanied by moral feelings, moral desires, and moral motivation: the authenticity of the agent and her belief/judgment. I argued that people who express moral beliefs but lack moral motivation are not genuine believers in what might seem to be their moral beliefs. They are either inauthentic in what they say, and thereby fail to use moral language in the spirit in which it is meant to be used, or they are authentic amoralists who don’t express their own moral beliefs at all. Authentic amoralists mention the beliefs of others, using moral language in an inverted commas sense, but do so consciously rather than in a self-deceptive, suspiciously motivated way.

In this chapter, I dealt with concerns that arise in relation to authentic internalism. I argued that we can have good grounds for calling someone or some judgment inauthentic, and that worries about containment and introspection do not prevent this. I argued that akrasia and inauthenticity are distinct types of defect, distinguished by a complex combination of criteria: the degree to which an agent acts as she says she morally ought to, the degree to which she is reflective, open to criticism, shows remorse for moral failures, expresses commitment to change, and takes steps to improve, and the
degree to which she shows actual signs of moral development or improvement. I also
tackled the question of how exactly to express the thesis of authentic internalism, and
opted for a flexible version, where what we choose to say about an agent’s or a belief’s
authenticity, and how idealist we choose to be, can vary from situation to situation,
depending on the way we think it is helpful to see the situation at the time and the
rhetorical point we want to make.

Now let us suppose a most unlikely thing, that what I have said came to be widely
accepted amongst the philosophers who concern themselves with motivational
internalism and externalism. What would this mean for the future of the debate?

I would hope that the old debate would be transformed into or replaced by three new
areas of focus and debate. Firstly, if I am right that authenticity is the key concept we
need to talk sense on the issue of moral belief and motivation, the nature of authenticity
will need to be explored more carefully than I have been able to within these chapters.
Authenticity was once a highly fashionable notion, but is now out of fashion, so we
need to reacquaint ourselves with it, and reflect on it. Perhaps there are different strains
of authenticity that it would be useful to distinguish. Perhaps there are distinct ways in
which an agent can be inauthentic, or social environments which make authenticity
difficult or impossible. Perhaps authenticity emerges through a process of stages that we
can identify. Perhaps, finally, there are ways in which setting a value on authenticity can
be dangerous, so that caution of a certain sort is needed. All this could usefully be
explored and debated.

Secondly, if we do find the label “inauthentic” useful and employ it more often, we will
need to become more self-conscious, reflective and explicit about how we judge when
this label is appropriate as opposed to labels related to akrasia, hypocrisy and
dishonesty. Because the difference depends on matters of degree, there will be ample
room for debate about where to draw the line in difficult cases. For instance, did Hitler
genuinely believe that the Jews were a moral menace, was this merely a useful belief to
pretend to have in order to achieve power, or was the belief self-deceptive? Different
historians have quite different views on this question.100 The difficulty of deciding what

---

100 See Ron Rosenbaum, *Explaining Hitler: The Search for the Origins of His Evil* (Boston, MA: Da Capo
Press, 2014).
to say here is like the difficulty of deciding a verdict in a court case. Jurors in a criminal case may all agree that they should not vote to convict if there exists reasonable doubt as to the guilt of the defendant. However, jurors can certainly disagree about what constitutes *reasonable* doubt. Criminal cases usually depend upon the prosecution putting forward individual pieces of evidence that, taken alone, are insufficient to remove reasonable doubt, but together may be sufficient. Two jurors can agree that the presence of the defendant’s fingerprints on the murder weapon is relevant, but not conclusive, that his being seen in the building at the time of the murder is relevant, but not conclusive, and that the fact that he bore a grudge against the victim is relevant, but not conclusive, but disagree about whether all three facts taken together are sufficient to make doubt as to his guilt *unreasonable*. The same is true when deciding on authenticity.

Thirdly and finally, if we do find the label “inauthentic” useful and come to use it more often, we will need to discuss the ethical and practical questions about how we ought to treat people who are inauthentic as opposed to akritic or dishonest.
PART B: Reason and Motivation
5 The Nature of Explanatory Reasons

Don’t say: “There must be something in common, or they would not be called ‘games’” – but look and see whether there is anything common to all. –For if you look at them you will not see something that is common to all, but similarities, relationships, and a whole series of them at that.\textsuperscript{101}

— Ludwig Wittgenstein

I Introduction

When someone does something – whether she ties her shoes or joins the Communist Party – we can ask why she did what she did. Philosophers like to say that in asking this question we are asking for an explanatory reason, or perhaps a motivating reason which, let us assume for now, is the same thing. Explanatory reasons are typically contrasted with justifying reasons, and motivating reasons are typically contrasted with normative reasons. These distinctions are routinely marked in philosophy because the question of why she did what she did is (arguably) not the same as the question of what counts in favour of doing it. The latter reasons (normative/justifying) are unaffected by whether or not she actually does the action in question, while the former (explanatory/motivating) require that she did in fact do it – at least, that is how many as philosophers use these terms.

This chapter will focus primarily on explanatory reasons and argue against explanatory reason constitutionalism – the view that all explanatory reasons are the same sort of

\textsuperscript{101} Wittgenstein, \textit{Philosophical Investigations}, §66.
entity, such as facts, or beliefs, or desire-belief pairs, or states of mind, or propositions. Section II will survey the variety of views that exist regarding the constitution of reasons, and Section III will outline a dilemma that one encounters when trying to choose between the competing views. In Section IV, I begin to move away from the thought that reasons have a shared constitution, by looking at the wide variety of things we allow to be called explanations of action, reasons why things happened, and reasons for action. Despite this variety, I accept that there is something which holds cases of acting for a reason together – namely that for any action done for a reason by an agent, there must exist a corresponding desire-belief pair with which we could explain the agent’s action, as Humeans claim. In Section V this claim is defended against two common objections. However, in the following section, Section VI, I undermine Humean as well as other forms of constitutionalism about what explanatory reasons “really” are. I argue that our everyday criteria for determining what counts as a reason or an explanation are pragmatic rather than ontological, so it is a mistake to say, for instance, “the real reasons that explain our actions are desire-belief pairs.” Indeed, this is the point to take away from the whole chapter. The pragmatism described also explains the permissiveness with which we treat “reason” and “explanation” talk. What remains is to return in Section VII to the dilemma of Section II and show that the points made in Sections III-VI dissolve the dilemma. I finish with the briefest conclusion, in Section VIII.

II A Variety of Views

What sort of entities are reasons, and are explanatory/motivating reasons the same type of entities as justifying/normative reasons? Although it is a widely held assumption among contemporary philosophers that we need an account of what reasons are, there is little agreement about what that account should be. Probably the most widely known constitutive account of explanatory/motivating reasons is the Humean account, which states that motivating reasons are composed of desire-belief pairs, where desires and beliefs are understood as mental states of the agent with causal powers. Many philosophers find this account too psychologistic, or have other difficulties with the

Within this anti-Humean group there is broad agreement in opposition to Humeanism but disagreement over the details and terminology of what to replace it with. They variously prefer accounts of reasons or explanations as being constituted by facts, truths, states of affairs, propositions, beliefs (without desires at all), or beliefs without independent desires. Jonathan Dancy opts for reasons as facts/truths rather than propositions, but not as facts about our beliefs and desires except in special cases. Rowland Stout, Ru Bittner, and Frederick Stoutland agree. Joseph Raz, Bernard Williams, Stephen Darwall, and John Skorupski are united in claiming that one’s reasons are the things one believes rather than one’s beliefs. That is, my reason for leaping out the window was that my house was on fire, not that I believed my house was on fire. While many see reasons as facts, they may mean different things by this. Some use “fact” to mean what a proposition states, which Stephen Everson calls a Strawsonian view, rather than what a proposition is about (i.e. a state of affairs), which he calls an Austinian view.

Whatever one’s view of explanatory/motivating reasons, there is the added question of whether normative/justifying reasons have the same constitution as explanatory/motivating reasons. Humeans regard motivating reasons as desire-belief pairs (or facts about desire-belief pairs rather than the pairs themselves). They cannot

---

104 In writing “reasons as...beliefs without independent desires,” I point to the sort of position one could attribute to Thomas Nagel, or that could be inspired by reading him, where a desire may always be required to explain an action, but need not be what Nagel calls an “unmotivated desire”. The most that is required to explain the action is that the agent has a motivated desire to perform it (or to pursue a related goal), a desire which is explained by whatever explains the agent’s “pursuit”. This explanation need not involve an unmotivated desire. See Thomas Nagel, *The Possibility of Altruism* (Princeton, NJ: Princeton Paperback, 1978), 29-30.
regard normative reasons as desire-belief pairs as well, because even Humeans agree that it is possible for an agent to have a reason to do something that they do not realise they have because they lack the requisite belief. If belief were required for a normative reason to exist, we couldn’t explain such cases. However, most Humeans think normative reasons are based in the agent’s desires, such that absence of a relevant desire (unlike lack of belief) entails absence of a normative reason to act accordingly, while the presence of a relevant desire entails the existence of a reason to act accordingly.\(^{109}\)

One conclusion that Humeans might draw from this is that genuine normative reasons simply are desires (either, the actual desires the agent has when the normative reason is said to exist, or the desires she would have if she were fully informed at that time). This is the closest that a Humean can get to holding that motivating and normative reasons have the same constitution – they are both constituted, either fully or partly, by desires. Meanwhile, the majority of non-Humeans also hold that motivating and normative reasons share a constitution, but not a constitution in terms of desire-belief pairs or desires, rather a constitution in terms of considerations that count in favour of an action. The third possibility is to deny a shared constitution for motivating and normative reasons. For instance, Michael Smith views motivating reasons as desire-belief pairs but normative reasons as truths – truths which, to his mind, are also propositions.\(^{110}\)

There are untold complexities here. Even if every philosopher were to accept that my reasons are or must contain beliefs, this could mean one of many things – it could mean that each of my reasons is (A) an act or state of believing, (B) the fact that I believe what I do (as opposed to the fact which I believe), or (C) what I believe, rather than the act of belief or fact of belief. As for this last, what is believed may be variously characterized as (C\(_1\)) the proposition that I believe, (C\(_2\)) the state of affairs that I believe to obtain, (C\(_3\)) the fact that I believe to be the case (and what indeed are facts?), or (C\(_4\)) the consideration that I believe to apply. Similar options to (A), (B) and (C) exist for accounts in terms of desires. Humeans can agree that desires are always part of my reasons for what I do but disagree over whether my reason for, say, φ-ing (for some

\(^{109}\) According to Mark Schroeder, any Humean about reasons (like himself) is committed to believing that: If doing \(A\) promotes what some agent desires, then there is a reason for that agent to do \(A\). See Schroeder, Slaves of the Passions (Oxford and New York: Oxford University Press, 2007), 88.

\(^{110}\) Smith, The Moral Problem, 95. Darwall shares a similar view, where reasons are propositional, and expressible with a “that” clause. See his Impartial Reason, 31.
action $\phi$\textsuperscript{111} is (α) my act or state of desiring to $\psi$ (for some action $\psi$), (β) the fact that I desire to $\psi$, or (γ) what I desire, namely (to) $\psi$\textsuperscript{112}

III The Personal/Impersonal Dilemma

While there is no agreement over what constitutes reasons for action, from mundane actions such as turning on the light and drinking a glass of water to more important actions such as voting, marrying, and murdering, it is in relation to morality that the debate becomes particularly problematic, most especially in relation to whether desires are a necessary part of the constitution of the reasons why we do what we do, and what this tells us about moral reasons. Here I will lay out what is problematic by presenting a dilemma.

As philosophers we seem to have two decisions to make. First, we must decide whether we think explanatory/motivating reasons are the same type of entity as justifying/normative reasons. Then, whatever we decide on the first question, we must decide what type or types of thing we think they are. Suppose, in answer to the first question, we feel that both types of reason are the same type of entity. We can now be guided in answering the second question either by what we take an explanatory/motivating reason to be or by what we take a justifying/normative reason to be, and then use the nature of the one to determine the nature of the other.

If we are guided by explanatory/motivating reasons cases, we will probably be inclined to think of reasons as desire-belief pairs, desires alone, or something similar. However, if we do so, and transfer this insight over so that justifying/normative reasons are also seen as desire-belief pairs, or desires, or something similar, we are forced to say some rather odd things about moral normative reasons. We may find ourselves forced into accepting that what we happen to desire marks the limits of what morality can demand (what moral normative reasons there are), which seems to distort the nature of morality.

\textsuperscript{111} Throughout this chapter, “$\phi$” and “$\psi$” refer to actions that are or might be performed by an agent, even if this is not explicitly stated.

\textsuperscript{112} Maria Alvarez discusses the act/object ambiguity in talk of belief and desire in Kinds of Reasons: An Essay in the Philosophy of Action (Oxford: Oxford University Press, 2010).
Or, if we admit that morality demands things that go beyond what can be accounted for on the basis of our desires, we are forced into saying that sometimes we have more reason to act immorally than morally and that morality is only reasonable for those with the right motivational makeup in the right circumstances. For many of us, neither position is acceptable. This is a problem. Either morality has become something too personal, too tied to individual psychology, or what we have reason to do has become too personal, too tied to individual psychology, and too disconnected from morality. Call this, then, the overly-personal problem.

Alternatively, while continuing to think that normative/justifying reasons and explanatory/motivating reasons share a constitution, we can be guided by normative/justifying cases. If we take this line, we are likely to see reasons as consisting of the facts, considerations, or states of affairs which should, but perhaps don’t, guide our actions. If we then transfer this insight over so that explanatory/motivating reasons are also seen as facts (or the like), we run into new problems. If what explains what we do is a fact, and not a desire-belief pair, then the presence or absence of any desire in the agent must be ultimately irrelevant to the action and its explanation – we simply act as we do because the fact is a fact, not because there is something that attracts us to acting in this way. If we take this line, acting becomes something mysterious – for how is it that a mere fact, or even belief in a mere fact, can lead us to act without our needing to have any motivation? Our actions are made to seem mechanical and involuntary, or if accepted as voluntary, they seem dispassionate and cold-hearted. Neither option is attractive, especially when characterizing what it is to act morally.113 This is a problem. Moral action – indeed all action – has become something too impersonal, too detached from normal and healthy individual psychology. Call this, then, the overly-impersonal problem.

Given this dilemma, it would seem then that we must reverse our earlier decision to see reasons as sharing a constitution, and hold that the constitution of explanatory/motivating reasons is not the same as the constitution of

---

113 The second option is certainly not attractive to sentimentalist philosophers, or Aristotelians, or anyone who believes authentic moral behaviour to be an expression of Buddhist compassion or Christian love.
justifying/normative reasons. So, we might say for instance, explanatory/motivating reasons are desire-belief pairs, while justifying/normative reasons are facts (or beliefs in facts, or beliefs in putative facts). However, if we take this line we run into a final problem – one identified by Jonathan Dancy amongst others. We want to be able to say that we sometimes act for good reasons. So suppose I perform some action \( \phi \). When I \( \phi \) for a good reason that means that the reason why I \( \phi \)-ed (the explanatory/motivating reason) is the same as the reason that there was for \( \phi \)-ing (the justifying/normative reason). As Dancy puts it:

[A] motivating reason, that in the light of which one acts, must be the sort of thing that is capable of being among the reasons in favour of so acting; it must, in this sense, be possible to act for a good reason.\(^{114}\)

Unfortunately, if explanatory/motivating reasons are one sort of thing (desire-belief pairs), and justifying/normative reasons are another (facts, or the like), it seems we cannot make sense anymore of the idea that someone acted for a good reason. This is a problem, which we can call the problem of acting for good reasons.

Defenders of different accounts of the nature of explanatory reasons will have their own solutions to the problem or problems that apply to their style of account.\(^{115}\) However, it is not my purpose in this chapter to evaluate whether these solutions work. Rather, I wish to argue that the quest for a constitutive account of the nature of reasons for action and explanations of action is a mistake. Certainly, we can give a stipulative definition of a certain type of reason or explanation, grouping the reasons by what constitutes them,

---


\(^{115}\) Michael Smith for instance, accepts a Humean account of motivating reasons as being constituted by desire-belief pairs (Smith, *The Moral Problem*, 92), and normative reasons as being constituted by truths (Ibid., 95) but claims that normative truths (such as that \( \phi \)-ing is the most rational thing to do) will cause a rational agent who believes it to have the related desire (the desire to \( \phi \)), and thus the normative truth can play a part in explaining what the agent does when she acts on this belief, but – in Smith’s terminology at least – the normative fact which creates the desire is not properly called a motivating reason. That term is reserved for the desire-belief pair that causes the action. In this way Smith may hope to escape the overly-personal problem and the problem of acting for good reasons.
but that is not – I think – what the philosophers I am concerned with see themselves as doing. If it were, there would be no debate.

What I wish to show is that it would be better to give up trying to articulate what reasons are – e.g. *reasons are beliefs*, or *reasons are facts*, or *reasons are desire-belief pairs* – and further, if we give up trying to articulating what reasons are we can avoid the overly-personal problem, the overly-impersonal problem, and the problem of acting for good reasons in a simple and attractive way.

IV Permissiveness

If we want to get escape the dilemma of Section III, the place to begin is by noticing our relative permissiveness when it comes to the usage of the terms “explanation” and “reason” (as well as “because” and “cause”, though I will say little directly about these two in this chapter). In this section I will try to map out some of the major permissible uses of these terms in relation to explaining things that have happened. I will begin with a very broad class of reasons and explanations and then move into narrower classes. I will not say why we are so permissive here, but I will try to account for it in Section VI.

Something happened: a volcano erupted, a man fell to his death, a child broke a window, a woman quit her job. One of the broadest uses of “explanation” and “reason” is when they are applied to cases where something, anything, happened. Of course we can also explain what is happening, will happen, might happen, or should happen, or why some non-event type fact is true (e.g. why \(\sin^2 x + \cos^2 x = 1\)), but these will be left aside for simplicity sake. We will focus only on things that happened. The reason why (or *reason that*, or *explanation why*)\(^{116}\) the volcano erupted is that pressure had built up beneath its surface. The reason why the man fell to his death is (or was) he slipped on the wet roof.

\(^{116}\) The Oxford Dictionary Online notes that some regard “the reason why” as objectionable due to the redundancy involved in the phrase. However, it argues that the phrase is “well established” and “generally accepted in standard English” even if “inelegant”. See “Reason,” Oxford University Press, accessed May 3, 2015, http://www.oxforddictionaries.com/definition/english/reason. Similar comments would apply to the somewhat less established phrase “explanation why”. We tend to prefer sentences and questions containing “explanation for” to those containing “explanation why”, just as many prefer those containing “reason that” or “reason for” to “reason why”.
The reason why the child broke the window is she was carelessly hitting a ball around. The reason why the woman quit her job is she was unhappy with her salary. This is the first type of permissiveness – we allow talk in terms of “explanation” and “reason” in all these cases. Call the reasons that are given here reasons-why-something-happened, or H-type reasons or H-type explanations for short.117

We allow “reason” and “explanation” to be used for any case where something happened, not just for cases where an agent used reasoning to determine what to do or might want to explain why she did what she did. It is we, the observers of the thing that happened, who reason out and explain why it happened. Thus even when reasoning participants are not involved in the event being explained, we can speak of “reason”.

A second type of permissiveness is that we allow different kinds of H-type reasons to count as the reason why a particular event happened, without contradiction – we can say that the reason why the man fell is because the roof was slippery, and he fell because he slipped, and he fell because he wasn’t paying enough attention to his footing, and he fell because he wanted to finish fixing the roof as quickly as possible, and he fell because he’s a fool. The first is a fact about the roof. The second is a fact about something the man’s body did. The third is a fact about something he didn’t do. The fourth is a fact about what he wanted. The last is a fact about his character or intelligence. All are, in the right circumstances, acceptable and compatible H-type reasons.

Within H-type reasons, there is a subclass of reasons: reasons-why-people-do-things. Call these D-type reasons. There is a broader and a narrower class of D-type reasons. Broad D-type reasons explain intentional actions people perform as well as things that people’s bodies simply do or cause, while narrow D-type reasons explain only the

---

117 I am purposely using this non-idiomatic, rather ugly expression so the reader is unable to unconsciously import any preconceptions about what is meant by H-type reason from her prior understanding of a familiar qualifying term. I want the reader to be forced to return to the initial definition: H-type reasons are reasons why something happened. This is why I do not speak, say, of event reasons. Similarly, I prefer to talk of D-type reasons and F-type reasons than action reasons, motivating reasons, agential reasons, or the like. For an example of where the use of more natural expressions creates problems, consider the various ways in which “motivating reason” is currently understood, and see the comments at the end of Section IV.
"He fell because he slipped" and "she broke the window because she was
carelessly hitting a ball around" are broad but not narrow D-type reasons, because the
slipping and breaking are (we assume) unintentional things that the body does or causes.
Of course, climbing on the roof and hitting a ball are intentional actions, but the things
that are being explained are not. By contrast, saying “she quit her job because she was
unhappy with her salary” is giving a narrow D-type reason, since when she quit her job,
she was intentionally quitting her job. Again, we are quite permissive when it comes to
what we allow to count as narrow D-type reasons: situational facts, beliefs, wants,
character traits, goals, facts about one’s upbringing, and habits can all be used to explain
why people intentionally do things.

A third type of permissiveness that is true of reasons in general and narrow D-type
reasons in particular is lexical/grammatical. There are a variety of ways in which to
finish a sentence such as “She quit her job...” in order to give a narrow D-type reason.
We can say:

She quit her job...
...because she wasn’t happy with her salary. (because + a clause)
...to get a better salary. (an infinitive)
...because of her poor salary. (because of + a noun phrase)

We also allow, for instance:

She quit her job...
...in order to get a better salary.
...because of the fact that she wasn’t happy with her salary.
...due to her poor salary.
...for the sake of her income.

It should be noted that whether an action is intentional or not depends on how it is described. Under
one description, it may be intentional, under another it may not. However, we can still distinguish actions
that are intentional under some description from “actions” that are not intentional under any description. In
a case if the former type, the person involved is the agent of the action, and in a case of the latter she is
not. As Davidson says, “a man is the agent of an act if what he does can be described under an aspect
that makes it intentional.” See Donald Davidson, “Agency (1971).” in Essays on Actions and Events
…out of dissatisfaction with her salary.

Any of the above ways of ending the sentence is a way of giving a narrow D-type explanation of what she did.

The distinction between narrow and broad D-type reasons, and between D-type reasons and non D-type reasons within H-type reasons, is a distinction between the types of things that are being explained: intentional actions as opposed to unintentional actions, things done by people as opposed to things that are not done by people. However, we can narrow our focus even further by making a distinction not in what is being explained, but what kind of information is being offered as an explanation. Within narrow D-type reasons/explanations, there is the subclass of reasons-for-doing-things, or F-type reasons, with a broad and narrow sense. In the broad sense, an F-type reason explains by making clear either directly or comparatively indirectly what it was about doing the action that appealed to the agent – for instance, that she quit her job because she was unhappy with the salary or because she wanted a better paying job, or because the salary was low. The narrow F-type reasons are those where the content of the explanation is straightforwardly the consideration that the agent saw as counting in favour of acting, not some other thing from which we can infer such a consideration. I am stipulating that both facts and beliefs are to qualify as considerations in this formulation. In the case where the agent acts on a mistaken belief, that belief can count as a narrow F-type reason, and is obviously not a fact (though the fact that she believed it is). In the case where her belief is correct, one can say the narrow F-type reason is her belief or the fact which she believes. The consideration may even be a fact which does not count in favour of the action, but which the agent believes counts in favour of the action. Not all actions that admit of narrow D-type explanation also admit of narrow F-type explanation. The group that do not are the actions that are performed intentionally, but not for any reason: the arational actions.\textsuperscript{119} An example is when someone jumps up and down from excitement.

\textsuperscript{119} A strong argument has been made that there are cases of intentional action which are arational (and thus not done for any reason) – cases for instance where the agent’s action expresses her feeling but is not done in order to make something else happen. See Rosalind Hursthouse, “Arational Actions,” \textit{Journal of Philosophy} 88, no. 2 (1999): 57-68.
So let us suppose a woman quit her job because she was unhappy with the salary, in the normal way. The fact/belief that her salary was low would be a narrow F-type reason for her quitting, as would be the fact/belief that there are other jobs with higher salaries available. However, the fact that she wanted a better salary would be a broad, not a narrow, F-type reason. It implies, rather than simply states, the consideration that motivates her.

One question that arises at this point is whether the fact that the agent has a desire could be not just a broad F-type reason but also a narrow F-type reason. If so, when we say “His reason for φ-ing was the fact that he wanted to ψ”, we directly state the consideration that motivated him, rather than a fact from which such a consideration can be inferred. There are indeed examples of this, but they are atypical. A typical case will not do. “His reason for eating the snails was the fact that he wanted to enjoy them” gives a broad but not narrow F-type reason. The narrow reason will be something like this: “His reason for eating the snails was that (he believed) snails are enjoyable to eat.” Of course, from this statement we can infer that he had a desire, but this statement is not a statement of his desire. Nor will a case of following a mere urge do. If I play around with a loose thread on my jacket just because I feel like doing so and for no other reason, the fact that I feel like doing so is not my reason for doing so – I have no reason for doing so. Of course, if I enjoy doing it, or it pains me not to do it, then these facts can be my reason for doing it, but the narrow F-type considerations will not be claims about my desires, but about (what I think are) the pleasures of playing with a thread, or the pains that (I think) can be avoided by playing with a thread.

A case where the fact that an agent has a desire is his narrow F-type reason will generally be one where he does not reflectively endorse the desire. For instance: “His reason for visiting the psychiatrist was the fact that he desired to go on a killing spree.” In these atypical cases, the action explained is done to battle, escape, reduce, or limit the damage that might be caused by the desire detailed in the explanation.
Finally, we can add one more layer to our taxonomy. We can define T-type reasons as those narrow F-type reasons that are facts or true beliefs as opposed to false beliefs. This completes the taxonomy (see figure 1).

From what has been said, it should be clear that each reason type is a subclass of the previous type, but no previous type is a subclass of the next. All T-type reasons are narrow F-type reasons, all narrow F-type reasons are broad F-type reasons, all broad F-type reasons are narrow D-type reasons, all narrow D-type reasons are broad D-type reasons, and all broad D-type reasons are H-type reasons, but not vice versa.

Figure 1
We should note that though we allow “explanation” and “reason” to be used very broadly, there are limits to the acceptable use of some phrases that include “reason”. One example is the limits to the use of “reason for”, as in “her reason for φ-ing was that…” When explaining why something happened, the phrase “reason for” needs to be matched with a phrase giving an F-type reason. Thus, although one can say “Her reason for quitting her job was that the pay was low” one cannot say “her reason for breaking the window was she was carelessly hitting a ball around” or “the reason for the volcano erupting was there was a blockage in the crater.” In these cases we had better speak of “reason (why)” rather than “reason for”. A similar limitation holds true of “a reason to”, “his reason to”, “the reason in favour of”, “the reason in support of”, and “by reason of”, but not “for some reason”, “reason why”, or “reason that”. The presence of a blockage in a crater is not a reason to erupt, nor is it the volcano’s reason to erupt (as low pay can be someone’s reason to quit), but a volcano can erupt for some reason. The first group of phrases must be matched to F-type reasons. The second group need not, and often cannot, be matched in this way.

Having given a quick run through of the different types of reason, we can diagnose one source of problems in the philosophy of the explanation of action. Suppose I offer a theory of the nature of action explanations. The fact that it is action to be explained means I must at least be restricting the scope of my theory to D-type reasons within H-type reasons, and probably to narrow D-type reasons. However, I had better know and make explicit whether or not what I am offering is simply this, or whether it is a narrower claim about F-type reasons (broad or narrow) within D-type reasons, or a claim about some other sort of reason/explanation. If I am not clear about this, unnecessary confusion will result. I must also be careful to say whether I am truly theorising about the nature of some class of reasons, or simply giving a stipulative definition of some class of reasons. Again, if I am not clear, confusion will result.

120 One does sometimes hear sentences like “The reason for the eruption was a blockage in the crater.” Given that we know volcanoes lack intentionality, there is little danger we will misunderstand a misplaced “reason for” here. Still, I think this is a potentially corrupting turn of phrase – accepting and engaging in this way of talking may lead us astray when we come to consider cases that do involve intentionality. Of course, we never hear “The volcano’s reason for erupting was that it was blocked.”
These dangers are especially apparent when we take one of the most popular terms in current debate, *motivating reason*, and try to find a place for it in the taxonomy I have given. Non-Humeans usually take motivating reasons to be something like what I have called narrow F-type reasons, although they may not be so liberal as to allow both beliefs and facts to be motivating reasons.\(^\text{121}\) Meanwhile Humeans see motivating reasons as desire-belief pairs.\(^\text{122}\) Typically, neither Humeans nor non-Humeans are clear enough about whether they are (A) stipulating what they will take *motivating reason* to mean, (B) following someone else’s prior stipulation, (C) arguing for a substantive philosophical conclusion about the nature of motivating reasons characterized in some neutral way, or (D) expanding on the (supposedly) obvious meaning one gets when the two common English words “motivating” and “reason” are placed side by side.

For instance, this is how Smith introduces the notion of a motivating reason in his book *The Moral Problem* (and, almost identically, in the earlier article “The Humean Theory of Motivation”):

> The Humeans in fact seem committed to two claims about motivating reasons, a weaker and a stronger….According to the stronger claim – the claim that is, as I understand it, crucial to the Humean theory – motivation has its source in the presence of a relevant desire and means-end belief. The claim finds more formal expression in the following principle….: P1: \( R \) at \( t \) constitutes a motivating reason of agent \( A \) to \( \varphi \) iff there is some \( \psi \) such that \( R \) at \( t \) consists of a desire of \( A \) to \( \psi \) and a belief that were he to \( \varphi \) he \( \psi \).\(^{123}\)

One interpretation of this passage is that Smith is defining *motivating reason* as whatever the source of motivation is, and then going on to make a substantive

\(^{121}\) Dancy, *Practical Reality*, 1-6; Maria Alvarez, *Kinds of Reasons*, 34-36. Derek Parfit is another key philosopher who talks of “motivating reasons”, but it is hard to judge whether he holds the same view as Dancy and Alvarez. Parfit states, for instance, that “while motivating reasons require that we have some belief, whether or not this belief is true, normative reasons are provided by some truth, whether or not we believe it.” Parfit may believe motivating reasons are (partially) constituted by beliefs, or only that motivating reasons require beliefs. See Derek Parfit and John Broome, “Reasons and Motivation,” *Proceedings of the Aristotelian Society, Supplementary Volumes* 71 (1997): 113.

\(^{122}\) See footnote 102.

philosophical claim that the source of motivation is (the presence of) a desire and means-end belief, rather than something else, such as a normative fact. A different interpretation is that he is defining a motivating reason as (the presence of) a relevant desire and a means-end belief, and then making the substantive philosophical claim that it is these desire-belief pairs that are the source of motivation. A third possibility is that the whole thing is stipulative: motivating reasons are being defined as desire-belief pairs which, in Smith’s stipulation if not common parlance, we can also call the source of motivation. I suspect that Smith would maintain that his theory should be understood as conflicting with later non-Humean accounts of motivating reasons, such as Dancy’s. If so, Smith was rejecting the claim that considerations – what I call F-type reasons – are the source of motivation, and claiming that desire-belief pairs are the source. That is, I incline towards the first or second interpretations. Even here, though, it is not clear what to take “source” to mean. Certainly, a broad sense of “source” (in “source of motivation”) has room for what I call broad F-type reasons. If a more restricted sense of “source” is in play, such as a narrowly causal one, the Humean and non-Humean will only be talking past each other.

Because there does not seem to be a consensus on the meaning of “motivating reason”, and because I am not sure what exactly is being asserted by Smith, I will hereafter avoid the term. However, it will be useful to have a word for the type of thing Smith finds motivating reasons to be – the desire-belief pairs that are appropriately related to the performing of an action φ. I will simply call this pair a Humean pair for φ, or HP(φ), defined by modifying P1 as follows:

Definition: R at t constitutes a Humean Pair of agent A to φ (or HP(φ)) iff there is some ψ such that R at t consists of a desire of A to ψ and a belief that were he to φ he would ψ.

Or, more elegantly,

Definition: A Humean Pair of agent A to φ (or HP(φ)) consists of a desire of A to ψ and a belief that were he to φ he would ψ, for some ψ.
One of the consequences of the permissiveness I have detailed in this chapter is that, if there exists some Humean pair $HP(\phi)$ related as the definition states to the action $\phi$ that was performed, then we can say that $HP(\phi)$ is both an explanation of the agent’s $\phi$-ing and her reason for $\phi$-ing. In Section V, I will argue that there always is such a pair for any intentional action, but in Section VI I will argue that one cannot make the stronger claim that the Humean pair constitutes the real, true, primary source of motivation and explanation of intentional action while other so-called reasons (e.g. claims about facts, goals, character) are merely coincident, ancillary, or secondary to the true source of motivation or true explanation, i.e. Humean pairs. It is on this last possibility that the dispute over the nature of reasons must hinge if there is to be a genuine dispute rather than a mere difference in usage.

**V Ellipticality Defended**

We will now examine some of the necessary connections between the permissible reason-statements outlined in Section IV. The thought in this section, roughly speaking, is that we are as lenient as we are in what we let be called a reason or an explanation in narrow D-type cases because we take it that when someone offers an explanation of a particular sort for an action, she would also accept the truth of other claims that could be used as explanations of the action by others. That is, we do not require explainers to express all permissible explanations of an action when they explain an action, but we do require that they would endorse certain truths about the case at hand without which their explanation would be faulty. These other truths are not explicitly stated, and in that sense we can call most if not all explanations elliptical.

We are familiar with this type of ellipticality from the discussion of Humean theories of motivation. For instance, Nikolaj Nottelmann argues that a Humean is committed to two elliptical theses ER and ET:
Elliptical Rationalization (ER): In so far as an action explanation purports to explain an action via its internal rationalization, it is at best elliptical, unless it includes reference to discrete states of belief as well as desire.

Elliptical Teleological Explanation (ET): In so far as an action explanation purports to explain an action via the agent’s aims, it is at best elliptical, unless it includes reference to discrete states of belief as well as desire.\(^{124}\)

Let us see what this means by considering an example, first in relation to ER, then ET. Suppose Quorra quits her job as a teacher, and we are interested in her “internal rationalization” for quitting.\(^{125}\) Suppose we say that the explanation of (or reason for) her quitting was that she believed other jobs are better paid than teaching. If we are correct in saying this, then according to ER it must be the case, and we must believe, that Quorra wanted a better paid job and also that she believed that by quitting it would be more likely that she would have a better paid job, even though we did not mention these things in giving the explanation. If either one of these were not true, our explanation would not be correct. In that sense, our explanation in terms of her belief that other jobs are better paid than teaching is elliptical, as ER states.

Similarly, consider Quorra’s case in relation to ET. Suppose we say that Quorra’s aim was to get a better paying job, and that is the explanation for her quitting. If we are correct in saying this, then according to ET it must be the case, and we must believe, that Quorra wanted a better paid job and also that she believed that by quitting it would be more likely that she would have a better paid job, even though we did not mention these things. If either one of these were not true, our explanation would not be correct. In that sense, our explanation in terms of her aim of getting a better paid is elliptical, as ET states.


\(^{125}\) I take it that explaining an action via its internal rationalization is basically the same as giving a broad F-type reason for her action. Recall that F-type reasons were either considerations that counted in favour of acting, or explanations from which such considerations could be straightforwardly inferred, such as a statement of what the agent wanted.
ER and ET are related, but not identical, to Smith’s Humean principle P2, a principle weaker than P1, which was discussed in Section IV. Here is P2:

Agent A at t has a motivating reason to φ only if there is some ψ such that, at t, A desires to ψ and believes that were he to φ he would ψ. \[126]\n
I neither endorse nor reject P2, because I am not sure how Smith is defining a motivating reason. However, I can endorse something similar to P2 which is related to ER and ET, using the idea of the Humean pair for φ:

Agent A φ-ed for a narrow F-type reason only if there existed an HP(φ). \[127]\n
We can call this the Humean existence thesis, HE. \[128]\n
Although there are a number of objections which can be made against the ER, ET, and HE, I think with a sufficiently broad notion of “desire” and “belief” and a neutral reading of “elliptical”, these theses are correct. A broad enough notion of “desire” will be one that does not require an agent to feel any particular emotions or yearnings related to φ-ing prior to acting in order to say that at the time he genuinely had a to desire to φ, and a broad enough notion of “belief” will be one that does not require the agent to actually think the thought “φ-ing is a way of ψ-ing” (or something similar) prior to

---

\[126\] Smith, “The Humean,” 36; The Moral Problem, 93.

\[127\] In an early draft of this chapter, I had a different version of this thesis: “Agent A φ-ed intentionally only if there existed an HP(φ).” However, this version of the thesis is untrue if what are known as arational actions exist. See footnote 119. There does not exist either an F-type reason or a Humean pair for arational actions.

\[128\] The Humean Existence thesis is similar to Neil Sinhababu’s desire belief theory of action, DBTA: “Desire is necessary for action, and no mental states other than a desire and a means-end belief are necessary for action.” However, as I make clear at the end of section V, I believe the mental state of intention/intending is also required on top of a desire and a means-end belief. See Neil Sinhababu, “The Humean Theory of Motivation Reformulated and Defended,” Philosophical Review 118(4) (2009): 465.
acting in order to say that at the time he genuinely had a belief that by φ-ing he could ψ (or something similar). Michael Smith’s account of desires and beliefs as dispositions is one plausible way to satisfy this condition.\textsuperscript{129} A neutral reading of “elliptical” will be one that does not entail that an explanation’s being (in some sense) elliptical is a defect, nor that this explanation is inferior to some superior, imagined, non-elliptical explanation. For this reason, the one thing I dislike about Nottelmann’s formulation of ER and ET is the inclusion of the phrase “at best”, which suggests the implications I wish to avoid. The thought that Humean explanations are superior explanations will be explored (and rejected) in Section VI.

Nottelmann identifies two common objections to the elliptical theses, both of which I believe fail to undermine them. There is only space in this chapter to briefly run through the objections and the responses that I think are available to a Humean, but this will suffice since my true purpose is to show that even if ER and ET are true, this does not mean that desire-belief pairs are what constitute reasons for action, nor that we are stuck with the overly-personal problem. If I am wrong and one or both of ER and ET is in fact false, then this only makes arguing the work in Sections VI and VII easier.

The first objection Nottelmann discusses is what I shall call the false-belief objection.

\begin{quote}
Sometimes, an anti-Humean might claim, we need not invoke the agents’ desires in order to explain an action. It suffices that we invoke her false beliefs about her desires. Imagine, e.g. a student whose upbringing has induced in her a strong belief that she desires to pursue a career in art, and also believes that by going to art college, she will promote satisfaction of this desire. However, at the bottom of her heart, really she hates the idea of being an artist and desires no such career. This situation seems psychologically possible. Now, in so far as the student actually attends art college in a diligent manner, is not the best explanation of
\end{quote}

this that her false belief about her desires has taken over the action-guiding role, which, according to the Humean, only a desire could fill?\textsuperscript{130}

How can the Humean respond? Here is my take. She can begin by looking further into what is involved in the back-story, wherein “[the student’s] upbringing has induced in her a strong belief that she desires to pursue a career in art.” How could an upbringing induce such a belief? Presumably, what happened was one of two things, depending on the character of the child. Let us contrast a critically minded child Agatha with her uncritically minded sister, Bertha. Agatha and Bertha’s parents recommended art college for both of them, and both girls – let us agree – came to believe that they desired to go to art college. Agatha, being critically minded, had weighed what her parents said and found it convincing, and gone on to attend art college. In this case, I think we would say that she did have a desire to become an artist, and that’s why she attended the college. Of course, once at the college, when she found out more about what studying to be an artist actually involved, Agatha could decide that she hated the idea of being an artist, but this would only mean that she had lost her former desire, not that she had never had the desire in the first place. Thus we still have a Humean explanation of Agatha’s action, one that includes her former desire to become an artist, a desire that she correctly believed herself to have had at the time.

Now consider uncritical Bertha, the sort of child who is overly keen to please her parents, and who accepts quite uncritically what they say. Bertha was disposed to do what her parents recommended and believe what they told her, so when her parents told her that she should pursue a career in art because an artist’s career is a desirable one, that was enough for her. Bertha, acting rather blindly, may have had absolutely no idea what art was beyond the name or why art was worth pursuing – all she knew was that whatever art was, it was something that her parents recommend as a career, and so she said to herself that she wanted to become an artist. In this case, I think we could admit that at least in one sense Bertha was wrong to think she desired to become an artist, and therefore this desire (which she lacked) cannot be the explanation of her action. This of course doesn’t, says the Humean, mean we cannot point to a different desire as the

\textsuperscript{130} Nottelmann, “Belief-Desire Explanation,” 918.
explanation of her action. That is, *Bertha wanted to please her parents*. The Humean account of Bertha’s action is that she wanted to please her parents, and she believed by going to art college she would please her parents, so she went to art college.

In general then, as Nottelmann observes, either the agent in question actually does have the desire she believes she has (Agatha) and this is part of the explanation of her action, or she has a different desire that is part of the explanation her action (Bertha). Either way, the Humean account is safe.

The second objection to ER and ET which Nottelmann discusses I shall call the *evaluative judgment objection*.

The argument from evaluative judgments typically proceeds thus: Often we hold evaluative beliefs with contents such as *<it would be highly desirable to take off my jacket>* or *<it is ethically correct to help this poor stranger>*. And often we act as we deem desirable or ethically correct. Now, so the objection goes, not only do such evaluative beliefs provide us with an internal reason for acting; it also seems roundabout to explain such actions by citing some desire, e.g. an overriding desire to do what is desirable or ethically correct. Merely citing the relevant evaluative belief suffices to explain the action in violation of the Humean’s ellipticality theses.\(^{131}\)

The Humean should respond by first conceding that, in terms of conversational pragmatics, citing relevant evaluative beliefs often does suffice to explain action, and explanations that explicitly refer to agent’s desires can seem rather “roundabout”. However, this does not mean that the explanation can work without us assuming the existence of a desire, unstated. The same kind of attention to pragmatics can also help the Humean to deal with cases where we would say, for instance, “I did the right thing because it was the right thing to do, not because I wanted to.” When we understand such claims, says the Humean, we do so by interpreting what is being said to mean:

\(^{131}\) *Ibid.*, 919.
(1) I did the right thing because I desired to do the right thing and didn’t want to do the wrong, (and I believed that doing what I did was doing the right thing and not the wrong).

(2) But if I hadn’t believed it was the right thing to do, I wouldn’t have wanted to do it, because it is unpleasant, time-consuming, etc.

What we cannot understand is a claim like this:

(3) I did it because it was the right thing to do, but I had no desire at all to do the right thing. I don’t just mean that I didn’t find the prospect of doing the right thing attractive. I simply don’t care about doing the right thing.

This simply makes no sense, and what this shows is that desires understood in a broad sense are necessary to the explanation of action even in cases where we cite an evaluative belief as the explanation of the action. This necessity does not mean, though, that we must explicitly mention the desires when we explain such actions.

Having accepted the Humean elliptical theses, it seems only fitting to ask whether there is anything beyond a desire-belief pair that must exist for an F-type explanation to be correct. I will briefly indicate why I think intentions are also elliptically required.

Explanations in terms of desire-belief pairs alone are partly elliptical, for they make no mention of the agent’s intentions and do not tell us whether she intentionally acted on her desire or merely as one would expect someone with this desire to act, and acted for some other reason. Suppose there are a number of things Quorra disliked about her job, things she wished were different. She did not like the pay, she did not like her boss, and she did not like her commute. Quorra knew that quitting a job and taking up another job she had been offered would mean an improvement in pay, in her boss, and in commute. If Quorra quit, then although we can say that she desired a better commute, and believed

---

132 Some, such as Shafer-Landau, object to seeing caring as a type of desiring, preferring to classify it as (for instance) a kind of valuing instead, but for the purpose of this thesis I am allowing the Humean to use the term “desire” in a broad sense, the sense of Davidson’s “pro-attitude”. See Russ Shafer-Landau, Moral Realism: A Defence (Oxford: Clarendon Press, 2003), 126; Davidson, “Actions, Reasons, and Causes”, 686.
by quitting she would get a better commute, that does not mean this desire-belief pair was her reason for quitting, or one of her reasons for quitting, unless she actually intended to shorten her commute by quitting.\textsuperscript{133} If she did not have this intention, only the desire and the belief, this was not her reason for quitting. There will be a Humean style explanation (since ET, ER and HE are true), but this is not it – instead, she may have quit in order to change her boss, or in order to get better pay. In general, the fact that the agent φ-ed, combined with the truth that she desired to ψ and believed that were she to φ she would ψ, does not entail that this desire and belief are why she φ-ed. Unless she intended to ψ by φ-ing, the desire-belief pair are not the explanation of her φ-ing. Thus, a desire-belief pair is also an elliptical explanation of an action.

Humeans may feel there is no need to talk of intention to complete a Humean explanation. They may feel we can do the job by stipulating that we pick out only those Humean pairs that cause Quorra to quit – these pairs are the reasons why Quorra quit. This may be fine, if we allow a rather loose use of “cause”, but it is acceptable only because it effectively states the truism that a pair is the reason for her action only if it is the reason for her action. This is true but not informative. If “cause” has a narrower, more scientific meaning here, then the causal condition is more objectionable – and I do not think we have any clear notion of what it means for beliefs and desires to cause actions in the scientific sense. I do allow that certain brain events, against the background of certain brain states, can cause certain muscle movements, but beliefs and desires are not brain events or brain states, even though it is true that they are not possible without brain events and brain states.\textsuperscript{134}

\textsuperscript{133} There is a weak sense in which anything that we knowingly allow to happen as a result of our actions is “intentional” or “our intention”, whether it was actually our purpose (in a strong sense) to produce that effect or it was merely an effect that we knew would be produced. The weaker intentions are what Jack W. Meiland calls “non-purposive intentions,” in The Nature of Intention (London: Methuen & Co Ltd., 1970), Chapter 1. However, I am arguing that intentions in the strong sense (purposive intentions) are always at least implied, if not stated, in any case of successfully explaining an action done for a reason.

\textsuperscript{134} See Chapter 6.
VI  Constitutionalism Undermined

I have argued for and accepted the Humean elliptical theses ER and ET, and the Humean existence thesis HE. We now turn to the constitutional question of whether desire-belief pairs, beliefs alone, or some other entity such as facts, events, states-of-affairs, propositions, goals, intentions, character traits, brain states, causes, or combinations of these, constitute the reasons why people do what they do. If we are careful we will recognize that this is really a cluster of questions: we may be asking about the constitution of narrow D-type reasons, broad-F type reasons, or narrow F-type reasons, or (in Smith’s idiom) the source of motivation, motivating reasons.

However, the view I am combating says that there is a way to reduce this complexity, so that only one question is really of concern. Although in Section IV we saw that we are quite permissive when it comes to what we call explanations of action and reasons for action, someone who wants an answer to the constitutional question aims to get beneath this permissiveness – to uncover what really explains action, as opposed to what we loosely allow to be called an explanation or a reason. We can mark what she is asking after with a variety of adjectives: real, primary, genuine, fundamental, proper, or true reasons or explanations. I will simply talk of real explanatory reasons. The constitutional question concerns what constitutes our real explanatory reasons and, says the thought, we can leave the question of what this means for the taxonomy of D-type reasons and F-type reasons (etc.) for later. Let us run with this thought for now.

I have accepted the elliptical theses and HE and so I might be expected to accept a Humean constitution of real explanatory reasons for action. That is, given that desire-belief pairs are always required for actions done for reasons, they must be what really explain these actions. The apparent range of types of explanatory reasons is little more than an accident of the English language, says this view, as a desire-belief pair is what really does the explanatory work. We can call this view the real constitution thesis, RC:

Whenever an agent φ-ed for a narrow F-type reason, there must have existed a Humean pair HP(φ) which is the real explanatory reason for her action. (RC)
However, I do not accept RC, or any non-Humean alternative account of what our real explanatory reasons for action are. Furthermore, I think it is a mistake to search for the constitution of real explanatory reasons. To show why, I begin with a quote from Ralf Stoecker:

Reasons are basically something put forward or given. Someone gives a reason for or against doing something. The cook…may give several different reasons for firing the oven, for example:

1. It saves energy to preheat the oven.
2. The cake will rise if the oven is already warmed up.
3. It is my duty to make sure that the guests get cake at tea time.
4. I like working in the kitchen.
5. I don’t want to be told off by the boss again.
6. I intend to bake a cake.¹³⁵

According to Stoecker, what unifies the items 1-6 is that they could each, in the right circumstance, be given as (part of) arguments for firing the oven, and accepted as such. What separates them is whether they are, express, allude to, or refer to, general facts (see 1), future states (2), values or norms (3), emotions or feelings (4), and what are known as intentional attitudes (5 and 6). What makes items 1-6 reasons is not, therefore, that they are (or refer to), say, norms as opposed to something else, since not all are (or refer to) norms, but that they are one of the many kinds of things that can be offered as a reason. What limits what can be offered as a reason is that it must play a role in helping the recipient of the reason make sense of what has happened, or sense of what to do, not its occupying some ontological category.

In this way, the nature of a reason is like the nature of something else we offer and give, help. If one looks at all the various things that one does when one gives help and tries to find the common core, one will not find any common core. If Tony has pulled his car to

the side of the motorway because of a flat tire, and I come along and give him help, I may help him by doing any of these things:

1. Loaning Tony a spare tire from my car.
2. Fixing the puncture in Tony’s tire with some sealant.
3. Explaining to Tony how to change the tire.
4. Setting up traffic cones so that no one will drive into Tony while he attends to the situation.
5. Calling the Automobile Association.
6. Giving Tony a lift to a petrol station.
7. Taking the old tire and disposing of it for Tony.

What is important is that what I do somehow fills Tony’s need or needs, or part thereof, but filling his need may be constituted by loaning something, fixing something, taking something, taking someone somewhere, placing something somewhere, saying something, asking for something, and so on. The common core is that items 1-7 function to fill (or are meant to fill) his need, not that they are all one type of activity (except of course, that they are all a form of helping).

In making these points, I mean more than just to point out an analogy between the giving of reasons or explanations and the giving of help. When I give reasons or explanations, I actually give a kind of help – I say something that is intended to help the recipient (which may be myself, if I am considering matters alone) to gain a better understanding of the situation, or some aspect of it, and what is key is that what is said is intended to fulfil that role and intended to be recognizable as such by the recipient, not that it is of a certain type independent of that role.

It is this pragmatic rather than constitutional essence to explanations and reasons that explains our general permissiveness in calling things reasons or explanations, examined in Section IV. In the right situations, both “He intended to bake a cake” and “It saves energy to preheat the oven” can help a recipient to understand why he fired the oven, and therefore both can be called the reason why he fired the oven, or his reason for
firing the oven. The shape of a recipient’s need, and what will fill it, will depend on at least three factors – the state of his prior knowledge, his area of concern or interest, and the depth of understanding sought. When the explanation offered does not fill the need of the recipient, he will indicate his dissatisfaction…

That just doesn’t make sense.
That can’t be the reason.
Yeah, yeah, I know that. But that still doesn’t explain why….

…and/or remaining need:

Interesting. But why did he want that?
OK. But what’s good about that?
Why on earth did he think that though?
I see. But how did he come to be a person like that?
All right. But why is that true?

What brings a line of questioning to a close will depend on the case.

I seem to be saying that any statement that fills the recipient’s need counts as an explanatory reason. Someone may object that this pragmatic approach to explanation and reason is far too lax. Not everything that satisfies a recipient’s need to understand the situation counts as a statement of a reason or explanation. Take this (admittedly silly) example. Suppose Billy fires the oven two hours before service, and Suzy wonders why he is doing it. If the following day Suzy attends a chemistry lecture on endothermic and exothermic reactions, her lecturer’s comment about the chemistry of wood combustion might trigger in Suzy the realization that cakes have a certain sort of chemistry, and thus the realization that Billy must have been heating the oven so as to prepare for baking a cake. We wouldn’t want to say that the lecturer had told her Billy’s reason for firing the oven, even if what the lecturer had said did actually lead to the end of Suzy’s puzzlement and the satisfaction of her need for understanding. There must be something more to reasons than their helpful effects, and (the objection continues) this
is best captured by some account of reasons as being constituted by, say, the desire-belief pairs which drove the action.

However, although it is quite true that not every statement that somehow leads someone to an understanding of an action counts as a statement of a reason, we do not need to resort to talk of constitution to see why. This objection is easily dealt with if we notice the distinction between the effects of making a statement and the intended function of a statement. Not every effect is intended, and not every effect is indicative of function. The lecturer did not intend to explain Billy’s action in giving her lecture. If she’d had that intention, she would have formed her comments into a more recognizable form so that Suzy would recognize that this was her intention and recognize the relevance to Billy’s case, removing or greatly reducing the need for an intuitive leap from Suzy.

Let us return to the main issue. Many will grant that at the level of everyday discourse there is nothing that unifies the various ways we talk about reasons apart from their shared function as aids to understanding, but will still want to object that underlying the variety there must be something that can be called the true or real reason for an action, as opposed to all the things we can call a reason in some loose conversational sense. It is this thing which one is helping the recipient of the explanation to grasp. I must say something in response.

What are the real explanatory reasons for action? How are we to pick the real reason from its lesser fellows amongst the elements of the story behind an action? To see the difficulty, let us suppose that we all agree with the following story of action, for at least some actions: when it is a normative fact that X is desirable or valuable and a rational agent believes this fact, this will lead her (or, perhaps, cause her) to desire X, and this desire in turn will lead (or, perhaps, cause) her to act accordingly, in the right circumstances. If we accept this story, we might choose to say that the real reason for the action was (A) the initial normative fact, (B) the agent’s belief in this fact, (C) the desire that this belief led to, or a combination such as (B) and (C) without (A).
We must pay special attention to what criterion is employed to separate *real reasons* from their supposedly secondary companions. If we are interested in tracing *the chain of causes or influence back as far as we can*, we will probably resist calling desires the *real reasons*, because in the story of action sketched above these desires come late on the scene and are explained by something prior – the belief or, even better, the fact (when it is a fact). That is, we will choose (A) over (B), and (B) over (C). If however we are interested in grasping *whatever it is in the lead up to the action that comes closest to guaranteeing that the action will actually occur*, we will probably prefer desire accounts of reasons to belief accounts, and belief accounts to fact accounts - that is, (C) over (B), and (B) over (A). The normative fact does not produce the action if it is not believed, and the belief does not produce the action unless it creates the desire, but (the thought goes) the desire does produce the action unless some stronger desire interferes, or the agent is prevented from acting on the desire by her circumstances. In addition, we also might note that the desire could exist without the belief in the normative fact or without there even being a normative fact, and still produce the action. Or, instead, if we are interested in understanding *the agent’s perspective on what she was doing*, we will probably prefer accounts in terms of mental states of belief and desire – that is, (B) and (C), over (A). This is often our focus when we think the action was a mistake, such as when the putative normative fact is false. If, finally, we are interested in understanding *what made the action appropriate or right for one to do*, that is, appropriate or right for any agent including (but not especially) this agent, we will want an account in terms of the normative facts that made it right, preferring (A) over (B) and (C).

The point is that the criteria we apply in choosing what to call the real explanatory reason are simply a reflection of what particularly interests us about actions and their explanation. They reveal the *bias* in our attention but nothing more, because all of the above criteria are legitimate things to focus on and care about, and all have their importance in rounding out a full picture of what explanations can tell us. Consequently, it is a mistake to speak of one sort of thing rather than another as constituting the *real* explanatory reasons for action.
To conclude, we can commit to ER, ET and HE without committing to RC, and in general, what unifies explanatory reasons is their intended function in relation to an audience’s need for understanding, not a shared essence distinct from that function.

VII  Resolving the Dilemma

In the previous three sections we have covered a lot of ground: our permissiveness in reasons talk, the ellipticality of non-Humean and Humean explanatory reasons, and the constitution of reasons question. What remains is to return to the dilemma outlined in Section III and see if the observations of Sections IV to VI shed any light on the three problems that constitute the dilemma. What is particularly attractive about rejecting explanatory constitutionalism (the idea that explanatory reasons share a nature) is that it makes avoiding these problems comparatively straightforward. I do not claim that this is the only way to avoid the problems, but I think it is the best. We will deal first with the overly-personal problem, then the overly-impersonal problem, and finally the problem of acting for good reasons.

The overly-personal problem

The overly-personal problem is the problem one gets if one accepts that explanatory and normative reasons share a constitution, and believes explanatory reasons and therefore normative reasons are both constituted (whether partly or fully) by desires. Let us put the overly-personal problem in explicit form so that we can see how rejecting constitutionalism avoids the problem. I will present the problem as arguments forcing us to accept one (but not both) of two conclusions: uncomfortable conclusions A and B.

1. An explanatory reason for intentional, non-arational action is constituted by a desire-belief pair (or a desire), whether we act for a good or bad reason.
2. When we act for good reason, the explanatory reason for the action is a normative reason that supports doing the action in question.\(^{136}\)

\(^{136}\) This premise is essentially Dancy’s maxim. See Section III and footnote 114.
(3) Therefore, a normative reason, such as a moral reason, is a desire-belief pair (or desire) that supports doing the action in question.

At this point, we can go two different directions, both resulting in uncomfortable conclusions, depending on whether we accept (4):

(4) Sometimes an agent has no desire to do the moral action or anything that doing the moral action promotes.

Putting (3) and (4) together yields:

(5a) Sometimes an agent has no normative reason to do the moral thing – that is, when she has no relevant desires. (Uncomfortable conclusion A)

Here is the second alternative, which manages to avoid the path to (5a) by denying (4), but which is again based on (3):

(5b) If some action is morally obligatory for someone, (such that there is a moral reason for her to perform it), then she must have a desire that would be served by doing that thing. If she does not have such a desire, the action cannot be a moral obligation for her. Similarly, if some action would be immoral for someone to perform, (so there is a moral reason not to perform it), then she must have a desire that would be served by not doing that thing. However, if she does not have such a desire, the action cannot be an immoral thing to do. (Uncomfortable conclusion B)

It should be clear that I reject premise (1) for the reasons stated in Sections VI, and thus both of the above arguments are, in my eyes, unsound. The arguments fail to establish (3) and also (5a) and (5b), the uncomfortable conclusions.137

Given that I endorsed the elliptical theses ER, ET and HE, it is worth examining whether the arguments for the uncomfortable conclusions can be salvaged by using

137 I do accept (4).
these theses or something like them as premises, rather than any constitution theses. To see why they cannot, it will be enough to consider the first argument.

Although I accept…

(1’) If R can be called the explanatory reason for A’s action, then either R is a desire-belief pair, or there exists an appropriately related desire-belief pair HP, that can also be called the explanatory reason for A’s action.

…and…

(2’) When we act for good reasons, some of the things that can be said to be our explanatory reasons can also be said to be our normative reasons – in particular, certain facts in the light of which we act.

…and…

(4) Sometimes an agent has no desire to do the moral thing or anything that doing the moral thing promotes.

…one cannot conclude on the basis of (1’), (2’) and (4) either of the following:

(3) A normative reason, such as moral reason, is a desire-belief pair (or desire alone) that supports doing the action in question.

(5a) Sometimes an agent has no normative reason to do the moral thing – that is, when she has no relevant desires. (The uncomfortable conclusion A)

Thus by rejecting constitutionalism we straightforwardly avoid the overly-personal problem, even if we accept (as I think we should) the elliptical theses and HE.

The overly-impersonal problem

Just as with the overly-personal problem, it will help to express the overly-impersonal problem in the form of an argument with an uncomfortable conclusion. In this case, the uncomfortable conclusion will be a set of unattractive options: either moral reasons fail
to explain actions except by making them involuntary, or moral action is heartless, or moral action is motivationally mysterious.

(1) Normative reasons, including, moral reasons, are constituted by non-psychological facts.¹³⁸
(2) When we act for good reason, the normative reason we act on is the explanatory reason for the action.
(3) Therefore, explanatory reasons, (at least when we act morally or rationally), are also facts.¹³⁹
(4) The agent’s explanatory reason is the thing which motivates the agent to act.
(5) Therefore, what motivates the agent to act in moral cases is the normative fact that is their explanatory reason, and nothing else, in particular not any desire of the agent that will be served by acting accordingly. If there is such a desire that will be so served, it is irrelevant to the explanation of the action and is not the agent’s reason for acting.

There are various further conclusions we might draw from (5). Either…

(6a) When an agent acts for moral reasons, the agent acts involuntarily.

Or, somewhat differently, …

(6b) When an agent acts for moral reasons, the agent acts voluntarily but heartlessly, mechanically, like a passionless rational automaton.

¹³⁸ In place of “facts”, the reader can insert preferred non-Humean constituent: propositions, or states of affairs, or truths, or beliefs as objects rather than mental states. For instance, “killing is wrong”, “one should pay one’s taxes”, “she needed help.” What the non-Humean will not insert are mental states such as desires or beliefs (as mental states rather than things believed) or desire-belief combinations – for instance, his desiring to help her, or his belief/believing that paying taxes is just combined with his desire/desiring to act justly.

¹³⁹ I have omitted the phrase “when we act for good reasons” which occurred in the previous argument from this argument since, if reasons share a constitution, that constitution does not depend on the situation. In particular, whether one is acting for good or bad reasons, those reasons should still be the same sort of entity.
Alternatively, if we are not satisfied with either (6a) or (6b) we seem forced to conclude…

(6c) Acting for moral reasons is ultimately motivationally mysterious.

Whether we conclude (6a), (6b) or (6c), we have a problem.

However, as with the overly-personal problem, the overly-impersonal problem dissolves if we do not believe that reasons share in a constitution. That is, if we reject (1) and (2), neither (3) nor (5) is supported, and we are not forced to choose between (6a), (6b) and (6c). Unlike with the earlier problem, we do not have to contend with the possibility that the argument can be salvaged using elliptical theses, since there is no non-Humean equivalent to the Humean elliptical theses.

I would like to say a little more at this point about moral action. If we endorse the Humean existence thesis HE, we believe that for every intentional action there is a Humean pair that can be used to explain that action to a suitable recipient – a recipient interested in the psychology of the agent. This means that every case of acting morally is also describable as a case of acting according to one’s desires (broadly construed). If John visits his sick grandmother in hospital, to the right recipient we can explain the fact that John did this by citing John’s desire to do the moral thing, and this marries unproblematically with saying that the narrow-F-type reason for his action was that visiting a sick grandmother is the moral thing to do. Likewise, we can explain his choice, to the right recipient, by citing John’s desire to make his grandmother happy, and this marries unproblematically with saying that his narrow F-type reason was the fact that his grandmother would enjoy a visit.

It should be added that there is no problem with maintaining that the reason why John performed his action was both that it was the moral thing to do and that it would make his grandmother happy. It is possible that these two reasons act together as narrow F-type reasons in support of his acting. Alternatively, one of these reasons can be the narrow F-type reason while the other is a background narrow D-type reason. For
instance, he may consciously choose to visit his grandmother because it would make her happy, never thinking about the fact that it is also the moral thing to do. If he is a moral man, the fact that this kind of thing is the moral thing to do is something he was raised to respect, and it played a role in producing and strengthening the kind of desire he now feels to make his grandmother happy. Thus it is a narrow-D type reason for his action even though he does not think on this occasion about morality. Whether both reasons are F-type, or one is F-type and one is D-type, we are not forced to choose between saying it was a moral consideration rather than a desire that she be happy, or a desire that she be happy rather than a moral consideration, that explains his choice.

The problem of acting for good reasons

We want to be able to say that when a person acted for a good reason, the reason which explains why she did what she did is the reason there truly was for doing that thing. That is, the explanatory reason and the normative reason are identical. If one holds that normative reasons and explanatory reasons have different constitutions (e.g. facts as opposed to desire-belief pairs), one cannot say the two reasons are identical except as a kind of metaphor or shorthand for a different relation than identity, such as the relation of cause and effect.

However, if we reject constitutionalism, we can speak straightforwardly of the identity of an explanatory and a normative reason, but only if the explanatory reason is a narrow F-type reason. When we talk about a reason there is for doing something (normative reason) and want it to be potentially the same as someone’s reason-for doing it (an explanatory reason, when she has done it), we want the normative reason, which is a consideration that counts in favour of acting in some way, to be the same as the explanatory reason, characterized as the consideration that the agent recognized as counting in favour of acting and acted upon. The latter is the narrow F-type reason that explains her action, defined in Section IV as the consideration that the agent recognized as counting in favour of acting, not some other thing from which we can infer such a consideration. Whatever considerations are taken to be – situational facts, psychological facts, states-of-affairs, propositions, truths, or beliefs – we can assert without difficulty
that the narrow F-type explanatory and normative considerations are the same when we act for good reason.

Although this solves the problem of acting for good reasons, a little additional comment may be called for, because the notion of acting for good reason is somewhat ambiguous. In outlining a taxonomy of reasons in Section IV, I stipulated that both facts and beliefs are to qualify as considerations that can constitute F-type reasons, and extended the taxonomy further to include a place for only those F-type reasons that are facts or true beliefs, calling these T-type reasons. The distinction between T-type reasons and the narrow F-type reasons which are not T-type reasons is reflected in two ways in which we use the word “reason” in cases where the agent acts on a false belief. To take a familiar example, suppose that Roy has mixed some petrol into his martini, thinking it was gin. If we want to know the narrow F-type reason for his action, we can cite his false belief that the liquid was gin. In that sense, he had a reason for pouring the petrol into his drink. However, if we want to know a T-type reason for his action, we must admit that there was no reason for pouring petrol into his drink. Thus in one sense, he had a reason to do what he did, while in another sense, he did not. I disagree therefore with Bernard Williams, Rowland Stout, and Maria Alvarez, who argue that when I act on the basis of a false belief, I simply have no reason for doing what I do. By reflecting on how we actually talk, the reader can see for herself that Williams, Stout and Alvarez are only half right.

There is a similar ambiguity in talk of “good reason” when we say “he had good reason for doing what he did.” One sense is where we use “good” to emphasize that we do not in any way blame the agent for what he did, even if his action was a mistake in the sense that he acted on a false belief. If I try to douse a dangerous flame with petrol that some prankster has put into a water bottle, one can say I had good reason for doing so, because I believed it was water and I believed this liquid would put out the fire. Both beliefs were false, but I am not to be blamed for acting on them unless we think I should

---

have known the bottle contained petrol. I may even be praised for acting as I did. However, another sense of “good reason” is one where we require that no false beliefs of the agent are part of the explanation of his action. Here our concern is with what the best thing to do was, all things considered, in light of all the facts. In that sense, I had no good reason to throw petrol on the fire. Thus in one sense, I had good reason to pour petrol on the fire, and in another sense, I did not.

VIII Conclusion

To conclude, rejecting constitutionalism allows us to solve the overly-personal problem, the overly-impersonal problem, and the problem of acting for good reasons. Explanations of action are not limited by ontology, but by pragmatics.
6 In What Sense Are Reasons Causes?

I Introduction

In the previous chapter, I made some brief remarks to the effect that there is little sense to talk of beliefs and desires causing behaviour if the type of causation in question is meant to be of a scientific sort.\textsuperscript{141} I chose not to develop that thought there, but to leave it for separate consideration. Likewise, in the previous chapter I dismissed attempts to characterize the real nature of explanatory reasons, including characterizations that saw the real reasons for action as whatever causes the action.\textsuperscript{142} I did not in that chapter pay specific attention to whether or in what way reasons might be causes, and that is something that is worth doing if my position (or anti-position) is to be attractive. Thus I now turn to this issue.

We can begin by posing a familiar question that arises in relation to the performance of actions for reasons. The question is:

\textit{What is it that distinguishes the reason (or reasons) for which an agent acted from the many possible reasons for so acting which she may have known of but which were not her reason for acting at the time?}

I will call this the pro-reason question because it concerns the reasons in favour of performing the action. One answer that is commonly given is a causal answer or causal solution:

\textsuperscript{141} Chapter 5, Section V. 
\textsuperscript{142} Chapter 5, Section VI.
The difference between the reason (or reasons) for which an agent acted and the reasons for which she did not is that the former caused her to act as she did while the latter did not.

Typically, the pro-reason question is raised, and the causal solution offered, as a way of arguing that reasons for action are causes. The idea is that those philosophers who reject an account of reasons as causes will be unable to satisfactorily answer the pro-reason question, and thus we seem to have good grounds for accepting that reasons are causes.

One sees this argument, of course, in Donald Davidson’s famous essay, but also in more recent work. For instance, here is Arpaly and Schroeder:

The simplest case for the thesis that rationalizing reasons must be causes is that there is no other way of explaining which, of multiple available rationalizing reasons for an action, was an agent’s actual rationalizing reason: the reason for which she actually acted.

Arpaly and Schroeder offer the example of a father, Richard, who punishes his daughter’s refusal to wash her dirty hands by sending her to her room. They note that there might be many reasons for someone to do this – some more acceptable than others. What settles what Richard’s reasons actually were for doing what he did? They answer:

Not the manner of speech or facial expression. That can be evidence to the observer, but it is only evidence, and can be misleading (a snarling facial tic might make a punishment seem to be given out of a desire for dominance when it was not). Not Richard’s beliefs about the matter. Since Kant, at least, we have been aware that we do not always know our actual motives. Not by holding that every rationalization available to Richard plays an equal role. If it had occurred Richard five minutes earlier that his daughter will never become a doctor if her hands are forever full of germs, it would hardly follow that five minutes later the punishment was given partly in order to secure a good career for the daughter (it could have been, but that would indicate an unusual parenting style). The causal

143 Davidson, “Actions, Reasons, and Causes.”
role played by his beliefs and intrinsic desires has always been the best
candidate to fix such facts.145

The Davidsonian view that reasons are causes became the orthodoxy in the 1970s, and
perhaps continues to hold that position today. However, around the turn of the last
century, a movement began against causalism.146 This movement was largely inspired
by ethical realism. According to ethical realism, when one acts for a good reason, one’s
reason for acting is some truth about the world, some fact, or some state of affairs,
which is independent of or external to our psychology in a certain sense. Ethical
realism clashes with the Davidsonian attempt (repeated by Michael Smith) to equate
reasons with psychological entities having causal powers – that is, desires and beliefs.147

The ethical realist criticism of the Davidsonian orthodoxy relies in part on a distinction
realists mark between acts or states of believing and what one believes, and a similar
distinction can be marked between acts or states of desiring and what one desires. The
psychological act or state of believing that the toast is burning is distinct from the fact
that the toast is burning (when it is a fact) and also distinct from the idea that the toast is
burning (whether it is burning or not). Ethical realists hold that it is the contents or
extensions of beliefs (and, perhaps, desires), not the states or acts of belief and desire,
which explain action, and these contents are not causes.

It would seem, though, that the causalist can readily adjust her position so as to take
account of the above distinctions and remain a causalist. She can argue that, when we
act for good reason, a state of affairs has caused it to be the case that we hold the true
belief that the state of affairs obtains, and our believing so together with our desiring
such-and-such, in turn causes our action. We would have causation and explanation of
action in two steps rather than one, and a view that an ethical realist could seemingly
accept. There would merely be a difference of emphasis or focus: the Davidsonian is
more interested in the primarily causal (but also explanatory) relationship between the

145 Ibid., 68.
146 Giuseppina D’Oro and Constantine Sandis, “From Anti-Causalism to Causalism and Back: A History of
the Reasons/Causes Debate,” in Reasons and Causes: Causalism and Anti-Causalism in the Philosophy
states of believing and desiring and the resultant action, while the ethical realist is more interested in the primarily explanatory (but also causal) relationship between the state of affairs that is the object of belief and the resultant action. In this scenario, the Davidsonian causalist would have the nominal advantage that the relationship she focuses on exists whether or not the agent is acting for a good reason. The ethical realist would need to admit that her reason-for-relationship does not always exist – and therefore, as she uses ‘reason’, when we do not act for good reason, we do not act for a reason at all, and have no reason for doing what we do.148

However, I mention this way of conceiving of the dispute in order to set it aside – ultimately, I think causalists and non-causalists can at least in this respect find a happy middle position that makes space for ethical realism and the distinction between acts/states and contents/extensions.

Another and to my mind more important distinction that concerns reasons for action and causes is also drawn in the writings of the anti-causalists: the distinction between actions and motions of the body.149 Most philosophers would accept that when we act, motions of bodily parts are caused by brain states and events. However, if actions are not equivalent to bodily motions, one is not required to accept that brain states cause actions. Similarly, if one draws a distinction between mental states such as believing and desiring on the one hand and brain states that accompany mental states on the other, one has a second way to accept brain-motion causation without necessarily accepting mind-action causation. These distinctions mean there is space for an anti-causalist view of reasons for action, even if reasons for action are conceived of as mental states of

148 Some philosophers accept this position – that is, that when we do not act for good reason, we do not act for a reason at all, and have no reason for doing what we do. See for instance Maria Alvarez, *Kinds of Reasons*, 135-147, and my comments in Chapter 5, Section VII criticising Williams, Stout and Alvarez. I think we do sometimes talk in this way, as if bad reasons were not reasons at all – “I know he insulted you. But that's no reason for killing him!” One would assume the sense of “no reason” in the claim was, in context, “no good reason.” However, there is another way of talking about reasons that is also acceptable, where bad reasons for acting are still genuine reasons, just bad ones. “Your reason for killing him was that he had insulted you, but that's just not a good reason at all for killing someone!” Alvarez's claim is, in other words, at most conditionally permissible (see Chapter 1, Section IV) but ought not to be presented as a philosophical thesis that is true (or false) taken out of any context.

149 For a discussion of this distinction, see Alvarez, *Kinds of Reasons*. 
desiring and believing. But what type of anti-causalism does this leave space for? Might what we get be an anti-causalism in one sense while a causalism in the other?

This leads me to what is of primary interest in this chapter. I will argue that if we permit a way of talking which treats states of affairs and mental states as causing actions, we must be very clear that the type of causation in play is importantly different from the type of causation in play when saying that brain states cause motions. I will call the former type of causation, from states of affairs or mental states to actions of agents, *soft causation*. I will call the latter, from empirically identifiable physical states or movements (such as brain states) to states or movements of physical objects (such as movements of bodies), *hard causation*. Likewise I will talk of soft and hard causes, and soft causalism as opposed to hard causalism. Hard causalism is the view (often effectively assumed rather than articulated) that all causes, including reasons for action, are hard causes. Soft causalism is the view that some causes, in particular reasons for action, are soft but not hard causes.

Hard causalists tend to equate mental events or states with brain events or states. Accordingly, they make claims such as this:

[S]uppose the thought of meeting her one o’clock class fills Emma with dismay. One neural event, her thinking of the class, causes another neural event, her being dismayed.\(^{150}\)

Likewise, hard causalists typically conceive of actions as motions caused (in the right way\(^{151}\)) by mental events or states.

Here is the plan of this chapter. I believe soft causalism is true and hard causalism is false, and that the causal solution is only true under a soft causal interpretation. To

\(^{150}\) Arpaly and Schroeder, *In Praise of Desire*, 57.

\(^{151}\) The “in the right way” condition is added because of the phenomenon of deviant causation – an example is Davidson’s nervous climber in “Freedom to Act (1973),” in *Essays on Actions and Events* (Oxford: Clarendon Press, 2001), 79. It is an interesting question whether adding “in the right way” might actually be enough to make the Davidsonian view soft-causal rather than hard-causal. However, Davidson does not say enough to suggest that this was in fact his view.
argue this, I will explore the distinction between hard and soft causes in relation to the epistemology of causation (Section III), the ontology of causation (Section IV) and the deontology of causation (Section V). I will thereby show the sense in which Davidson and others are right that reasons for action are causes, i.e. under a soft-causal interpretation, and the sense in which they are wrong that reasons are causes, i.e. under a hard causal interpretation. Before any of this, I will look in Section II at our ordinary language use of “cause”. Facts about our usage seem to support a straightforward argument against saying that reasons are causes. However, the argument is too swift, and needs to be undermined in order that my distinction between hard and soft causes is seen as worth marking.

II “Cause” in Ordinary Language

Let us begin by looking at a few examples of how the verb “cause” is used in non-philosophical contexts to describe why someone did what she did. I will provide some typical examples, drawn from online news headlines amongst other sources. First, there are those cases where an agent is caused to do something unintentional or non-intentional:

(1) The cold salt water on his face caused him to come up gasping.
(2) Joan Rivers’ Unplanned Throat Biopsy Caused Her To Stop Breathing

Similarly, there are cases where the agent claims more or less credibly to have done something unintentional:

(3) “The dog tripped me”: Dad accused of killing 10-week-old son claims his miniature dachshund caused him to fall on the boy.

These are not cases of doing things for a reason – assuming we accept the Dad at his word – which is to say, they are not, to use a term from the previous chapter, F-type

---

152 I have elected to focus on the use of the verb “cause”. Sentences which use the noun “cause” to explain action follow the same pattern.
reason cases. As such they needn’t concern us, except that they are interestingly related to other groups of cases.

The second group of cases are those where the action performed is intentional, but is either performed reluctantly, or at least with the sense that given the circumstances the agent had little or no choice but to act as she did. These cases often involve the giving up of something the agent values:

(4) Charlie Hunnam Says Nervous Breakdown Caused Him To Quit “Fifty Shades Of Grey” Movie
(5) Becker’s photograph was in the German newspapers yesterday, accompanied by lamentations about his physical condition after a knee injury had caused him to abandon the Classic on Thursday.

Included in these cases are ones where the claim is being used to indicate or argue that the agent has diminished responsibility or no responsibility for doing what she did, though she acted intentionally or at least not unintentionally. For instance:

(6) Oscar De La Hoya claims drugs and alcohol caused him to be unfaithful.
(7) Woman Claims Her Brain Injury Caused Her to Become a Dominatrix

A quite different group of cases are those where the agent is definitely responsible for acting as she did, and the word “cause” now seems appropriate because she lacks control over herself, particularly her emotions and the effect they have on her, and culpably so. These cases often involve a fit of temper.

(8) Irritation caused her to snap at him.
(9) But when Andy announced he wanted to cool their relationship a little, Frances felt a panic and dread inside which caused her to lash out in unexpected anger.

Based on the above, we seem to be on the verge of a rather simple rebuttal to any kind of causalism, and to the causal solution to the pro-reason question. That is, we might
argue that the word “cause”, used either as a verb or noun, is reserved for cases that are importantly different to the typical cases of action where we act for reasons. When we do something ordinary such as walk down the driveway to the mailbox to collect the mail, we typically do not act (a) unintentionally, (b) reluctantly, (c) as if we had no choice, (d) with diminished responsibility for our actions, or (e) from lack of self-control. And if only in these special cases could we say that a desire to get the mail caused the action, then in the normal case the desire does not cause the action, even though it is the reason for the action. Therefore reasons cannot be causes of action, or at least not always so.

But we can ask whether it is the case that “cause” is and must only be used in the ways I have outlined. We can ask whether it is truly a misuse of language to say, for our ordinary case, “My desire to get the mail caused me to walk to the mailbox,” or “the cause of my trip to the mailbox was my desire to get the mail.”

I certainly find such locutions rather odd, and find I must consciously resist automatically interpreting the case as special in ways (a) to (e). For instance, I must avoid interpreting the above as if the situation were like this: throughout the day I had been excitedly anticipating the arrival of an important piece of mail, which never arrived. That night my desire caused me to sleepwalk to the letterbox. However, even if these locutions strike us as odd and easily misinterpreted, this does not necessarily mean that there is an important truth behind our reaction such that there is good reason not to allow a broad use of “cause”. The fact that it is odd to describe the colour of someone’s complexion as sherbet doesn’t mean that there is good reason to disallow such a description. Furthermore, a supporter of the causal solution can argue that she does not wish to be confined by the details of our usage of the word “cause” in ordinary language. Her usage, she may tell us, is related to but different from ordinary usage.

In the previous chapter I explained that the terms “reason” and “explanation” can be used for any of the cases covered in the reason taxonomy, and they are generally used interchangeably. For the purposes of this chapter, I will be equally permissive over the

---

153 Chapter 5 Section IV
usage of “cause” – that is, I will allow any reason or explanation why something happened to be called, in some sense, a cause of that thing happening.

This does not of course settle the question of whether reasons are causes in the sense that Davidson and his followers mean, nor whether reasons are what I have called soft or hard causes. In allowing “cause” to have a very broad use, I only mean to say that I will not be using an argument from ordinary language to object to the causal solution. In Section VI, having explored the epistemology, ontology, and deontology of soft and hard causation, I will however draw a connection between how “cause” is used in ordinary language (to refer to what is done in ways (a) to (e)) and the version of the causal solution I find objectionable.

III The Epistemology of Causation

In this section, I will outline the difference between how soft and hard causal claims are established or come to be known, and deal with a major objection to soft epistemology. Our examples of hard causal claims will be neurological causal claims, and these will be compared with the soft causal claims, if they can be called “claims”, which are involved in the giving of one’s own reasons for action. This section will be brief. For a lengthier discussion of the epistemological distinctions between causes (that is, hard causes) and reasons, I direct the reader to Paul Johnston’s Wittgenstein and Moral Philosophy.¹⁵⁴

Let us begin with neurology. Neurological causal claims, like other empirical claims, are established – to quote Johnston – “by experiment, through statistics, or by tracing a mechanism.”¹⁵⁵ When causal claims are established by experiment, this will require repeatability, such that we can say that the pattern of events in a single experiment is in accordance with a general law, whether a strict or statistical law. That is, it must be possible for the scientific community to test a purported causal claim regarding a particular brain or muscle event by producing sufficiently similar experimental initial conditions and seeing if they observe the same outcome, or seeing if they observe at

¹⁵⁵ Ibid., 38.
least a statistically reliable tendency for the same outcome to occur. For instance, a causal claim might cite an increase in sodium ion concentration as the cause of the firing of a particular individual neuron. For this to be valid scientifically, it must be possible to demonstrate that equivalent increases in sodium ion concentrations at equivalent points on the surface of other neurons (of equivalent initial charge) are reliably followed by the firing of those neurons.156

Importantly, the agent within whose brain and body the events occur usually has no special insight as to what the neurological cause of a neurological or muscular change is. Though he can know why he raised his arm in terms of what he hoped to achieve by raising it, he almost never knows which neurons in his body are responsible for triggering the lifting of his arm. For hard causation, it makes no difference that it is his arm as opposed to someone else’s arm that has risen. His knowledge of this causation is, like anyone’s, from outside.

By contrast, reason-for-acting soft causes are typically determined by asking the agent herself what she is or was doing. She is in a special position to know the cause of her action because she is the one doing it. More or less equivalently, she is in a special position to tell us which re-descriptions of her action are true or not, for many if not all re-descriptions. She can tell us, for instance, whether in raising her arm she was hailing a taxi or not, and whether in winking she was indicating to a confederate that she had lied or not. She can also tell us whether a particular movement was intentional or not. Her understanding of reason-for-acting causation of her own actions is from inside.

Unlike with neurological causal claims, we do not regard the agent’s claim about why she acted as a mere hypothesis which awaits experimental confirmation before being justified. Unlike with particular neurological causal claims ("the firing of neuron 246b was caused by an increase of relative internal charge to -50mV at location Ω"), the agent is not indirectly asserting that a law of causal connection exists to cover this and other cases. When she gives her reason for acting, she tells us only what happened here

---

156 Or something like this. What scientists would probably demonstrate is that there is a law-like relation between the relative internal potential of any neuron and its tendency to fire.
and now. She is not asserting that in the same circumstances she would always do the same thing, though of course she may do the same thing in the same circumstances, and perhaps reliably so.

Some philosophers doubt that there exist empirical laws which could accurately map the relationship between reasons for action, the associated states of belief and desire, the associated brain states, and actions themselves. If laws of this sort existed, that would seem to undermine the notion of free will or rational agency. So, for instance, Davidson concludes that “nomological slack between the mental and the physical is essential as long as we conceive of man as a rational animal.”¹⁵⁷ I agree, though I am not sure if my reason for thinking so is the same as Davidson’s. I think that there is no suitably empirical procedure for determining an agent’s exact mental state, a procedure or procedures that could determine, amongst other things, what is believed, under what description(s), with what confidence, with what degree of understanding, and with what degree of authenticity. But an such an empirical procedure is surely a precondition for constructing testable laws relating the mental and the physical.

I would now like to consider, from an epistemological perspective, a view I call hard reductionionism or hard eliminativism. According to this view, talk in terms of soft causation is merely a poor substitute for what genuinely explains action: hard causation and hard causes. Accordingly, what settles the reasons why someone has acted are the facts as determined by empirical science (such as neurological facts), not the agent’s opinion about why she is doing what she is doing. Indeed, the agent may well be in error about why she is doing what she is doing. This possibility for error is recognized, for example, in the earlier quote from Arpaly and Schroeder:

What will establish the facts here [of why Richard has sent his daughter to her room], other than the causal role played by each rationalization?…Not Richard’s beliefs about the matter. Since Kant, at least, we have been aware that we do not always know our actual motives.¹⁵⁸

¹⁵⁸ Arpaly and Schroeder, In Praise of Desire, 68.
The thought that traces back to Kant “at least” is of course more fully developed and more familiar in the work of Freud. Freud introduces the unconscious, in which repressed desires may reside, ready to spark agents into actions which they will mistakenly attribute to other, purer, motives, or claim to be unintentional. Freudian psychotherapy is amongst other things the process of uncovering the true causes of such actions. The existence and power of the unconscious seems to offer support for a view that the agent does not have anything better than modestly reliable access to her own motives or the well-springs of action.

This is a clue to what is objectionable about the causal solution of Section I. At least one thing causalists might mean in saying reasons are causes is that what someone’s reasons for acting are is not established by what they say on the matter. Instead, what they say on the matter is to be judged as according with, or failing to accord with, the truth on the matter as determined in some other way. That is, even if an agent does have some kind of first-person authority over why she does what she does, there is something else of greater authority. Therefore, to say with these causalists that the reason why an agent φ-ed is also the cause of her φ-ing (for some action φ) is to say that there is a fact which explains why the agent φ-ed which is deeper than and possibly inaccessible to introspection, and it is this fact that both explains why the agent φ-ed and that also rationalizes her φ-ing. However, this is a view that we should reject.

While accepting many of the ideas of Freud, we can question what Freudian cases actually show about the epistemology of the causes of action. It will help to turn to some remarks of Wittgenstein in which he considers the Freudian interpretations of dreams. We imagine a therapist has suggested to her client that the meaning of his dream is this or that.

Now I could imagine that someone …might exclaim “Yes, that is the solution, that is what I dreamed, but without gaps & distortions.” It would then be this acknowledgement that made this solution the solution. Just as, if you are searching for a word while writing & then say: “That’s it, that says what I wanted
“to say!” – Your acknowledgement stamps the word as having been found, i.e. the one you were looking for. (In this case it might really be said: only when you have found it, do you know what you were looking for – much as Russell said about wishing.)

What is interesting in Freudian dream cases (and cases of self-deception) is that the therapist’s confidence in her interpretive solution’s status as a solution is dependent on what the client can come to recognize and endorse about himself. It is the client’s “acknowledgment that made this solution the solution.” It is not, note, empirical data that is supporting the therapist’s interpretation, even if empirical data led her to propose the interpretation. It is good old fashioned self-reflection, introspection, or first-person authority of the client. The client is being asked whether the interpretation makes sense to him, and thus he is being asked to make sense of his mental life in a new way – to acknowledge thoughts and feelings that he may have formerly ignored or actively repressed.

Arguably, it only makes sense to talk of ignoring or repressing something one is consciously aware of and therefore knows of. Ignoring and repressing are forms of conscious activity with associated bodily activity (clenching of the jaw, fixing of the gaze, and so on). The notion of repression within or by a largely inaccessible unconscious/subconscious is simply a confusion. It is however possible to forget (in some sense) what one is doing by habituation – e.g. that one is fixing one’s gaze on one thing in order to avoid looking at something else. In this way the actions and the desires that motivate them become “sub” conscious, just as other habitual activities like tying shoelaces are often “sub” conscious, yet perfectly accessible to consciousness, and intentional, without being explicitly intended as such. When it is difficult for someone to access his “sub” conscious, it is not because it is buried deep beyond reach, but because he is engaged in burying it, and uncovering it requires that he desist in a largely forgotten activity of holding it down. The therapist’s couch is a safe place in which he

---

159 Wittgenstein, *Culture and Value*, 78e.
160 The active, physical nature of repression is major focus of Gestalt therapy, a more conceptually coherent approach than Freudian therapy. In Gestalt theory, the agent is the one who represses her true feelings etc., not a quasi-agent like the ego, superego, or id.
can recover awareness of what he is doing, but the fact that he is doing it means it is accessible. That which is truly inaccessible to consciousness is not the sub-conscious but the non-conscious – e.g. the metabolism – the activity of which is not an action of ours, though it makes action possible.

Some readers will remain unconvinced. They will argue that the moment of therapeutic recognition experienced by a client still rests for its authority on something else. They will say that analysis of neurology or physical behaviour will eventually tell us our true motives, so as to make the potential acknowledgment of the client irrelevant. Taken to its natural conclusion, this is to affirm non-belief in persons as persons and minds as minds – to accept an error theory of persons. It is to see a human being as a complicated arrangement of chemicals or dispositions, but not as an agent. This is the hard reduction (the complete reduction of soft causation to hard causation) laid bare, undisguised and ugly.

At the very least, we cannot have our cake and eat it too. If we want our claims about what caused someone to act to have an authority that is independent of what the agent thinks about her own action, then we are no longer interested in making claims about rationalizing reasons, and whatever reasons we discover will not be anyone’s reasons for acting, which I have called F-type reasons. If we are interested in F-type reasons, in what rationalizes action, we must accept that there is nothing deeper or firmer to find than what is accessible to the agent herself.

IV  The Ontology of Causation

Supporters of the reasons-are-causes position like to say that their position has an ancient pedigree. Davidson, for instance, when discussing the “mystery” of the connection between reason and action, mentions “Aristotle’s attempt to solve the mystery by introducing the concept of wanting as a causal factor”, and concludes that “[f]ailing a satisfactory alternative, the best argument for a scheme like Aristotle’s is
that it alone promises to give an account of the “mysterious connection” between reasons and actions.”\textsuperscript{161}

Likewise, Alfred Mele remarks that the “idea that our actions are to be explained, causally, in terms of mental states or events is at least as old as Aristotle.”\textsuperscript{162} In support of this, he cites the \textit{Nicomachean Ethics}: “The origin of action – its efficient, not its final cause – is choice, and that of choice is desire and reasoning with a view to an end.”\textsuperscript{163}

There can be no doubt that Aristotle saw there as being a causal connection of some kind between reason and action, but there is a danger that we read into Aristotle a modern conception of causation or causality, one that is drawn from a modern scientific perspective which Aristotle never held. A detailed exegesis of Aristotle’s position is not appropriate or required here, but I would stress that Aristotle’s notion of “cause” (and, for that matter, “efficient cause”) is broader than that employed in the hard sciences. Each of Aristotle’s four causes is a possible answer to a “why” or “how” question, and many of these “why” or “how” questions are not properly speaking scientific questions.

So what \textit{is} causation as we see it today, which is to say, what \textit{is} the causal relation and what are its relata? I call these ontological questions. I will begin by sketching what I think might be at the heart of the modern, scientific notion of cause, causation, and causality (i.e. hard causation), using neuroscience again as an example, before contrasting this with what I take to be the nature of cause, causation, and causality as pertains to understanding the \textit{whys} of human intentional action (i.e. soft causation).

The nature of causation as employed in the sciences is a hotly contested topic, and it would require at least a book to tackle the subject satisfactorily. However, it will be enough to present here a plausible account of empirical causation to stand in contrast to the account I offer of soft causation. If the reader prefers an alternative theory of

\textsuperscript{161} Davidson, “Actions, Reasons, and Causes,” 693.
empirical causation, she can contrast it with soft causation herself. Still, the account of causation that I think gets closest to the working definition employed in the sciences is the *manipulationist* or *interventionist* account, as described by James Woodward:

> According to this account, causal and explanatory relationships are relationships that are potentially exploitable for purposes of manipulation and control. They are also relationships that are *invariant* under interventions.\(^{164}\)

Let us take an example from neurology. We have a general theory of what causes a neuron to fire which can be applied to particular cases of neuronal firing: if enough positive ions enter a neuron so that the relative internal charge exceeds -55mV, an impulse will travel down the axon of the neuron. There are two aspects to what is meant by calling this a causal explanation of firing. Firstly, it means that we can trigger a neuron to fire at will if we can bring about the required internal voltage, and we can prevent a neuron from firing if we can keep the voltage below the threshold, other things being equal. This is manipulability and control. Of course, in making a correct causal claim, it does not matter whether we actually have the ability to control whether the causal conditions obtain – only that if we *did* have that ability, we could exercise control over the effect. Thus we can make causal claims about things that are actually beyond our control, such as what causes stars to go supernova, without thinking for a minute that we have the means to effect the right conditions.

Secondly, our ability to trigger a neuron to fire at will by altering its internal charge is not affected by changes to certain other conditions. It does not matter, for instance, and so far as I know, whether the neuron to be triggered is located in the front or back of the brain, or whether the additional charge is introduced with Sodium or Potassium ions, or whether the experiment is being done at 20 or 30 degrees Celsius, or in Sweden or New Zealand. This is what is meant by “invariance under [some, not all] intervention” in Woodward’s summary of the manipulationist account.

Woodward’s account chimes with a broader conception of modern science we see in later Heidegger. In Heidegger’s most famous late essay, “The Question Concerning Technology”, his focus is – as the title indicates – technology, with technology characterized not merely as items of equipment but as a way of engaging with and relating to the world. “Technology is therefore no mere means. Technology is a way of revealing.”¹⁶⁵ Thus what Heidegger says of technology applies equally to scientific theory, which involves the same way of revealing. The revealing that modern science embodies is a different type to (or a particular modification of) the revealing of an earlier era, which Heidegger refers to as poiesis, which I will explain in due course.

[T]he revealing that holds sway throughout modern technology does not unfold into a bringing-forth in the sense of poiesis. The revealing that rules in modern technology is a challenging, which puts to nature the unreasonable demand that it supply energy that can be extracted and stored as such.¹⁶⁶

In Heidegger’s view, when we look on the world from an empirical frame of mind and attempt to determine what causal relationships there are, we are effectively looking for points of manipulability, points that could be used in the grand plan to force nature to supply us with what we need – principally, with storable energy.

This is not to say that science is wholly anthropocentric (nor that it is objectionable) – at least, this is not to say that scientists are only interested in investigating features of the world that can be of use to us in serving our practical needs. Rather, it means that the notion of cause that we employ in interested or disinterested investigation is one whose form is determined by a way of seeing focussed on supplying our need. Here is an analogy. A chess player may play for fun, be a good friend of the other player, and a pacifist, but nonetheless the logic of the chess game she plays is the logic of war. What counts as a good or bad move is determined by what promotes or does not promote victory. She plays under this scheme – she defends and attacks, and eventually wins or loses against her opponent, her ludic if not actual enemy. To play in another way is not

¹⁶⁶ Ibid., 14.
to play chess. Similarly, to describe relations between events as softer and less forceful than relations of potential manipulability and invariance is not to play the game of hard science. (This is why History properly and sensitively done is not a hard science.)

Contrast this “challenging” with poiesis, or bringing-forth. Poiesis is not a matter of manipulation and control, and does not require invariant reliability in the production of effects. The classic examples of poiesis are artistic (i.e. the artist’s bringing an artwork into being is a poietic activity), but Heidegger wishes to emphasise that the way of revealing of poiesis encompasses more, and stretches to include a way of seeing natural phenomena as well as human activity, a way that is different to the modern scientific way of seeing.

It is of utmost importance that we think bringing-forth in its full scope and at the same time in the sense in which the Greeks thought it. Not only handcraft manufacture, not only artistic and poetical bringing into appearance and concrete imagery, is a bringing-forth, a poiesis. Physis, also, the arising of something from out of itself, is a bringing-forth, poiesis…e.g. the bursting of a blossom into bloom.167

Whereas with hard causation a human agent is (at least implicitly, potentially) able to reliably force certain effects by controlling certain conditions, with soft causation the role of a human agent (when he has a role) is to nurture a largely independent and uncontrolled process to fruition. There is no guarantee that our endeavours will be successful, and these endeavours are engaged in with a respect for the natural agency (if I may call it that) that exists in the object of our efforts. We are gardeners, not manipulators.

I have spoken of the causation of art and craft and of natural events such as flowering and fruiting, and contrasted it with causation of science, but the two contrasting ways of revealing are also seen in our relations with each other. One need only think of the

167 Ibid., 10.
contrast between the use of military force and international diplomacy, or between a rape and a marriage proposal, to see a causation achieved in fundamentally different ways. In both cases we bring something about, but the manner in which it is brought about it radically different in spirit.

Most importantly for our topic, these two ways of revealing and causing are seen in the contrast between the causation of human action and the causation of bodily motions seen only as bodily motions rather than actions. When we act or perform an action our bodies move accordingly (or remain still accordingly\textsuperscript{168}), but the motion (or stillness) of our bodies is not in and of itself an action.\textsuperscript{169} Accordingly, a proper investigation of the causes of bodily motions is different to a proper investigation of the causes of actions. The first is an investigation of hard causes, and concerns the relations between neurological states and neurological events and the subsequent flexing or relaxing of muscles, which together constitute or cause the motions of the body. The second is an investigation of the soft causes of actions, which are (a) the features of the situation that the agent responds to, (b) the things the agent believes are features of the situation but may not be, and (c) the desires, purposes, goals or intentions that the agent acts on.

An agent does not have power over her own desires and beliefs (for the idea seems to require a homunculus with its own desires and beliefs), but neither do they have power over her. Rather, they are a part of, or expression of, her nature. When she acts on the basis of her desires and beliefs, she acts on the basis of a part of her nature, but her nature is more than just those particular desires and beliefs. There is always space to choose to act differently, to express a different part of herself. Whether or not the agent makes use of that freedom by surprising us, or whether she continues to behave in the way she has so far, is her choice.

\textsuperscript{168} Refraining from doing something or holding one’s body still can be actions just as much as can the actions that involve bodily movement, at least as I am using the word “action”. So, for instance, refusing to acknowledge a greeting is an action, as is standing motionless on one leg for a minute.

\textsuperscript{169} Constantine Sandis argues against the conflation of action with what our bodies do, amongst other related confluations, in The Things We Do and Why We Do Them (Houndmills, Basingstoke, Hampshire: Palgrave Macmillan, 2012). See in particular Sandis’ list of conflating views on pages xviii-xix.
One might think that what she ends up choosing is simply a function of the relative strengths of her desires, and could be explainable or predictable in virtue of this. However, what notion of “strength” would this be? Not phenomenological strength – i.e. how strong a desire feels to the agent – since we can resist our strongest felt desire. Not effectual strength – i.e. strength as judged by what the agent actually ends up doing – for that can explain nothing. Not physiological/neurological strength, because there is no scientifically legitimate way to pair physiological states with states of desire having intentional content such that empirical data could then tell us the strength of a desire. To say that the strongest desire (or set of desires) always determines what the agent will choose is thus, at best, to say what amounts to a tautology: the desire acted on is the desire acted on.

Before moving on, I must deal with an objection to the idea that there is something special about human action causation involving reasons for action that makes it different to the hard causation of the sciences. The hard causalist will point to cases where we manipulate people into performing actions, not by manipulating their bodies directly (as this would not produce an action), but by manipulating what the agent believes or wants, and thereby manipulating her actions. Suppose for instance that I want you to give me money. I could target your beliefs. If I know that you care about your child’s safety, I can manipulate you into giving me money by making you believe that your child’s safety will be endangered unless you do, even if this is not the case. Or, I could target your desires. If I can train you – say, by operant conditioning – to fear butterflies, or to desire sandalwood, I can manipulate you into paying me in my capacity as a butterfly catcher or sandalwood salesman. What we have in such cases is a causal chain: I act so that you believe or desire something, and then your desire or belief causes you, in concert with other desires and beliefs, to perform an action.

The causalist objection is therefore this. A hard causal connection is one of manipulability, where X is the hard cause of Y if by ensuring that X occurs one can ensure that Y occurs. In the cases mentioned, one ensures an action is performed by ensuring the agent believes or desires something. Therefore, the belief or desire of the agent must be the hard cause of the agent’s action. Furthermore, there is really nothing
different about the causal relationship between belief and desire and action in those cases where desires and beliefs have come about in the normal way, rather than by manipulation. Therefore, these cases are also examples of hard causation. This shows, says the causalist, that soft causation of action is simply hard causation involving beliefs and desires.

First we should set aside an unsatisfactory response to the objection. In this response, we point to the fact that our ability to manipulate someone into believing or wanting something is quite limited and the fact that we cannot be sure that the person will always act on the desire or belief in the way we wish, even if we do succeed in establishing the desire or belief we want.

I do not think this gets at the heart of the difference between soft and hard causation, since the difference is not primarily one of statistical reliability. The hard causalist would say that science often deals in causal relationships and predictions that are only statistical (e.g. in medicine), and of course work in quantum mechanics has forced us to accept that probabilities will always be a part of the science of the small. The hard causalist could also say that our ability to manipulate beliefs and desires and thereby actions might one day be greatly improved with the development of new techniques and technologies. If so, she would argue, there does not seem to be any real difference between the causation involved in human action and the causation involved in the control of other things we currently do not but might one day control well, such as the weather.

Here then is a more satisfactory response to the original objection. There is a distinction between grade-A, full-fledged, soft causation and lower grade, nominal soft causation. Grade-A soft causation involves acting for reasons in the fullest sense. The agent who is caused to act in the grade-A way is responding to real features of the environment, his action expresses a desire for something of real value, his thinking (if thinking is involved) is logical and coherent, the process by which he came to act as he did was a rational or appropriate one, and so in every sense and every way his action came about.
in the right way, makes sense and was the right thing to do. In the fullest sense, he is acting for a reason.

In contrast, cases of manipulation are cases of lower grade or nominal soft causation. Here we cannot say that in every sense the action comes about in the right way, makes sense, and is the right thing to do, nor that he is acting for a reason in the fullest sense. We will say that the action comes about in the right way in one sense, but not another, that it makes sense in one sense, but not another, and/or that it is the right thing to do in one sense, but not another. To see this, suppose I manipulate you into paying me money as a butterfly wrangler by using operant conditioning to produce in you a fear of butterflies. When you pay me the money, what you do obviously makes a kind of sense – the action is comprehensible against the background of your beliefs and newly-formed fear. However, it also does not make sense – butterflies are not to be feared. Likewise, your action came about in the right way in one sense – your belief and desire combined in the proper fashion to produce an action that makes sense given the belief and desire – but in another sense your action came about in the wrong way – your action resulted from a fear that was not produced by the sorts of things that would make it a rational fear, such as personal experience of the danger of butterflies, or being taught to fear butterflies by someone trustworthy and well-informed. Similarly, paying me the money was the right thing to do in one sense, but not another, and you had a reason for paying me the money in one sense, but not in another.

It is in virtue of their similarity to grade-A cases that nominal cases can be described in the same vocabulary, using “reason for”, “rational” and “soft-cause”. However, this should not blind us to the difference, which is that nominal cases are also similar to cases of hard causation in ways that grade-A cases are not. In a case of hard causation, such as pushing someone so that he falls, the person who is caused to do something lacks responsibility for, ownership of, and rational control over what his body is made to do. In a full-fledged case of soft causation, the agent has full responsibility,
ownership, and rational control over what she does.\textsuperscript{170} Cases of manipulation fall somewhere in between.

We can now see what is wrong with the hard causalist objection. While it is true that in many cases of manipulation a belief and/or desire is the hard cause of an action, or very much like a hard cause, since we can virtually ensure that the action occurs by manipulating the agent into having a desire and/or a belief, this is not true for cases of acting for reasons where manipulation and other kinds of defective rationality are absent – that is, for full-fledged soft causation cases. Thus soft causation is not simply a form of hard causation.

\section{The Deontology of Causation}

Deontology is the study of one’s duties, and I am concerned in this section with the duties that go with the agent perspective and the type of causation associated with the agent perspective. I will not try to argue that we ought to conceive of ourselves as agents, with all that conception entails, nor that we ought to treat others as agents. Neither will I consider whether an agent-free perspective can be broadly maintained as a way of life, in contrast to the localized adopting of the agent-free perspective by the scientist in the course of her work. Instead, I wish to simply run through a few of the ways in which the agent perspective and its associated form of causation involve duties that are not found in the scientific perspective with its associated form of causation.

We can conceive of the everyday way of seeing the people around us as agents as involving participation in a language game. One is duty bound to obey the rules of the agent language game as one is duty bound to obey the rules of rugby or German grammar, assuming one is seeing and treating people as agents, playing rugby, and speaking German respectively.

\textsuperscript{170} I do not mean to suggest that all actions which can be explained by citing beliefs and desires are actions that result from conscious choice or deliberation. Much of what we do is an expression of our rational natures yet not consciously chosen or deliberated over, and the beliefs and desires that explain an action may not be objects of conscious awareness before or when we act. One example is the professional tennis player who returns a serve. Cases of this sort can be equally good examples of the soft-causation of action from desires and beliefs.
The core idea of agent duties with respect to causation of action is that an agent owns her action – it is her action, it is her acting. Likewise, she owns her reasons for acting, as they are her own reasons. Where desires and beliefs play a part in her reasons, they are her desires and her beliefs. Ownership is connected with authority, and with other concepts from the same aut/auth word family: author, authority, authorization, authenticity, authentification, and autonomy. A person seen as an agent is seen as the author of her thoughts and actions, she has a kind of limited authority over what it is she is doing, has done, or will do (including what she is thinking, and feeling), a limited autonomy with respect to how she lives her life, and she has the potential for authenticity (a high level taking ownership of the self) and inauthenticity (a high level fleeing from ownership by disappearing into a role, which is nonetheless a choosing of a self).171

The ownership of the self carries both powers and duties. I will begin with the conventional powers of the agent seen as agent. These powers, with respect to reasons for action, are a result of what I call a mental right of reasons for acting. Recall that in an earlier chapter I defined mental rights in the following way172:

**Definition:** An agent S has the mental right of X iff S is entitled to the benefit of assumption with respect to claims she makes about X-type matters in her mental life (unless C).

\[ X = \text{e.g. intention, belief, desire, meaning, …} \]

\[ C = \text{conditions that, if sufficiently satisfied as a group, would remove or shift the benefit of assumption.} \]

So, specifically, we can define a mental right of reasons for acting (MRA) as follows:

---

171 See Chapter 3.
172 See Chapter 4.
Definition: An agent S has the mental right of reasons for acting iff S is entitled to the benefit of assumption with respect to claims she makes about her reasons for acting (unless C).

The conditions C will be explored in a moment. For now, let us simply say that unless we believe those conditions are met, when someone explains her reasons for acting (which could be reasons why she did what she did, reasons why she is doing what she is doing, or reasons why she will do what she will do), we are obligated to treat what she says as settling the matter. If we do not treat her explanation as settling the matter, we have ceased to conceive of her as an agent.

I have created the concept of a mental right because I think it accurately models how we treat each other in the everyday mode of interacting with each other. The model can be accurate without requiring that agents realise their behaviour follows this pattern, just as a rule of descriptive grammar can accurately model the verbal behaviour of members of a language community without those members being aware that they follow the rule.

A community whose members treat each other as having MRAs is one which provides a space for persons to be persons, and for agents to be agents, as opposed to mere things. Of course, a community which does not effectively respect the MRA is not a community at all. A community which respects the MRAs of only some members is conceivable. One can imagine, for instance, a community in which women are routinely treated as less than full agents and persons by taking their explanations of their own actions as mere hypotheses about their internal dispositions.

The most important condition in C relates to duties that the agent must fulfil in order to be taken as holding an MRA with respect to her statement of her reason for acting. We are familiar with duties that go with making claims. For instance, whenever we express a claim to know something, we are entitled to make the claim only if we have done our duty with respect to determining and checking the truth of the claim. A person who says “I know there are 10 available rooms in the motel”, but who isn’t ready to back up her
claim with the right kind of support (e.g. “I called the manager to check”) is taken to have misused the word “know”.

Mental-right duties concern an integrative task: the role of being an agent involves a responsibility to ensure that, as much as possible, what I say about my thoughts, feelings and reasons makes sense given my actions and the other things I say of myself and the world. This is a requirement of agentive coherence, different to a requirement to check the truth of a belief by conducting an investigation before making a knowledge claim. In terms of the mental rights schema, where we have the right “unless C”, this means that C is a condition of agentive incoherence. One’s claim should be (and normally is) accepted unless one’s expressed reason for acting fails to make sense alongside the other things one does and says.

There is a small problem that must be dealt with before the account of mental rights is acceptable, which will require that we add another condition to C. To see what the problem is, suppose my friend Frank tells me this:

My reason for eating 2L of banana ice-cream everyday is that eating banana-rich foods will help me lose weight.

Even if I am inclined to believe that Frank is being frank and honest, because what he says meshes well with other things he says and does (so, for instance, I know Frank is not aware that ice-cream of any flavour is fattening), I still cannot accept that what Frank says is true – at least, not exactly. I cannot accept that his reason for eating the ice-cream is that eating banana-rich foods will help him lose weight, because I know not all banana-rich foods are like that. Reasons are factive. That is, any sentence that says “the reason why this happened is…” must finish with something that a person who accepts the sentence believes is a fact. Of course, in Frank’s case, we make a small adjustment and say his reason for eating the ice-cream is that he thinks eating banana rich foods help will him lose weight. In this way we preserve the spirit of the mental right by making the smallest possible adjustment to the reason claim. One way to incorporate this into the definition of the MRA is to add a second condition to C: The
mental right exists unless (a) what the agent says does not make sense given what we know about the agent or (b) what she says includes or entails something that is not true. When it does involve or entail something untrue, the agent has a mental right to establish the beliefs or desires driving the action, but not the facts driving the action.

We now turn to the deontology of hard causal claims and science in general, where things are less complicated and so will be described only briefly. The scientific view of causation (i.e. of hard causation) has nothing like the deontology of mental rights. In science, we play a much different game. The objects of description – events and their causes – are not owned in any sense by the scientist making an observation or making a causal claim. The scientist may be unique in virtue of being the only person to observe a particular event or series of events, but this is a contingent and unimportant type of uniqueness, and does not constitute ownership of the events or their causal connection. If anyone else had been in the same position she should have observed the same thing. Similarly, causal claims in science require that particular cases follow a general law, and such laws are established (or refuted) by the collective efforts of the scientific community as a whole, not established by one scientist with authority.

The deontology of science, if there can be said to be one, involves a number of obligations, of which I will list only a few. (1) Scientists are obligated to be honest, (2) scientists are obligated to remain detached until the weight of evidence supports a conclusion, (3) scientists as a group are obligated to question and test the claims of others, and (4) scientists are obligated to avoid ad hominem argument and argument from authority when doing science. One more aspect of the deontology of science is (5) scientists are obligated to express their core claims (if not all claims) in testable form – in particular, they must remain within the penumbra of vagueness of language and thereby make claims that could be falsified.173 This deontology has no place for first person privilege, and no concern with the kind of making sense that is agentive coherence.

173 See Chapter 1, footnote 11, and Popper, The Open Society, 21-22.
VI Conclusion

To finish, I’d like to suggest a relationship between the ordinary language use of “cause” detailed in Section II and the ontologies and deontologies of soft and hard causation outlined in Sections IV and V. I’ll also return to the pro-reason question and causal solution.

In Section II we saw that when “cause” is used in the explanation of things people do, the actions or motions explained are almost always done either (a) non-intentionally/unintentionally, (b) reluctantly, (c) as if the agent had no choice, (d) with diminished responsibility, or (e) from lack of self-control. It should be apparent that there is a close connection between (a) to (e) and the ontology and deontology of hard causation. That is, in using “cause” to explain what is done in (a) to (e) type cases, we indicate that we regard the action as resulting in a way that is either hard causal, making it actually not an action at all but a motion (see some cases of (a)), or as sharing important similarities with hard causal cases (true for other cases of (a), and (b) to (e)).

In fully fledged soft causal cases we have full authorship and ownership of an action and the reasons for which it is done. In hard-like cases, we still have some authorship and ownership, but our authorship and ownership is portrayed as reduced, limited, strained or troubled, such that we feel like or seem like the objects of internal or external forces rather than agents in control of ourselves and our lives.

Thus, we can accept the causal solution to the pro-reason question, but only under a soft interpretation. What makes maintaining this interpretation difficult, but not impossible, is that our ordinary use of “cause” is linked to hard causation, so it can be natural to read or misread the causal solution as involving hard causalism. Under that interpretation, human action and personhood are demoted to something more mechanical, less free, and less responsible, than the truth.
7 A Defence of McDowell-Style External Reasons

I Introduction

Bernard Williams published his seminal article “Internal and External Reasons” in 1979. A large literature has grown discussing the thesis Williams outlined in that article and expanded on in subsequent articles. 16 years after the initial article, John McDowell wrote a contribution titled “Might There Be External Reasons?” to a festschrift in honour of Williams titled World, Mind, and Ethics in which he challenged Williams’ thesis, and Williams wrote a reply to McDowell published in the same book. The purpose of this chapter is to examine both McDowell’s challenge and Williams’ reply.

Given that there are any number of more recent contributions to the debate, I should say why I want to return to considering an exchange that took place 20 years ago. The motivation comes from my belief that McDowell’s contribution is underappreciated and unfairly ignored in the literature – perhaps because it is thought that Williams’ reply conclusively refuted McDowell. I would like to show that this is not so. McDowell’s essay is typically either not dealt with at all, or roundly dismissed. To give just one example of the former, Richard Joyce spends an entire chapter of his 2001 book The Myth of Morality arguing, along with Williams, that external reasons do not exist, without mentioning McDowell or McDowell’s way of making sense of external

174 Williams, “Internal and External Reasons.”
176 John McDowell, “Might There Be External Reasons?”
reasons. When McDowell’s position does come up for explicit discussion by other authors (as in recent articles by Christopher Cowley and John Brunero), it is briskly rejected. Cowley claims McDowell’s position improperly ignores the particularity of agency, (see Section V of this chapter), and Brunero claims that McDowell had misunderstood Williams’ core argument for his position, (see Section IV), such that McDowell’s counterclaims miss the point. Again, I would like to refute these claims against McDowell.

Even though Williams offered arguments to refute McDowell, he also hinted at how he thought McDowell’s position might be developed in a defensible direction. Ultimately, Williams thought the prospects for developing such a view “discouraging” given “the problems that lie in that direction” but he finished his reply to McDowell by recognizing that “there are no doubt many undiscussed considerations to be pursued here.” My aim is to discuss some of the undiscussed considerations, and thereby say something in defence of McDowell-style externalism.

In Section II, I will explain Bernard Williams’ position on reasons (and throughout this paper by *reasons* I mean *reasons for acting* rather than *reasons for belief*), and in Section III I will explain McDowell’s response to Williams. In Sections IV, V and VI, I will outline and argue against three objections that Williams makes against externalism in general or externalism of McDowell’s sort, one of which is repeated in Brunero’s article and one in Cowley’s.

My overall aim is modest. I will not try to prove the existence of external reasons, but simply try to show that there is space for such reasons – that there indeed “might be” external reasons, as in McDowell’s title. I will not show that the internalist is making a demonstrable mistake in rejecting external reasons – only that he is mistaken if he thinks he has a convincing argument for his position and against the externalist’s.

II Williams’ Internalist Position

To begin with, consider two situations, adapted from cases Williams considers, and two related questions:

(1) A young woman Sarah has become depressed and sees no point in continuing her life. She is dangerously ill, but when we put medicine before her, she refuses to take it. Under which conditions can we rightly say that she has a reason to take the medicine?\(^\text{180}\)

(2) A married man Bill treats his wife terribly, for he does not care about her at all. Under which conditions can we rightly say that he has a reason to treat her better?\(^\text{181}\)

Williams argues that we can only say that a person has a reason to do something if there is a sound deliberative route from her current subjective motivational set S to the action in question.\(^\text{182}\) Such reasons are called “internal reasons” (as opposed to “external reasons”), and are the only reasons Williams believes exist. For the cases of Sarah and Bill, that roughly means that unless Sarah has some motivation furthered by taking the medicine, she has no reason to take it, and that unless Bill has some motivation furthered by treating his wife well, he has no reason to treat her well. By contrast, a believer in external reasons (and I am one) would say that that Sarah and Bill have (or at least might have) reasons to do the respective actions whether or not their motivations would be served or furthered by doing so.

Williams’ position, whether right or wrong, certainly has its strengths, one of which is the sophistication of the notions in play, and understanding sufficiently what he means requires that we explore the notions of a sound deliberative route and a subjective motivational set in a little more detail. A deliberative route is a route in practical reasoning that begins from a set of beliefs and a set of motivations, and ends in an act. What makes a route sound is that, firstly, the beliefs that the reasoning begins with are true, and secondly, that the moves from the beliefs and motivations to the action (or at

\(^{180}\) This example is adapted from one in Williams, “Internal and External Reasons,” 105-106.

\(^{181}\) This example is adapted from one in Williams, “Internal Reasons and Obscurity,” 39-40.

\(^{182}\) This is the mature formulation of Williams’ position. Ibid., 35.
least a motivation to perform the action) can be accomplished by rational deliberation. Importantly though, Williams’ conception of the sound deliberative route (or SD-route) is not merely one of simple instrumental rationality. An exercise of imagination, for example, can play a part in one’s following a sound deliberative route.

As for the other important notion, the agent’s subjective motivational set is not simply a set of desires, though the word “desire” could be used as philosophical shorthand to describe any member of the set. A set can include “dispositions of evaluation, patterns of emotional reaction, personal loyalties, and various projects, as they may be abstractly called, embodying commitments of the agent.”

We are not to assume that S includes moral or long-term prudential concerns, for while Williams believes everyone rational is motivated to be well-informed about any facts relevant to a decision, argument is needed, and he believes none has been provided, for saying that everyone rational is motivated by moral and long-term prudential concerns. Having said this, the desires in S need not all be egoistic – one might have the desire to help strangers in need or the desire to follow the ten commandments. Likewise, the desires in S need not all focus on short term gain – one might have the desire to be healthy in one’s old age. However, Williams does not believe there is necessarily anything irrational about a person whose motivational set consists only of egoistic desires or short-term desires.

A final important aspect of Williams’ position is that he recognizes that there will be some hard cases, where there is no determinate answer as to whether someone has a reason to do something or not. He explains this by saying there is “a continuum from rational thought to inspiration and conversion,” so whether some intermediate movement of thought on that continuum ought to be included under his notion of a sound deliberative route may be impossible to say.

---

183 Williams, “Internal and External Reasons,” 105.
184 Williams, “Internal Reasons and Obscurity,” 37.
185 Williams, “Internal and External Reasons,” 112.
186 Ibid., 110.
III McDowell’s Response

John McDowell’s response is given in the article “Might There Be External Reasons?”\(^{187}\) McDowell holds that when we understand what we have reason to do, we see things rightly, as an ideal deliberator (or phronimos\(^{188}\)) would. With respect to Williams’ notion of a sound deliberative route, McDowell argues that coming to see things rightly is not always accomplished or accomplishable by sound deliberation, or not sound deliberation alone. Firstly, we can be raised into seeing things rightly, and the process of being raised is not one achieved or achievable by deliberation alone. Secondly, we can undergo an adult conversion, which again is not achieved or achievable by deliberation alone. What matters is that we somehow come to see things rightly, and can thus deliberate soundly, not that we come to see things rightly by sound deliberation. Therefore, an agent may have a reason to do something even if there is no sound deliberative route to doing the action. This would happen when the agent is not seeing things rightly, when his or her perspective is immature or distorted and when rational deliberation of the type Williams’ allows is insufficient to overcome this immaturity or distortion.

In challenging Williams in this way, McDowell is careful not to over-commit himself to any specific metaphysical or ethical theory as to what constitutes being well brought up:

> I am using the notion of proper upbringing only to defuse the threat of metaphysical peculiarity [in the notion of considering matters aright], not as a foundational element in some sort of ethical theory – as if we had independent access to what counts as a good ethical upbringing, and could use that to explain ethical truth as a property enjoyed by the judgments that a properly brought up person makes.\(^{189}\)

This caution should be remembered. Some of Williams’ objections seem better targeted at a more definitive ethical theory than McDowell means to offer (i.e. the impersonality objection (Section V), and the objectivity objection (Section VI)).

\(^{187}\) McDowell, “Might There Be External Reasons?” 95-111.

\(^{188}\) This is Williams’ way of formulating McDowell’s position (in Williams, “Replies,” 186-94) but it is, I think, a fair characterization.

\(^{189}\) McDowell, “Might There Be External Reasons?” 101.
Nonetheless, it is worth thinking a little about the nature of ethical upbringing, and how it differs from a process of rational deliberation, so that – as McDowell says – his position does not seem too peculiar.

Here is a crude sketch of ethical/moral development and ethical/moral reasons which I think McDowell would accept.\(^\text{190}\) A child has certain desires in her subjective motivational set S: desires for physical pleasure, desires to avoid physical pains, a desire to please rather than displease her parents, a desire to fit in with how people around her behave, a desire to acquire useful things, a desire to be liked by her peers, and so on. None of these desires is essentially a desire for morality. However, these desires can be used to give her a reason to do the moral thing on a particular occasion. For instance, her desire to please her parents can be used to get her to do those modest “moral” tasks we require of a child – to express gratitude for a gift, to refrain from rudeness and bullying, to contribute to household chores, and so on. In employing such motivational levers, a parent is like a farmer, tilling and planting the ground in the hopes of a future harvest: a moral character. I call these motivational levers *tilling motivators*, both because they till the ground until character is developed, and because they till the ground *till* character is developed, when they are no longer needed. Tilling motivators are often, though not always, threats.

*If you don’t let your brother have his turn on the slide, I will be very angry with you.*

*If you steal cake from the pantry, you’ll have to go to time out.*

Repeated utilization of tilling motivators leads the child to form a range of habits, and it is the inculcation of these habits, along with explicit instruction, the learning of a moral vocabulary, and practice in the giving and defence of reasons for action, that promote a change in the child’s motivational set S, so that she comes to care about a number of things she did not before – many of which can be grouped together by saying that she now cares about living a moral life. The more she reaches moral maturity, the more the

she can be motivated by reasons that are truly moral, and the more the need to utilize tilling motivators drops away. Now she says to herself:

I would never shoplift – that would simply be wrong.

It’s both my duty to care for my ailing grandmother, and actually an honour (if not always a pleasure), so that’s what I’ll do.

If the former motivators were tilling motivators, these are harvest motivators.

In the process of raising a child, we will often express reasons to the child that would access harvest motivators in the morally mature:

Billy, you mustn’t steal. Stealing is wrong.

However, if such reasons are to have any motivational effect on a morally immature child, they must latch onto tilling motivators. This typically occurs via the tilling motivators of keeping a parent happy, and keeping out of trouble. For example, a child might think, effectively: “I have a reason not to steal because mummy says that stealing is wrong, and mummy will be angry if I do something wrong, and I don’t want mummy to be angry.”

Using this terminology, we can now see another way of expressing the debate between the internalist and externalist. The internalist maintains that an agent has no reason to do the moral thing unless (A) the harvest of a moral character has been achieved, so that she has a harvest motivator in her S to do it, or (B) a tilling motivator in S supports doing it. An externalist maintains that she can have such a reason, whether or not the agent currently has a tilling or harvest motivator in her motivational set that the reason can latch onto. The externalist also maintains that it is only with a mature moral character that the agent can see and understand that she has this reason. It is a reason that requires a certain sort of psychological makeup to see and understand, but not a certain sort of psychological makeup to exist.

191 Williams makes similar points about the role played by “the desire to avoid hostility” and the more nearly ethical “desire to be respected by people whom, in turn, one respects” in “Internal Reasons and Obscurity,” 41-2. Williams thinks that supposedly external reasons claims and related statements of blame may create new reasons for an agent by being directed at her and latching onto these sorts of desires. He calls this a proleptic mechanism. I agree that we sometimes motivate each other on the basis of such desires, but do not agree that the reasons we invoke were nonexistent prior to the claiming and blaming.
IV The Possible Explanation Objection

We will consider Williams’ three main objections to McDowell-style externalism, which I call the possible explanation objection (covered at length in this section), the impersonality objection (covered in Section V), and the objectivity objection (covered briefly in Section VI).

Firstly, Williams argues that an agent’s reasons must be able to explain and motivate her actions if they are to be her reasons, but so-called external reasons cannot do so without being unveiled as, or depending on, internal reasons that are doing the real explaining and motivating work. Thus such so-called external reasons are either not reasons, or are internal reasons.\(^{192}\) Williams’ argument rests on the premise that reasons must be able to explain and motivate action. Here is how Williams puts it:

[I]f there are reasons for actions, it must be that people sometimes act for those reasons, and if they do, their reasons must figure in some explanation of their action.\(^ {193}\)

This is somewhat vague and indirectly related to our question – it speaks of there being reasons for action, and of people acting for those reasons, rather than of some particular agent having reasons and those reasons explaining her actions. For the argument to work, Williams needs something firmer. Here is what I think he needs:

If an agent has a reason R to φ (for some action φ), then it must be possible for her to φ for reason R, and if she does so (that is, if she φs for reason R), the fact that she has this reason to φ must be able to figure in an explanation of her action.\(^ {194}\)

Call this premise the link to explanation requirement, or LER.\(^ {195}\) The question that concerns us is whether LER is true, and if so, whether it implies reason internalism.


\(^{193}\) Williams, “Internal and External Reasons,” 102.

\(^{194}\) I say “able to figure in an explanation” because there may be explanations of her action (seen, arguably, as a cluster of movements of her body) in terms of non-intentional states of affairs (such as neurological states), and such explanations do not involve reasons of the sort we are interested in. I suspect this is why Williams wrote that reasons “must figure in some explanation” rather than “figure in the explanation”.

\(^{195}\)
What makes deciding this problematic is that there are different ways to interpret the LER, because there are different ways of interpreting the crucial phrase “‘it must be possible for her to φ for the reason R’.” Even if we were to gloss the sense of possible as “within the agent’s power,” we have a problem of interpretation. In another work, Williams made precisely this point:

The question of what counts as in the agent’s power is notoriously problematical, not only because of large and unnerving theories claiming that everything (or everything psychological) is determined, but also because it is simply unclear what it means to say that someone can act, or could have acted, in a certain way.  

With this in mind, we need to tread carefully. One interpretation of the crucial phrase, one which produces a version of the LER that I do not accept, ties what is possible to the presence of elements in the agent’s current S that could ground an interest in φ-ing for reason R. Call this Interpretation 1. According to Interpretation 1, the fact that there is nothing the agent wants that could be gained by φ-ing means she cannot φ for any reason. If she does φ, it will only be by mistake. Of course, if we grant LER understood in these terms, the internalist conclusion quickly follows.

However, there is another interpretation, Interpretation 2, or in fact another two interpretations, 2A and 2B, according to which it is possible for the agent to φ for reason R even though her current S will not support doing so. The first of these (2A) is based on the fact that it is possible for the agent’s S to change. If her S were to change so as to support φ-ing in the right way, she could φ for reason R. We can agree with Williams that “nothing can explain an agent’s intentional actions except something that motivates him to act” but this only means that S must change in order for the agent to act in the right way. Of course, this interpretation is meant to give an account of a
necessary but not sufficient condition on having a reason. The mere fact that it is possible for an agent to come to care about something new does not mean she has reason now to act accordingly. Otherwise, we would all have reason to do almost anything at all.

Perhaps the change to S comes about because the agent has come to believe she has reason to φ. To quote McDowell:

Consider an agent who is not motivated by an external reason. It must nevertheless be true that some consideration constitutes a reason for him to act in a certain way; his not being motivated by it is a matter of his not believing, of the consideration, that it is a reason to act in that way. If he came to believe that, he would come to be motivated.\textsuperscript{198}

However, it should not be forgotten that McDowell accepts that the change in what the agent believes may depend on, or be necessarily accompanied by, a change in what she cares about. If we forget this, it will look suspiciously like rationality alone is driving the change, and that the person who does not see the reason she has to φ is therefore simply irrational. McDowell has explicitly rejected this view.\textsuperscript{199}

Here is an analogy to explain Interpretation 2A. Suppose we ask whether it is possible for Timmy, who is 6 years old and 3 feet 9 inches tall, to slam dunk in basketball. If we are meant to imagine that Timmy’s current height is fixed, then the answer is almost certainly “no”. If instead we are allowed to imagine that Timmy can grow, and of course we expect that he will, then the answer may well be “yes”. Indeed, we might say to Timmy “You can do it. You just have to wait till you’re taller.” Of course, if the question is what he can do now, with his current height, then that is a different story.

Similarly, for the case of the moral novice who lacks moral motivations but who could be brought to have these motivations, we can say that it is possible for her to act for moral reasons. She just needs to grow up (or convert) first. Of course, she cannot φ for those reasons now unless she converts now, but that is a different point, and does not mean she has no reason to φ.

\textsuperscript{198}McDowell, “Might There Be External Reasons?” 98.
\textsuperscript{199}Ibid., 103.
If the internalist finds such a rejoinder to Williams’ argument suspicious, we can remind her that she likewise takes an interpretation of LER with space for the agent to change. That is, the internalist allows that the agent may come to learn facts about the situation that she was ignorant of or in error about, and then act accordingly. For instance, suppose I am terribly thirsty, but I incorrectly believe the liquid in the glass before me is peroxide rather than water. Internalists and externalists agree that I have a reason to drink the water, given that I am thirsty and it will quench my thirst. But is it possible for me to drink this water for this reason? On one interpretation, one which holds my state of knowledge as fixed, it is not. If I do not learn that this is water, I cannot drink it for the reason that it, being water, will quench my thirst. This is not the interpretation any sensible internalist accepts. Instead, they operate with an interpretation of LER that allows for correction of factual error and removal of ignorance, under which interpretation it is possible for me to drink the water for the reason that it is refreshing water. Internalists and externalists only disagree about whether our potential to lose our poor characters, our immaturity, and our moral ignorance and error, should be regarded as relevant to what is possible in the same way as our potential to lose straightforward (that is to say, empirical) ignorance and error.

If we accept LER under Interpretation 2A, there is a limit on who has external moral reasons. If we think, as perhaps Aristotle did, that a man can be beyond redemption (i.e. beyond living a good life) because his upbringing has come to an end without producing a character of the right type, so that even a late conversion to a virtuous outlook is out of the question for the most part, then we cannot say that it is possible for him to come to believe he has reason to φ, at least in one sense of “possible”, and thus we cannot say that he has reason to φ.200 Likewise, and less contentiously, someone who was brain damaged enough to be un-teachable would have no moral reasons.201

200 A possible example of someone badly raised and beyond re-education and redemption is that of the convicted felon Robert Harris, detailed in Gary Watson, "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme," in Responsibility, Character and Emotions, ed. Ferdinand Schoeman. (Cambridge: Cambridge University Press, 1988), 256-286.
201 It should be noted that interpretation 2A of LER does not entail the ludicrous claim that babies and foetuses have moral reasons, even though babies and foetuses might one day grow up to accept moral reasons and act on them. In LER the possibility of φ-ing for reason R is a necessary, but not sufficient, condition for having reason to R.
Many externalists will be unhappy with letting all hopeless cases off the moral reasons hook. One option they can take up would be to reject LER and maintain that reasons need not be potentially explanatory to exist. However, to reject any version of LER and still maintain that the agent in question has the reasons may look at the very least like a misuse or unnatural use of language and at the worst as pure bluff. We might instead think it better to go along with Williams and disentangle talk of reasons from all the other things we might say of the agent, such as that he is vicious, cruel, evil, etc.\textsuperscript{202} Still, it also seems odd to say, as Williams would, “You are a cruel, heartless so-and-so for φ-ing, but you have no reason to behave otherwise” and more natural to say “The fact that φ-ing would be cruel and heartless is a reason anyone has, including you, not to do it. You might feel differently but you’d be wrong, and that’s what makes you a cruel heartless so-and-so.” If we feel this is right, we will need to look for an alternative interpretation of LER so as not to be forced in either of the other unnatural directions, where reasons are detached from possible explanation, or detached from talk of vice.

A second externalist interpretation of LER is possible, 2B, which may not let all hopeless cases off the hook, and may allow the externalist to say what she feels is natural. In Interpretation 2B, we explain the sense of possibility involved in LER in terms of what it is fair, or appropriate, or reasonable to expect or demand from an adult, rational agent.

Crucially, in this account of the sense of possibility in LER, I do not provide a necessary condition for \textit{all} cases of having a reason. I only intend to cover cases where the reasons involved are moral (and, perhaps, prudential), and where they involve obligations, rather than reasons to do things which it would be morally (or prudentially) good to do, but which are not obligatory. Moral reasons of the obligatory type are unlike more mundane reasons, such as the reasons we have to travel to foreign countries as tourists. For instance, since there are many wonderful things to see and do in Thailand, there is in that sense a reason (or many reasons) for me to visit Thailand. Nonetheless, we should not say that it would be \textit{fair to demand} that I go to Thailand. There is, absent special circumstances, no moral obligation to go to Thailand.

\textsuperscript{202} See Williams, “Internal Reasons and Obscurity,” 39-40. However, I am not alone in thinking it is a mistake to disentangle talk of reasons from these types of criticism. See Thomas M. Scanlon, \textit{What We Owe to Each Other} (Cambridge, MA: Belknap Press of Harvard University Press, 1998), 367.
So, suppose Bill, the cruel husband mentioned in the introduction, has no dispositions that would be served by treating his wife kindly. We can maintain that it is nonetheless possible for him to treat her kindly (for the reason that she is both his wife and a human being) even if he is a hopeless case beyond redemption, simply because it is fair to expect that he treat her kindly for those reasons, and if he were to treat her according to this reason (which is as much as to say, if he were a better person), that reason would be the explanation of his action. Bill cannot maintain that treating her kindly is a physical impossibility, as Timmy might complain that slam dunking is a physical impossibility for someone at his age. Demanding a slam dunk from Timmy at age 6 would be unfair, but it is not unfair to expect kindness from Bill. Of course, one does not expect, in the sense of predict, kindness from Bill. In fact, one may be almost certain that his character is beyond kindness. Nonetheless, we can judge him as guilty of failing to do and be as he should, and in this sense he fails our legitimate expectations, and fails to do what is possible. This is the interpretation of LER that I prefer when obligatory moral reasons are involved.

It is also an interpretation that takes on board Williams’ observation that what one has reason to do may depend on one’s stage of moral development. A novice may have reason to avoid situations full of temptation that a phronimos would not. As Williams remarks, the “homiletic tradition… is full of sensible warnings against moral weight-lifting.” If the LER is understood as requiring that in order for an agent to have a reason to φ, it must be fair to expect the agent to φ, we can allow that sometimes immature agents have different reasons to mature agents in virtue of their immaturity, for their immaturity affects what it is fair to expect of them. We will return to these points in the section on the impersonality objection.

One possible stance to take to moral matters is to accept the LER under both interpretations 2A and 2B by holding that there is a certain connection between them. That is, one can hold that if it is not possible for someone to change or convert to wanting to φ, for some action φ, it is not fair to expect her to φ. If φ-ing is typically the moral thing to do (say, graciously congratulating a victorious opponent after a tennis match) but she is not yet anywhere near having the right psychology to do it, or do it

---

203 Williams, “Replies,” 190.
properly, we may say that in her case she has no reason to do it, and that she is not acting immorally (but at worst immaturesly) by neglecting to do it. Of course, this stance is offered only as an option worth considering, and may be unattractive if one wants to leave space for criticizing those beyond redemption. I leave the issue open.

A more serious issue is whether the reasons internalist can argue that, for the LER under Interpretation 2B to be true, the LER under the internalist Interpretation 1 must also be true. The thought would be this: if it is not psychologically possible for the agent to act for reason R in the sense that there is no sound deliberative route to her acting for reason R, then it is also not fair to expect the agent to act for that reason. This would mean the LER is only satisfied if R is an internal reason, and thus that the LER, if true, implies reasons internalism.

I will not deny that this is a coherent possibility philosophically. However, it constitutes a substantive moral position, one that I reject and it seems Williams might also have rejected, given his comments about the applicability of vice terms such as “ungrateful, inconsiderate, hard, sexist, nasty, selfish, brutal” even in cases where there are no internal reasons to act otherwise. 204 The opposite position, that we can legitimately expect good behaviour from and criticize someone even if to exhibit such good behaviour is outside her character as it stands, is also a coherent position philosophically, and one that I think is closer to the grammar of our moral concepts.

So far I have argued that the LER can be satisfied without resorting to reasons internalism. In this regard, it is worth examining an objection to McDowell’s position from internalist John Brunero. The internalist will resist the idea in Interpretation 2A that what someone has reason to do now should take into account how S might change by subsequent upbringing or conversion, and resist the idea that it might be fair to expect certain behaviour from someone just because this sort of change is possible, as in one view of 2B. Brunero considers whether someone who was uninterested in philosophy would have a reason to read McDowell’s papers, and rejects the idea that the possibility of her converting so as to have such an interest meant she did indeed have such a reason whether or not she converted.

If the conversion involved the attainment of something like a desire to keep up with interesting work in philosophy, there now would be a desire in S rationally related to his reading McDowell’s papers. And so the internalist would happily concede that there is now a reason for him to read McDowell’s papers. He’ll just insist, against the externalist, that this reason was not there all along. It only came into existence when the relevant element in S came into existence.\footnote{Brunero, “McDowell on External Reasons,” 28.}

That is, what Brunero will not allow is that the agent had a reason “prior to the change in [her] S”\footnote{Ibid., 28.}. He goes on:

It does not matter to the internalist how this change in S occurs, whether by rational deliberation (from other desires) or by conversion. To be clear, it does matter to the internalist whether the agent can come to be motivated to \( \varphi \) only through a process of conversion or, alternatively, whether there is also a sound deliberative route proceeding from the motivations in his S, since, for the internalist, if the former is the case, an explanation could not be given which makes sense of the agent’s acting on the reason, while if the latter is the case, such an explanation could be given.\footnote{Ibid., 28.}

Brunero concludes that McDowell’s attempt to defend reason externalism by way of alternatives to sound deliberative routes is still vulnerable to the possible explanation objection, and thus that McDowell has missed the point of Williams’ argument.

Allow me to respond on McDowell’s behalf. The idea that reasons come into existence or disappear from it as desires (and other elements of S) come into existence or disappear, is quite plausible for a large range of everyday cases, but the question is whether this is true of all cases of having a reason. The plausible cases are ones where what is at issue is a matter of either personal preference or decision. If my dislike of the taste of chocolate ice-cream changes to an enjoyment of that taste, it seems fair to say I have come to have a reason to buy it that I did not have before – namely that I now enjoy its taste. Likewise, if there are 10 countries I would like to visit, all with their individual and incommensurable attractions, and I finally decide – because time is
running out – to visit France, I now have a reason to buy a plane ticket to France that I didn’t have when I was undecided – namely, that I have decided to go to France.

What separates the internalist and externalist is whether moral reasons are to be regarded in the same way. Internalists see moral reasons (if they exist at all) as equally dependent on the actual state of S as reasons of personal preference and decision, and thus effectively as reasons of personal preference and decision. An externalist accepts the existence of both types of reasons – the personal and the moral/universal, but sees moral reasons as special in that they typically do not depend on S.

To explain her position, the externalist can argue that the changes in S needed to support moral action also constitute or are necessarily accompanied by improvements in the discernment of matters pertaining to morality, such as discernment of what truly matters in life, of the value and sanctity of human life, of the rights we all share as human beings, as well as of details about how to act rightly in specific situations. One does not come to care about human life without coming to believe that human life has value, and the reverse is true as well, if one’s belief is authentic. As such, coming to care about human life (a change in S) means coming to believe that human life is valuable, and that is a transition to seeing the world aright, as McDowell would say. These things that one now sees aright were true all along – human life always had value, whether one saw it or not. Thus one always had a reason to treat people accordingly, whether one saw that or not. In general, moral reasons existed all along, and we have managed to satisfy McDowell’s requirement:

[W]e can preserve the [idea of an] external reason if we can make sense of something which would be true throughout the transition.209

What remain “true throughout the transition” are the familiar facts of morality. Crucially, Brunero has given us no reason to assimilate all reasons to the model of the personal. He is correct that, on that assimilating view, there are no external reasons, but

208 See Chapter 3’s discussion of authenticity.

209 McDowell, “Might There Be External Reasons?” 98.
the externalist is free to resist the assimilation and maintain that externalist reasons exist.

V  The Impersonality Objection

Even if we suppose that the issue of LER satisfaction is resolved, Williams has another major concern. Williams interprets McDowell as trying to understand “(R) A has a reason to $\varphi$” in terms of “(C) If A were a correct deliberator, A would be motivated in these circumstances to $\varphi$.” But (C), says Williams, is simply an application of the following to the particular agent A: “(G) A correct deliberator (i.e., a phronimos) would be motivated in these circumstances to $\varphi$.”

According to Williams, the problem with this analysis of (R) is that it “does not make a statement distinctively about A at all”, it does not relate “actions to persons, but types of action to types of circumstances”, and it is “impersonal, relative to particular agents.”

Something quite similar to this criticism is repeated in an article by Christopher Cowley, who I quote at length in order to fully convey the nature of the concern:

> It is the agent who has to act; and not just any agent, or any member of a given class of agents, but, simply, this agent. This is not a trivial detail; the agent is not a datum. By shifting the focus to (a third-personal depiction of ) the act, the externalist omits the agent’s particular first-person perspective on the contemplated actions, thereby rendering him passive, and ignoring the fact that he is necessarily and inescapably embodied in his particular S: the particular S is what unifies his experience into the first person as he uses it in speech and thought. He does not acknowledge an external reason and thereby the requirement that action X be performed by someone like him; instead, he acknowledges the internal reason and thereby the requirement that he (i.e. from his point of view, ‘I’) perform the action. So the very concept of agency underlying any discussion of reasons for action must be the particular agency of a particular person; an action does not get done until somebody does it – and chooses or decides to do it, or intentionally or willingly does it, on the basis of (but not uniquely determined by) the reasons available to and adopted by him at the moment of performance, where such a choice is internal to his point of view on the situation.

---


211 Ibid., 189-90.

Though this passage is informative, it will be clearer to tackle the problem using Williams’ characterization, which concerns the relation between (R), (C) and (G). Williams does not interpret McDowell as believing that (R), (C) and (G) are equivalent, only as believing or as effectively being committed to believing that “(R) gets its content from (C), and (C) gets its content from (G).”213 This thought is Williams’ attempt to formalise McDowell’s position, and not something to which McDowell has explicitly committed himself. I suspect McDowell’s position is subtler than such an equation or explication of content, and I will try to introduce some subtlety to the position later in this section when I draw on an idea from Michael Smith. For now though, let us accept Williams’ interpretation of the relations between (R), (C) and (G).

It is not immediately clear why the impersonal nature of a (G)-based (R) should be a problem. To quote T. M. Scanlon:

The universality of reasons judgments is a formal consequence of the fact that taking something to be a reason for acting is … a judgment that takes certain considerations as sufficient grounds for its conclusion. Whenever we make judgments about our own reasons, we are committed to claims about the reasons that other people have, or would have under similar circumstances.214

That is, the reasons an agent has are not distinctively about the agent who has those reasons – or at least, not typically. It is one of the attractive features of a belief in external reasons that it respects this fact. Though there are some cases in which individual preferences do make the difference between having a reason and not having one (such as whether one enjoys a particular literary genre and therefore has reason to buy a particular book) there is something wrongheaded about the notion that in general it is always possible that I might have a reason to do something which another person in the same circumstances would not, merely because of differences in what we desire. Instead, we can argue along with McDowell that we are all properly evaluated by how we respond to grounds for acting in terms of what a phronimos would take as grounds for acting, and would be motivated by, and therefore something that is a reason for one

213 Williams, “Replies,” 190.
214 Scanlon, What We Owe, 74.
individual is characteristically a reason for another individual in similar enough circumstances, whether he can recognize it as such or not.

However, Williams presumably feels that not enough work has been done by the externalist to show that what can be said of some ideal and non-specific agent has any relevance to what can be said of an actual, particular (and presumably non-ideal) agent, with his own preferences, commitments, idiosyncrasies, abilities and life history. These factors make the particular agent’s circumstances different from those of the ideal phronimos. Furthermore, and this would be the crucial step in the argument, in order to take reasons had by the latter agent and see them as applying them to the former, we would need to invoke the particular S of the agent, and that will mean the reasons in question will be internal. That is, the reasons of the phronimos will only transfer to being reasons for the particular agent if that agent’s S contains the right elements to allow those reasons to stick.

Consider the example of twelve-tone music, which both Williams and McDowell discuss. A certain man grows up listening to the work of Arnold Schoenberg and other twelve-tone music and comes to enjoy it, and is able to articulate to his fellow aficionados all the attractions he finds it to have. Still, he might find it difficult or impossible to explain to a non-afficionado what he sees in it. In this way he is like the moral phronimos in relation to the immature or badly brought up person.

Does the fact that the aficionado is motivated to listen to twelve-tone music mean that a non-afficionado has a reason to listen to it as well? We probably think not. We probably think, as Williams seems to, that most of the reasons to listen to this music only come into existence if and when one starts to enjoy it, and if there are reasons that the non-afficionado has to try it out prior to that enjoyment, they will be reasons such as these:

*She already likes another form of music with similarities to twelve-tone music, and so is likely to (come to) like twelve-tone music.*

---

She has a mathematical mind, and enjoys discerning mathematical patterns in nature and art, and so is likely to (come to) enjoy the patterns of twelve-tone music.

In both cases, elements of the agent’s S underpin the reason she has to begin listening to the music, making the reasons internal. Unless such elements exist, it seems we have no grounds for claiming she has a reason to listen to it. If this is true of the musical case, might it not also be true of the moral case?

The internalist’s argument assumes that the burden of showing that the reasons of the ideal agent (or phronimos) are also the reasons of our particular agent A can only be discharged by looking at A’s S. However, the externalist can accept that in many cases this is true, perhaps even in some moral cases, but maintains that there exist reasons for which it is not true.

We therefore cannot simply draw our conclusions from cases such as twelve-tone music. We will need to look at a moral case. To do this, let us consult Williams’ own argument for a difference between the reasons of the phronimos and non-phronimos.

In considering what he has a reason to do, one thing A should take into account, if he has grown up and has some sense, are the ways in which he relevantly fails to be a phronimos. ...If I know I fall short of temperance and am unreliable with respect even to some kinds of self-control, I shall have good reason not to do some things that a temperate person could properly and safely do.216

That is, (G) cannot be the correct account of what A has reason to do in those circumstances in which A is importantly not a correct deliberator.

It will be helpful to have an example of a case where a non-paradigm agent has reason to do something a paradigm agent wouldn’t have reason to do with which to evaluate Williams’ objection. Here is one Michael Smith adapts from Gary Watson. A player loses a game of squash, and finds he is so put out by the loss that he wants to smash his opponent’s face with his squash racket. He knows that this is not the rational thing to do, but he cannot shake his anger, and feels that if he went over to shake the other

---

216 Williams, “Replies,” 190.
player’s hand, he might succumb to temptation and hit him. Yet if he knows that if were behaving rationally, he would walk over and congratulate his opponent. This being so, does he have a reason now in his “uncalm and uncool state” to “stride right over and shake him by the hand” or does he have reason to hang back? Smith thinks the latter. 217

Smith’s case is certainly one in which the agent has reason to do something other than what a phronimos would have reason to do. Smith supports reasons internalism, arguing that

there is an analytic connection between the desirability of an agent’s acting in a certain way in certain circumstances and her having a desire to act in that way in those circumstances if she were fully rational. 218

Smith uses the case of the irrationally angry tennis player to argue against an account of reasons for action that uses what the rational agent would do as the example which imperfectly rational agents should follow. Instead, he argues that the perfectly rational agent’s advice, rather than example, should be followed. A perfectly rational agent would advise the irrationally angry tennis player against striding over to shake his opponent’s hand, even though the rational agent herself would do just that in the same circumstances.

Though McDowell is not an internalist, and would not accept that what one has reason to do is best captured in terms of what is rational as opposed to irrational, he can avail himself of Smith’s idea and argue that, at least sometimes, it is not what the phronimos would do, but what she would advise, that the non-phronimos should follow. If this is acceptable, it means that (R) is not simply a matter of (C) and (G), but something more complex, and that Williams’ objection can be met.

It is still possible that internalists will argue that in giving her advice the phronimos must take into account the agent’s motivational set S in such a way that any reasons enshrined in the advice are internal. The thought I suppose would be that a phronimos

---

218 Ibid., 109.
who did not advise in this fashion would be paternalistic and insensitive to the agent’s actual situation.

But let us actually look at the way in which the phronimos takes the agent’s S into account in advising against approaching the net at the end of the game. Her advice certainly does take S into account, because it is an element in S – the desire to strike the opponent – that prompts her to recommend something she herself would not do. However, it is in order to frustrate, not satisfy, this desire that she advises as she does!

Neither, I must add, is it because she believes the player is ashamed of wanting to harm his opponent that she would advise him against approaching the net. She would advise against it even if the player were unrepentantly keen to approach and strike, though her advice would fall on deaf ears, because (let us accept) there may be no sound deliberative route for this unrepentantly aggressive agent to withdraw from approaching the net and striking his opponent. Though some advice based in external reasons claims surely is paternalistic and insensitive, this advice is not.

What this shows is that the reason to withdraw need not always be internal. It also shows, in the case of the conflicted tennis player, that what grounds the reason is not the existence of some desire that might come or go on the basis of which withdrawing is attractive, but our sense about what one should and shouldn’t do independent of desire. Similar comments can be made to defend other (but not all) external reasons claims against the impersonality objection.

VI The Objectivity Objection

Recall the content of (R), (C) and (G):

(R) A has a reason to φ.

(C) If A were a correct deliberator, A would be motivated in these circumstances to φ.

(G) A correct deliberator (i.e., a phronimos) would be motivated in these circumstances to φ.
A third objection to externalism – one that is only touched on briefly in Williams’ reply to McDowell, but expanded on in a reply to Nussbaum – is directed by Williams at the thought that “(R) gets its content from (C), and (C) gets its content from (G)”\textsuperscript{219}. This is, that if “we take (C) in this way, we face questions about such Aristotelian conceptions, and how stable, objective and so on they are.”\textsuperscript{220}

Consider again the issue of whether one has reason to listen to twelve-tone music. Musical styles and tastes have changed dramatically throughout the history of Western music. The works of innovative composers such as Schoenberg are at first resisted. Some compositions are eventually accepted into the canon of great music. Others are rejected as mediocre or downright bad. Some find a small devoted following amidst widespread rejection or indifference. Others are fashionable for a time, before being replaced by the new new. Which of these occurs depends largely on historical contingencies that have little or nothing to do with objective quality. And, one might add, this is just to speak of Western music. A non-Westerner might find most of the Western canon unappealing. Would she thereby be making a mistake? “To each his own,” the tolerant observer replies. The same objection can be directed against those who say the beliefs and preferences of the Aristotelian phronimos constitute seeing things aright, and use this to make claims about the reasons of everyday agents. The view objected to requires that there be stable and objective truths about what seeing-things aright or being-well-brought up are, but surely these things change, says the internalist. To maintain otherwise only makes sense according to an Aristotelian conception of human nature, one that is pre-scientific and parochial, and if we have rejected this conception of nature we should also reject talk of reasons that are not relative to the agent and her concerns.

This opens up a complex topic with a huge literature, and I cannot hope in the space of this chapter to do justice to Williams’ concern. Rather than argue for the objectivity and stability of moral reasons and conceptions of the phronimos, I will say a few words about what the believer in objectivity in these matters might want to say and in so doing indicate in the direction of the fuller reply I cannot give.

\textsuperscript{219} Williams, “Replies,” 190. The related comments directed at Nussbaum can be found on pages 199-202 of the same work.
\textsuperscript{220} Ibid., 189.
The objectivist (if I may call her that) is not asserting that cultural conditions and personal taste are always or even generally irrelevant to what people have reason to do. Two very good examples of relativity to culture are the morality of etiquette and the morality of burial practices. In culture A, the right thing to do is shake hands when meeting someone. In B it is to bow. In culture C, the dead are cremated. In D, they are buried. At the funeral, they are celebrated…or mourned, depending on the culture. The right thing to do, and what one has most reason to do, is in this way relative to the culture, and to the feelings of the members of that culture, to what they find appropriate as opposed to disrespectful or repulsive. But of course this shows that in another sense it is not relative – in all cultures one has reason to act respectfully, politely, and sensitively. What constitutes the manifestation of a virtue or a vice is often relative to culture or to person, but the core notions of virtue and vice are not.

To relate this to reasons, suppose that an acquaintance of mine is planning to dispose of his recently deceased father’s body in a rubbish dump. I tell him this is awful, and that he ought to give his father a proper burial. He replies that he can’t really be bothered, and asks whether he really has any reason to go to all that expense and trouble. When I say he has reason to give his father a proper burial, whatever he feels about the matter, I should not be understood as claiming that any ideal deliberator would bury his own father or that I recommend burial to anyone who had lost a father. Rather, I am claiming that any ideal deliberator would treat his own father’s body with due respect and want to see any other treat his father’s body with the same respect, and that for this agent in this culture, that means organising a burial, or something else fitting. What constitutes due respect in this instance is culturally (maybe even individually) dependant, but the notion of respect is not.221

These few remarks indicate the kind of response I think appropriate to the third objection.

VII Conclusion

221 The moral externalist believes there is such a thing as a shared human nature, and a shared human form of life, but with multiple legitimate manifestations in culture, practices, etc. A contrasting view is the Nietzschean one, where for each type of human there corresponds a different morality: master morality for the masters, herd morality for the herd.
In conclusion, the externalist can meet the possible explanation objection, the impersonality objection, and the objectivity objection. None of Williams’ or the other internalists’ arguments against externalism and McDowell that I have discussed succeed in showing that there are no external reasons. I have not conclusively demonstrated that external reasons exist, but I hope I have shown that we can make sense of the notion and defend it against certain objections. At the very least, I hope I have persuaded readers that it is worth looking again at the Williams-McDowell exchange.
PART C: Morality and Reason
8 Morality, Irrationality, and Unreasonableness

I Introduction

In this chapter, I begin by introducing and defending the idea that moral reasons are grammatically basic, meaning that morality ultimately lies – by its nature – unsupported by extra-moral reasons. A more detailed explanation of this idea is given in Section II. In Section III, I respond to an opposing view I call external moral rationalism, which tries to see morality as justified by the demands or nature of an independent practical rationality. The section takes a recent rationalist argument from Julia Markovits’ Moral Reasons as its principle target, but the essence of my criticism is meant to be applicable to all versions of external moral rationalism. In Section IV, I respond to a related view, internal moral rationalism, according to which morality or goodness is a part-determinant of practical rationality, entailing that immorality must always be irrational. I focus on Philippa Foot’s internal rationalism as presented in Natural Goodness, and argue that “unreasonable”, not “irrational”, is the right word for describing someone who fails to see basic moral reasons as reasons. Section V summarises the middle position I have taken on the relations between morality, rationality, reasons, and reasonableness.

II Moral Reasons as Grammatically Basic

Moral judgments are, undeniably, value judgments. Coming to embrace morality involves coming to embrace new value judgments. For instance, one learns to value
truth-telling, and to despise lies. One learns to prefer justice to injustice. One learns to see violence, except in special circumstances, as the wrong means to one’s ends.

Can one be argued into these value judgments? Certainly, argument can be used to lead someone to a new value judgment. For instance, I can bring a child to see the value of eating his vegetables by explaining how these vegetables provide the type of nutrition without which he will become sick, or weak, or stunted. Similarly, I might persuade someone with a blanket opposition to divorce to change her mind by explaining how the inability to escape a marriage often forces someone who otherwise would leave to suffer continuing emotional or physical abuse. In both cases though, the ability to argue for the conclusion that X has value – that vegetables, and the ability to divorce, have value – is accomplished by using some other value that the audience endorses as leverage. In the first case, it is his concern to be healthy. In the second, it is her concern to see that women and men are not ill-treated.

To generalize, any valid argument whose conclusion is a value judgment must have another value judgment within the premises. Naturally, if that value judgment is not endorsed by our audience, the argument will be fruitless, even if sound. So, if there is to be an argument that is irresistible to all rational beings for some value judgment or cluster of them, such as the judgments involved in morality – a universally attractive argument for morality – we must find something that all rational beings value as a basis to argue from, or failing that, a set of things such that any rational being must value at least one of them, and any one of them is a sufficient basis to argue from.

One would be hard pressed to find a universally attractive argument for any value – the importance of brushing teeth regularly, the superiority of democracy to dictatorship, etc. – because people are so diverse in background, situation, temperament and outlook. The depressed and suicidal, for instance, throw a spanner into almost any attempt at a universally attractive value argument. What do the deeply depressed and suicidal care about healthy teeth or the greater long-term flourishing of their country under democratic leadership? Universally attractive arguments and universally accepted practical reasons are rare if existent.
It would therefore seem that morality is unlikely to yield to universal argument. However, I do not mean to argue that this is unlikely. I wish to argue that it is, in a certain sense, impossible. However, what I am taking to be impossible will need to be put carefully. What this impossibility hinges on is the fact that moral reasons for action are grammatically basic.

Let me explain. Consider the following discussion regarding illegal file-sharing:

A: I’ll just download the film from a Peer-to-Peer website. Then I can watch it tonight.
B: No, no, no. You can’t do that.
A: Why shouldn’t I?
B: Because you’re not paying for it, so it’s stealing.
A: But so what?
B: What do you mean “so what”! It’s stealing. It’s wrong. You can’t do it.
A: But so what?
B: It’s stealing. It’s wrong. End of story.

There are things that B could have said to A but did not say, and some of them she purposefully chose not to say. She could have spoken of the risks of A getting caught. She could have said that the movie wasn’t very good so it wasn’t worth downloading. She could have threatened to contact the authorities. However, though these options were open, B plainly thought that the immorality of the action was in itself sufficient reason for A to refrain from downloading the film. Plainly, she felt that she did not need to say more beyond this. So long as A could see that this was a genuine case of stealing, there was no need for further argument.

When I say moral reasons for action are grammatically basic, I mean that an endorsement of a moral reason for action is genuine only if the person who endorses the reason and the action sees moral behaviour, moral character, and a moral life, as having intrinsic value. She endorses the action on the basis of this intrinsic moral value, not
some extrinsic value that this type of action may sometimes also have. She will often be disinclined to call on extra-moral reasons to supply extra force to her argument, because she does not think this should be necessary. When moral claims state what is morally required or prohibited, they are seen by moralists as ending the discussion – or at least as claims that should end discussion. They mark points where there is nothing more to be said – or nothing more that needs to be said – in support of, or against, performing some action.²²² Saying something is morally required or prohibited is, in a moral person’s eyes, saying enough – even when what is at issue is what a less than moral or outright immoral or amoral person ought to do.

Of course, when a moral person discovers that she is trying to persuade an immoral or amoral person, or even just a morally immature person, she may decide for pragmatic reasons to change tack. She may stop arguing that illegal file-sharing is immoral, and start arguing that it puts the file-sharer at risk of prosecution. She may threaten to call the police. She may offer to pay for the file herself. In doing this, she in a sense gives the other person a reason to do the moral thing, but in another sense she does not. She gives the person a reason to do something which, as it happens, is also the moral thing to do, but that reason is not a moral reason and that action is not, strictly speaking, a moral action.

There can be no argument for the conclusion that one respect some moral reason, or in general respect morality, founded on extra-moral values – not because there is nothing of extra-moral value connected with things of moral value, but because the grammar of moral claims does not allow for such arguments. If an argument is formed, it cannot strictly speaking be an argument for morality. It can only be an argument for acting to some approximation as a moral person would – and a moral person needs no such argument.

My point applies not only to moral reasons, but to any other sort of reasons for action that are basic by their nature – if there are any basic reasons outside of moral reasons. In

²²² Not every moral reason is of this sort. I have reasons to do many morally good things which I am not morally required to do. In this sort of case, the fact that I have a reason to do some morally good thing need not end the discussion about whether or not I should do it, even if I am moral.
general, there can be no argument for treating or seeing something as a basic reason for action.\textsuperscript{223} There can only be arguments for treating something as a derivative reason. This of course is tautological, for a derivative reason for action is a reason that is respected only because an argument supports its being respected.

Suppose one is offered this argument:

\begin{itemize}
  \item P: People who act morally are more likely to prosper financially and socially than people who don’t.
  \item C: Therefore, one should be moral.
\end{itemize}

I hope that premise P is true in our society, but whether or not it is, the argument is only valid if the conclusion is interpreted in a very loose way. One is supposed to be moral \textit{whether or not} it leads one to prosper financially and socially because moral reasons are basic, not derivative. What the argument shows, at best, is that one should be prudent by copying the behaviour of moral people. To see this, we need only imagine the possible (or sadly actual) world in which P is not the case. Anyone whose basic reason for being “moral” is given by the above argument will have no reason to be moral (or even “moral”, i.e. to copy the genuinely moral) in that world. It is part of the grammar of morality that in calling someone moral we do not only reference how she acts in the actual circumstances of her life, but also how she would act in other circumstances. If someone only avoids doing the wrong thing when she risks being caught and punished, she does not genuinely avoid doing the \textit{morally wrong} thing but the \textit{dangerous, imprudent} thing. As prudence and morality are conceptually separate, there are possible situations in which they are practically separate, and therefore prudence cannot form the basis for genuine morality.

Morality is grammatically insulated from the possibility of extra-moral support by argument. This means it is impossible that there could be a universally attractive argument for morality and impossible that there could be an external justification of

\textsuperscript{223} Although one cannot be argued into accepting something as a grammatically basic reason, one can be raised into accepting it in that way. See Chapter 7, Section III, on tilling and harvest reasons.
morality. To recognize this, one need not be a person who values morality. One need only understand the nature of moral claims and moral reasons, claims and reasons that one is then free to reject.

In order to see what it means to say that moral reasons are grammatically basic from another angle, it may be helpful to examine a toy-version of morality which I call morGality, inspired by a consideration of rudimentary Game theory. Consider the infamous one-off n-person prisoner’s dilemma, in which the payoffs come in some morally neutral coin, such as money, rather than some morally infused coin, such as justice. I will define two types of player:

* A morGalist is a player who sees the fact that the game being played is a prisoner’s dilemma as a basic reason to cooperate, and will therefore cooperate.

* An immorGalist is a player who does not see the nature of the game as a basic reason to cooperate, and therefore will at least sometimes defect.

Given that we are talking of a one-off dilemma, we can ignore any information about previous play, about reputation, and risk of punishment – the kinds of considerations that make TIT-FOR-TAT a good strategy in repeated play. Although the morGalist believes there is reason to cooperate in a one-off game, both for herself and her fellow player(s), and in favouring cooperation disapproves of defection by herself and others, she has no opportunity to affect anyone’s play with promises, threats or punishments. Likewise, the immorGalist cannot make promises (which might or might not be kept later) or threats (which might or might not be carried through on later).

The morGalist will articulate her reason for cooperating as follows: “It’s a one-off PD. Therefore, the thing to do is cooperate.” If she is questioned “but why?” her only reply can be to articulate the features of the game that make it a prisoner’s dilemma – that is, she can point one by one to the features of the payoff matrix, such as (for a 2-person game), “if we both cooperate we’ll both do better than if we both defect. But if only one
of us defects, that person will receive the greatest payoff while the other will receive the worst,” and so on. “So, in consequence, we should cooperate.”

It is important to note that the morGalist is not insensitive to what is known as the temptation payoff (the payoff for defecting while the other cooperates). If she were, it would not be the largest payoff in her matrix. Also, she is aware that the other player may well be tempted into defecting by his or her own temptation payoff. However, despite wanting what that payoff offers, and knowing the other wants it as well, she sees no real reason to defect, because the situation is what it is — a prisoner’s dilemma.224

It is also important to note that the morGalist is aware of the risk of getting the sucker’s payoff (the payoff for cooperating when the other defects), and does not want to get this payoff. Likewise, she knows her fellow player fears the sucker’s payoff, and may defect to avoid it. Nonetheless, despite fearing that payoff and knowing the other fears it, she sees no real reason to defect, because the situation is what it is.

If things do not pan out – if the other player defects – the morGalist will not look back over her choice to cooperate with regret. She will not regard herself as a sucker. She will see what she did as the right thing to have done, and what the immorGalist did as the wrong thing to have done, and she will retain her disposition to cooperate in future games. (Recall again that there is no opportunity to inflict a deserved punishment on the defecting immorGalist – this is a one-off dilemma.) She will regard having got the sucker’s payoff as the price for keeping open the possibility of mutual reward. She is, in this sense, willing to offer herself as a potential sacrifice for the greater good.

224 Strictly speaking, the morGalist is not playing a straightforward prisoner’s dilemma game at all. She cares about something besides the payoffs of the matrix – she cares about behaving morGally – and this trumps any potential payoff in the morally neutral coin of the apparent prisoner’s dilemma. We can explain her disposition to cooperate in apparent prisoner’s dilemmas by saying that she is actually playing a different game where the payoff for defecting is always lower than the payoff for cooperating, whatever the other players do – because cooperating is valued. According to this strict characterization, if the morGalist recognizes she is in a potential prisoner’s dilemma situation, it will not be a prisoner’s dilemma situation for her. It follows, also, that if any immorGalist player knows he is playing with a morGalist who knows she is in a potential prisoner’s dilemma situation, then he is not playing a prisoner’s dilemma either. He knows that the morGalist will cooperate, and so he will think the only rational option for him is to defect, since he is not a morGalist. So, strictly speaking, prisoner’s dilemmas only exist when all players are immorGalists, or when some players do not know whether or not the other players are immorGalists.
However, this is not some morGalist version of the supererogatory. She believes every agent has reason to do the same, and that any agent who does not do the same is doing the wrong thing, the morGally wrong thing.

What I am saying is that morality is similar to morGality, and the moralist is similar to the morGalist. The morGalist does what’s best for the group even knowing the risk involved, and despite desiring the temptation payoff. The morGal requirement to cooperate is a categorical rather than hypothetical imperative: the fact that she wants the temptation payoff, and that it would be in her interest to receive it (narrowly conceived in terms of maximizing morally-neutral payoffs), does not change the fact that she sees herself and others as having compelling reason to cooperate. Similarly, any moral requirement to pay one’s taxes, to help the poor, or to keep one’s marriage vows (etc.) expresses a categorical imperative to a moral agent that she regards as reason to do those things irrespective of the fact that cheating on her taxes, ignoring the poor, or ignoring a vow, might be the thing she most feels like doing at the time. She may know that the payoffs in some morally-neutral coin are (statistically speaking, or with certainty) greater for behaving immorally, and she may be sorely tempted to do so, but that changes nothing for her if she is moral and rational.

Does any argument exist that the morGalist could use to persuade the immorGalist to come over to her side? It should be apparent that there is none. The morGalist’s commitment to cooperating is not derived from anything else – she does not calculate that cooperating is the best way for her to achieve some other end. Neither is her commitment to cooperating merely part of her commitment to choosing rationally. We cannot accuse the immorGalist of choosing irrationally without, in his eyes, begging the question. Of course, the morGalist can strike a deal with the immorGalist, and threaten defection if the deal is not adhered to, but even if the immorGalist signs up to the deal, this does not mean he has been given a reason to become a morGalist. The definition of “morGalist” refers to what one does in one-off prisoner’s dilemmas.

Just as there is no argument available to the morGalist that would sway the immorGalist to join her, there is no argument available to the moralist to sway the immoralist or
amoralist either. We can reason with the immoral and amoral, using promises and threats, to act in accordance with morality, but even if they therefore act in accordance with morality, they will not have been persuaded to be moral. They will at best have been persuaded that it is prudent to act as we wish. To bring them to morality requires something beyond argument, and beyond threats and promises.

III    External Moral Rationalism

Although some moral reasons rest on other moral reasons, whatever moral reasons are at the bottom, this is the place where our spade is turned, if we are digging for reasons. To dig for deeper justification – say, “what’s in it for me?” – is either to misunderstand the grammar of morality or to reject morality.

Some philosophers will disagree. They will think there is deeper yet to dig, not down to some desire beneath morality (for wealth, longevity, success or happiness) that might be served by acting morally, but down to something intrinsic in the nature of practical rationality itself. According to this view, immorality is always and necessarily practically irrational. I will call the doctrine that there is such a necessary connection moral rationalism, or rationalism for short. This doctrine should not be confused with a different but in many ways related doctrine of moral rationalism that is sometimes discussed, according to which moral truths can be known by reason alone. That doctrine is epistemological, but the one I mean to examine is not.

Before I can present and critique arguments for moral rationalism, we should note that the doctrine can take two forms, depending on whether practical rationality is seen as lying underneath morality and justifying it from apart and below, or whether moral concerns are seen as simply part of how practical rationality is correctly characterized, so that practical rationality does not justify but rather contains morality. To establish the truth of the former sort of rationalism, we would need an argument to show that a commitment to being practically rational, characterized in morally neutral terms, necessitates a commitment to being moral, acting morally, or pursuing a moral life. This type of rationalist argument would need to be one that, to quote Scanlon, “appeal[s] to
considerations that are as far as possible independent of the appeal of any particular ends.\textsuperscript{225} In contrast, whether or not we end up believing the latter form of rationalism depends on whether or not we think it is accurate to say that morality is already (at least implicitly) part of our existing notion of practical rationality, or a perfected version thereof. If it is accepted that morality is simply part of the notion of practical rationality, then the appeal of rationality is not independent of the appeal of particular ends; it is dependent on the appeal of moral ends. Let us call the independent form of rationalism \textit{external rationalism} and the dependent form \textit{internal rationalism}. I will critique external rationalism and internal rationalism in turn.

The most famous external moral rationalist is Kant. In more modern times, the position has been defended by such figures as Christine Korsgaard, Alan Gewirth, and Julia Markovits, though not under the name I have given it.\textsuperscript{226} I will devote most of my attention in this chapter to Markovits for two reasons. First, her argument is the most recent and so has not been much commented on, and second, Markovits bases her argument on the Williams-style internalism about reasons which we examined in Chapter 7, and is interesting to anyone, like me, who is interested in that position. However, before dealing with Markovits’ argument I should like to make some brief criticisms of an argument of Gewirth’s, so that I may later generalize about what is faulty in all arguments for external moral rationalism.

Alan Gewirth’s argument for external rationalism – which is too long to examine in its entirety here – hinges on a leap from the claim (1) my “freedom and well-being are necessary goods [to me achieving my ends],” through (2) “I must have freedom and well-being,” to (3) “I have rights to freedom and well-being.”\textsuperscript{227} An agent who accepts (1) must, thinks Gewirth, accept (2) and (3). Furthermore, she must also accept that other agents have these rights, and thus that she ought to act so as to respect those rights. What this shows is that, since we all believe (1), we all ought to act in a rights-respecting way. Otherwise, we are irrational.

\textsuperscript{225} Scanlon, \textit{What We Owe}, 150.
\textsuperscript{227} Gewirth, “The Justificatory Argument,” 91-94.
Unfortunately, the argument either trades in equivocation or is not actually an argument in support of external rationalism. Let us assume that it is interpreted as an externally rationalist argument – i.e. that the sense of the initial premise (1) is not already morally infused. If we work forwards, the necessity and goodness of my having freedom and well-being are patently goodness-for-me, necessity-for-me, goodness-in-my-eyes and necessity-in-my-eyes. Likewise, the sense of (2) in which I must have these goods is that they are necessary for me and in my eyes. However, this does not entail that I have rights to these goods, as (3) states. One can accept the earlier claims about goodness, necessity, and what I must have, but reject the third about rights, simply by being an error theorist about rights. I take it we are all error theorists about rights for some living things with needs: flowers must have sunshine and water, but do not have rights to sunshine and water. As it happens I believe in human rights, but the fact that there is a logical possibility of being a rights error theorist and accepting Gewirth’s earlier premises means the argument is invalid.

Equivalently, if we work backwards, the conclusion (3) about rights can only be validly reached by taking “must” in (2) to have a morally infused sense. That is, my having freedom and well-being is how the world ought to be arranged, morally speaking. Moving further back, this means that the first premise must also be morally infused. That is, my freedom and well-being are morally speaking good things, and are morally speaking necessary. However, if this is what the argument rests on, we do not have an argument for external moral rationalism.

In brief, Gewirth’s argument either trades in equivocation, or is not an argument in support of external moral rationalism. The same criticism can be applied, I believe, to all such arguments.

A more recent attempt to ground morality in the nature of practical reason is found in Julia Markovits’ *Moral Reasons*. Her goal is to show that failure to act morally is a manifestation of procedural irrationality. Markovits is correctly characterized as an

---

external rationalist in that she sees herself as offering a formal account of the importance of moral reasons. A formal account is, she says quoting Scanlon, an account that “appeal[s] to considerations that are as far as possible independent of the appeal of any particular ends.” Acting immorally, she goes on, is “failing to live up to the demands of procedural rationality” because “we behave irrationally when we fail to recognize others like us as our equals, in the sense that their goals and needs matter as much, objectively, as ours do.”

Here is the central argument of her book for external rationalism:

(1) I value the ends I rationally set myself, and take myself to have reason to pursue them.
(2) But I recognize that their value is only conditional: if I did not set them as my ends, I would have no reason to pursue them.
(3) So I must see myself as having a worth-bestowing status.
(4) So I must see myself as having an unconditional value – as being an end in myself and the condition of the value of my chosen ends – in virtue of my capacity to bestow worth on my ends by rationally choosing them.
(5) I must similarly accord any other rational being the same unconditional value I accord myself.
(6) So I should act in a manner that respects this unconditional value: I should use humanity, whether in my own person or in the person of any other, always at the same time as an end, never merely as means.

The second premise deserves some comment. It expresses a form of reason internalism, one that is related to but somewhat different from the reason internalism argued for by Bernard Williams, which I discussed in Chapter 7. Markovits explains her internalism as follows:

---

229 Ibid., 195. Markovits is quoting from Scanlon, What We Owe, 150.
231 Adapted from Markovits, Moral Reason, 102-3.
[A] reason for an agent to φ is a consideration that counts in favor of φ-ing – that throws its justificatory weight behind φ-ing – in virtue of the relation it shows φ-ing to stand in to the agent’s existing ends (for example, by showing that φ-ing is a means to one of those ends, or constitutive of it, or valuable in consequence of the value of that end).\(^\text{232}\)

Williams restricted himself to arguing that if a reason for an agent to φ existed it was necessary that the agent’s motivational set contain the right elements (i.e. elements from which a sound deliberative route to a conclusion that she should φ could be reached), but Williams also thought that the existence of such a route was sufficient to ensure that a reason to φ existed.\(^\text{233}\) Nonetheless, Williams only argued for a relation of necessity.

However, Markovits’ argument must use a relation of sufficiency. To see this, note that in (4) we are told that as an agent I have the “capacity to bestow worth on my ends by rationally choosing them”. That is, by wanting or caring about something I give it worth, and by giving it worth, I give myself a reason to pursue it – so long as my wanting or caring is “rational”. Thus, rational desire is sufficient to generate reasons for action.

What, according to Markovits, is involved in my rationally wanting or choosing or caring about something? Basically, my wanting (etc.) must not be based on factual error or logical error, or – I think we could add – on ignorance of relevant fact. Here is Markovits on the restricted nature of her internalist thesis:

\[\textbf{[N]}ot all and not only our desires give us reasons. Not all of our desires give us reasons because unjustified false beliefs or bad reasoning can give us desires we have no reason to fulfill…And not only our desires give us reasons, because we value and act for the sake of many things we can’t properly be said to desire, because they aren’t the kinds of ends we could achieve or come to possess; for example, we often act for the sake of other people. Our “motivational set” contains everything for the sake of which we act, everything we pursue, promote, protect, and respect. Secondly, and relatedly, linking what we have reason to do with what satisfies our desires suggests that reason plays a purely instrumental\]

\(^\text{232}\) Ibid., 52.
\(^\text{233}\) Williams, “Internal Reasons and Obscurity,” 35-6.
role. I’ll follow Williams, however, in allowing for the possibility that we have reason to act in ways that serve our ends non-instrumentally – perhaps the action in question is constitutive of some end or commitment, or expresses that commitment\textsuperscript{234}

So I take it, using Williams’ terminology, that what is involved in Markovits’ notion of my rationally desiring, valuing, respecting, etc. an end is either that there is a *sound deliberative route* from the contents of my *motivational set* to the conclusion that I ought to pursue the end in question, or the desire for the end already a member of that set (so that no deliberative route is needed).

In summary, Markovits’ argument rests on a form of reasons internalism: an agent’s rationally desiring (or choosing etc.) something is sufficient to entail that the thing she desires has worth and that she has reason to pursue it.

To an unsympathetic reader, Markovits’ argument from (1) to (6) can look somewhat like a series of non-sequiturs. In order to see it as a potentially valid argument, we need to do some work. Understanding the role of internalism in Markovits’ argument allows us to give a clearer reconstruction:

(7) What any agent rationally desires (with “desires” broadly construed) can be valuable purely in virtue of her rationally desiring it, and any agent has the power to bestow value on the valueless/value neutral by coming to rationally desire it. (Sufficient-for reasons internalism)

(8) An agent who can bestow value on an end by rationally desiring it, as in (7), has a special kind of value herself: value-bestowing value.

Therefore,

(9) Any agent has value: the value of being a value-bestower.

\textsuperscript{234} Markovits, *Moral Reason*, 5.
Therefore,

(10) I (and everyone else) should act to respect that value, by treating the agent as an end in herself, i.e. as a valuable value-bestower.

I think the argument from (7) and (8) to (9) is valid. Before looking at the move to (10), let us question the truth of premise (7) and whether Markovits can give rationalist grounds for accepting it.

In order to see that (7) requires some justification, one need only note that it is most certainly controversial. There are at least two positions opposed to (7) that could be rationally coherent and would need to be ruled out by a moral rationalist whose position rested on reasons internalism. The first is the position of what we might call the value nihilist or value error theorist. The value nihilist holds that nothing has any value because no sense can be given to the notion of value, if by the notion of value we mean something more than the notion of being desired. The value nihilist admits that various people desire various things, and so that in that sense various people value various things, but nonetheless the universe is ultimately value-free. We do not value things because they have value. Nothing has value.

The second position opposed to (7) is that of the value objectivist. The value objectivist holds firstly that certain things – such as human survival, pleasure, beauty, family, or power – have intrinsic value, secondly that nothing has value of any kind except in relation to what has intrinsic value (such as by being a means to its attainment), and thirdly that the satisfaction of human desire does not have intrinsic value, but is only valuable if what is desired has objective value by being intrinsically valuable or related to something that has intrinsic value.

As far as I can see, the only justification someone could plausibly give for believing (7) is that human rational choice has value, no matter who is doing the choosing or what is being chosen, because every human person’s life is of intrinsic value and a successful human life necessarily involves the person’s achieving or attaining many, though of
course not all, of what she desires or chooses. But the claim that every human person’s life is of intrinsic value, it is apparent, is itself a substantive moral claim. As such, it cannot be used as a premise in an external rationalist justification of morality.

The only other possibility is that the rationalist has a different, rationalist, argument for accepting premise (7). This, plainly, has not been given in Markovits’ argument. Until it is given, and I am confident it cannot, this means we are no nearer to having a rationalist justification of morality than we were to begin with.

Now let us consider the move from (9) to (10). The success of this move hinges on whether the value bestowed is relative to the agent that bestows it, or not. That is, when an agent bestows value on an end, does she only bestow value-for-her, or does she bestow value simpliciter? Likewise, and as a result, when we say in (9) that an agent has value, do we mean value-for-herself, or value simpliciter?

The problem is for the argument to (10) to be valid, the value involved needs to be value simpliciter, but only a claim about what has value-for-(the agent)-herself can plausibly be said to follow as a requirement of the agent’s being practically rational.

In Markovits’ original argument, the equivocation over “value” is supported by an unfortunate use of, or equivocation over, the word “status” in premise (3): “So I must see myself as having a worth-bestowing status.” There are two sorts of status: social status and objective status. Usually, talk about status is talk about a standing in the eyes of others, a standing that is seen as legitimate according to some generally shared view about what grounds or creates that status. Social status can only exist if there is this shared view. For instance, one had the recognized status of a knight in a medieval kingdom if one had been knighted by the legitimate king or queen, but without the existence of a community that recognizes the status of knighthood, it is not be possible to have that status.

Markovits’ argument needs to show we have a worth-bestowing status independent of the views or practices of such a community. This would be an objective status. We may
accept that there is such a thing as objective status. Many of us accept that there are, for
instance, objective human rights. But a supporter of universal human rights, rights
which are independent of whether any community exists that respects those rights or
enshrines them in law, takes up a substantive moral position in her support of these
rights. The belief in these rights, or in any sort of objective status, is not a requirement
of pure practical rationality.

Thus, to use the word “status” in (3) is a problem, since the status is either only social,
and dependant on the endorsement of a community, or objective, in which case
Markovits has once again imported a moral conception into the premises, thereby
undermining the status of the argument as a rationalist justification of morality.

I omitted talk of status from my reconstruction of Markovits’ argument in order that we
might get as close as possible to a valid and non-misleading argument. For the same
reason, I excised talk of unconditional value in my reconstruction – compare (4) to (9).
Markovits sees most of our ends as having conditional value – their value depends on
the fact that we want or need these ends. According to her version of internalism,
whatever end we pursue, “[w]e are the condition of its value,” and, supposedly, “their
mattering to us could make them matter only if we matter—if we have value.”235 Our
value cannot, she thinks, be conditional (on others valuing us as ends) because this
would involve us in “regress or circularity”. It remains (it would seem) an easy step to
go from the idea that we each have unconditional value to the idea that we ought to
treat each other with due respect. However, the most Markovits’ argument can show is that
we have unconditional value for ourselves, not unconditional value simpliciter. Her
argument shows at best that each of us matters to ourselves not because we matter to
others but because if we didn’t matter to ourselves, we couldn’t find our ends valuable,
and we do find our ends valuable (to ourselves). It does not show we have objective
value – which is also, in some sense, an unconditional kind of value. Since talk in terms
of unconditional value in Markovits’ sense can easily be confused with talk of objective
value, the argument is less misleading without such talk.

235 Ibid., 94-5.
To conclude, my critique of external moral rationalist arguments – whether Gewirths’, Markovits’, or some other’s – is that they either trade in equivocation, so that they are invalid, or they are not arguments in support of external moral rationalism, but instead arguments that take for granted the value of morality in their premises.

IV Internal Moral Rationalism

One natural response an rationalist who feels the bite of my criticism can make is to retreat to internal rationalism. Internal rationalism may have other things to recommend it, but chiefly it allows us to say that immorality is always irrational without needing to worry that we have illicitly imported moral considerations into our premises or our conception of rationality. Internal rationalism makes what was illicit licit. However, I believe internal rationalism is – if not straightforwardly wrong – at least highly problematic. In order to make my case, I will examine the view of a particular internal rationalist, Philippa Foot.

In *Natural Goodness*, in a section inspired by Warren Quinn’s “Rationality and the Human Good”, Foot claims that while trying to show the rationality of acting morally many philosophers have been viewing the relationship between morality and rationality backwards. She argues that they have been trying to show that morality is “consonant” with an independently characterized practical rationality, when they should have simply accepted that we ought to see “goodness as setting a necessary condition of practical rationality and therefore as at least a part-determinant of the thing itself.” That is, Foot rejects what I have labelled external rationalism and offers internal rationalism in its place.

To some readers who respect the involved (if doomed) attempts of the external rationalists to justify morality, internal rationalism will seem like a cheat. For instance, Charles Pigden comments:

---

“[According to Foot] it is no objection to her conception of natural goodness that considerations about what is naturally good may have no bite for the rational gangster. If the gangster is unmoved by considerations about natural goodness, then what this shows is that he isn’t really rational at all, since being rational entails being moved (or being movable) by such considerations. Foot has a lot of praise for Quinn (also an admirer of Nagel) who opened her eyes to the possibility of defining her way to victory in this cheap and easy manner. Quinn’s method, to be sure, has many advantages for the moral realist – they are analogous to the advantages of theft over honest toil.”

Those who share Pigden’s feeling of being argumentatively short-changed may be reminded of one of Williams’ criticisms of reason externalism. In Williams’ opinion, the reason externalist “seems concerned to say that what is particularly wrong with the [amoral or immoral] agent is that he is irrational.” However, “the suggestion that the agent is being irrational … once the basis of an internal reason claim has been clearly laid aside, is bluff.” Though Williams did not have Foot in mind when he wrote, indeed Foot’s position was different at the time, the charge might seem to be appropriate.

I shall not go so far as to agree with Pigden (or Williams), but the charge that Foot is “defining her way to victory” and engaging in “bluff” points us in the direction of what I think is a fair criticism: Foot is using the word “rationality” in an unnecessarily misleading way.

I will try to make plain my discomfort with Foot’s use of “rationality”. Let us focus on the case of an amoralist – which I define here as a person who does not care about morality and does not govern her behaviour in the light of moral considerations. I think

---

239 Williams, “Internal and External Reasons,” 110.
240 Ibid., 111.
241 For Foot’s earlier position, see Philippa Foot, “Morality as a System of Hypothetical Imperatives,” The Philosophical Review (1972): 305-316.
the natural way for a moralist (i.e. a believer in and pursuer of morality) to describe the amoralist is as someone who does not appreciate the genuine importance of moral considerations and as such as being someone who is ignorant of or in error about the importance of moral considerations. When the amoralist acts immorally, her behaviour demonstrates this ignorance and/or error. Now consider a different case of someone acting wrongly because of ignorance or error. Suppose I am following a recipe, but unfortunately I wrongly believe that “tsp” is the abbreviation for “tablespoon.” On this basis, I put a tablespoon of salt into my mix when I should have put a teaspoon, and ruin my cake. Certainly I have made a mistake, and perhaps I should have known better, but what I have not done is act irrationally. I have acted rationally, but on the basis of a false belief. Likewise, suppose I am ignorant of the fact that fugu (Japanese blowfish) is poisonous unless correctly prepared. If I serve my guests some poisonous blowfish, and one of them dies as a result, I can be accused of many things, but one thing I would not be accused of is acting irrationally. In both cases what I lack is not rationality but knowledge (and, depending on the facts of the case, due care).

Moral cases of ignorance or error are, I’ll admit, special and different to these ordinary cases of ignorance and error. In particular, in ordinary cases it is usually quite straightforward to show someone that she is in error by consulting an encyclopaedia or asking an expert. What makes the amoralist’s error different is that there is no way to show her that she is missing an important truth – e.g. that all human life has value – since that truth claim lies at the ground (i.e. is grammatically basic) rather than being supported (or contradicted) by evidence as most other claims are. The amoralist simply does not see the ground as ground and there is nothing we can do to persuade her otherwise, at least within the confines of rational debate.

It is useful to have a special word to describe the amoralist who acts without respect for morality and moral reasons, but the word is not “irrational”. Though likely not ideal, I think a much better word is “unreasonable.” The amoralist is an unreasonable person. She does not listen to reason when it is ultimately a moral reason being given for acting or not acting in a certain way. She cannot be reasoned with when the argument hinges

---

on moral considerations. Similarly, she treats her own happiness and welfare as *unreasonably important* relative to the happiness and welfare of others. Nonetheless, though she may act irrationally at times, she is not necessarily irrational just because she is amoral and unreasonable.

T. M. Scanlon, like me, is careful to distinguish the rational from the reasonable. He has given the meaning of “reasonable” his attention because the notion of principles that one cannot *reasonably* reject is central to his contractualism. As he says,

> A claim about what it is reasonable for a person to do presupposes a certain body of information and a certain range of reasons which are taken to be relevant, and goes on to make a claim about what these reasons, properly understood, in fact support.  

In particular, we often take moral reasons as relevant when we judge what it is reasonable for someone to do. We may also take financial considerations, or considerations of safety etc. as relevant reasons.

Though Scanlon’s comments are correct, I think we can go a little deeper into how unreasonableness is distinct from irrationality. Each concept has a different focus of criticism. In paradigmatic cases of calling someone unreasonable, we point to a flaw in his character which prevents, impedes, or undermines our reaching agreement with him on issues requiring coordination, such as business deals. The reason why it is important that we presuppose “a certain body of information and a certain range of reasons which are taken to be relevant” is that agreement cannot be reached without them. What makes someone unreasonable is that he will not listen, that he is stubborn, or that he is selfishly only thinking of his own needs.

Let us contrast this with irrationality. In paradigmatic cases of calling someone irrational, we point to the fact that his lower, desiderative, cognitive functions are preventing, impeding, or undermining his higher cognitive, decision-making functions,

---

Scanlon, *What We Owe*, 192.
and to the fact that those higher functions are not properly restraining those lower functions from expressing themselves and exerting influence. The terminology of Chapter 2 is useful here. When someone is being irrational his beliefs are having undue influence over his beliefs about what is true or likely or about what to do, or over his actions if not his beliefs. Examples of the former type are racial prejudice or other types of bias. An example of the latter is this: if I know that international airplane travel is safer than and faster than long distance train travel, but allow my tendency to panic at the thought of air travel to lead me to travel by train, I am acting irrationally, on the basis of irrational fears.

If these characterizations of irrationality and unreasonableness are adequate, it follows that the amoralist is better described as unreasonable than as irrational.

I believe unreasonableness is a very important, and distinct, concept. However, care is needed. It is worth distinguishing cases where we use “unreasonable” to describe the fixed character of an agent from when it is used to describe the current emotional or mental state of an agent or the particular behaviour or reactions of an agent. Amoralists are unreasonable people – they have unreasonable characters. Agents who behave akratically are not unreasonable people, but they may get caught in unreasonable moods, and they usually act unreasonably when they act akratically. Incontinent agents do in general listen to moral reason, they can be reasoned with from a moral basis, and they generally try to act according to moral reason, but sometimes they fail to act accordingly, and when they fail to act accordingly they are usually being unreasonable or behaving unreasonably. In addition, akratic action is also irrational, since acting according to one’s deliberative conclusions must surely be a requirement of practical rationality, and the akratic agent fails to do this.

Though I think the distinction between the rational and the reasonable is what we need, there is an alternative way of talking which may be attractive. This is to talk of two types of rationality. Jürgen Habermas draws a distinction between strategic and communicative action, each with its own form of rationality.²⁴⁴ Strategic action is

²⁴⁴ Habermas, The Theory of Communicative Action.
governed by a concern with the choice of efficient means to achieve ends. Communicative action is governed by an ideal or telos of unforced communal agreement. Thus the kind of rationality presupposed in communicative action is much like my notion of reasonableness, while the kind of rationality required for strategic action is good old (neo-) Humean practical rationality. Whether or not the reader prefers this way of talking (strategic rationality vs. communicative rationality) or the earlier (rationality vs. reasonableness), she should not be happy with the internal rationalist thesis as it stands.

Having made the preferred alternatives to internal rationalism clear, we can turn to three separate arguments that Foot seems to make in support of her position, to which I will respond. Firstly, Foot notes that we regard there as being a necessary connection between prudence and rationality such that acting imprudently – “against desire and self interest” – is always irrational. For prudence to be related to rationality in this way yet morality not to be so related, there would need to be something “special” about prudence that morality lacked. However, she thinks there is no difference between prudence and morality that could justify them having a different relation to rationality. Therefore, internal rationalism is correct.

Secondly, Foot sees practical rationality as something of great importance. If there were no moral limitations on practical rationality, if it were “indifferent to nastiness or even disgracefulness in an agent’s purposes,” practical rationality would not be “so important”. Therefore, there must be moral limitations on practical rationality, and internal rationalism is correct.

More tentatively, there may be a third, largely implicit, argument Foot is making. There must be a way to show that acting immorally is acting irrationally. Arguments which try to show this using a notion of rationality characterized independently of morality fail in this task and are doomed to fail. However, if rationality is characterized

---

245 Foot, *Natural Goodness*, 63,
246 Foot, *Natural Goodness*, 62. Likewise, Quinn says this sort of rationality would be a “minor virtue, like neatness” and a “petty excellence”. See Quinn, “Rationality,” 213 and 226.
247 It may be unfair to Foot to think she would make this argument, since the argument has an air of circularity to it, but it is worth stating nonetheless.
as being limited by morality, we do have a way to show this. Therefore, it must be
correct to characterize rationality in this way. And, given this correct characterization, it
follows that internal rationalism is correct.

Let me respond to each argument in turn. With respect to prudence, it is debatable
whether there is a special connection between prudence and rationality. It depends on
how one conceives of the target of prudence: self-interest. Take survival for instance.
Under one conception of self-interest, it can be in one’s interest to survive even if one
wants above all things to die, simply because survival is in most cases an objective good
and therefore in most cases in one’s interest. Given this conception of self-interest, I
would argue it need not be irrational to do what is not in one’s self-interest. The person
who commits suicide would be throwing away something objectively of value, but her
mistake would not be one of irrationality. It would be one of failing to see the value of
life. Possession of rationality alone does not guarantee an ability to see what is of
importance, but only guarantees that one will act in a way that makes sense given what
one thinks is important (whether or not it actually is important). In this view, prudential
failures sit alongside moral failures as not necessarily instances of rational failure.

The alternative conception of self-interest is more subjective: it is in an agent’s self-
interest to act so as to satisfy her desires (or as many of them as possible, or the
strongest one or ones, etc.). If she wishes above all things to die, then the prudent thing
for her to do would be to find a suitable means to kill herself. With this conception, it is
irrational for an agent to act against her self-interest, for in acting against self-interest
she would be consciously acting contrary to what she most desires, and that is clearly
irrational. She would be irrational in very much the same way as an akratic agent who
professes her conclusion about what it is morally best for her to do, but then fails to do.
However, unlike the first conception, this conception does make prudence “special” in a
way that morality is not: what is prudent depends on what the agent desires in a way
that moral reasons do not.

Foot claimed that we were improperly treating prudence as special compared to
morality. However, as we have seen, either prudence is not special, and not tied to
rationality, or prudence is special, but we have reason for treating it as special, in relation to rationality.

The second argument concerns the importance of practical rationality, which would be reduced if it had no moral constraints. I concede that this is correct, but I am quite happy to see the importance of practical rationality reduced to something conditional. When a person with a healthy, morally mature view of the world uses her reasoning to decide what to do, her possession of rationality is a clear good. When a person with an unhealthy, amoral or immoral view of the world decides what to do, her possession of rationality is not such a clear good, and may even be thought of as regrettable if her rationality enables her to inflict more harm than she would have otherwise been able to inflict. What has more importance than rationality is reasonableness. Reasonableness is constrained by moral considerations. Reasonableness has tremendous unconditional value. It is reasonableness that is “so important”, rationality less so.

The third (possibly unfair) argument is easily dispensed with. The argument starts from a premise that there must be a way to show that acting immorally is always acting irrationally, and concludes that internal rationalism is the only way that will work. I simply dispute the premise – we do not need to show that acting immorally is always acting irrationally. If we do not care about trying to show this, it does not matter that external rationalism fails, and we have no need of internal rationalism.

V Conclusion: The Middle Way

I have presented a middle position on the relations between morality, reason, and rationality. This middle position stands between two other positions that are, to my mind, inferior. The first is Foot’s position: to act immorally is to fail to act as one has best reason to act, and is also to act irrationally. The other position is Bernard Williams’: to act immorally is not necessarily to fail to act as one has best reason to, nor is it necessarily irrational. The position I have presented, which I take to be John McDowell’s position and (perhaps) Scanlon’s as well, is that there are always reasons not to act immorally, namely moral reasons, so that immoral action is always
unreasonable, but acting immorally is not necessarily irrational. I have also argued that it is a mistake to look for anything deeper than morality that could be a justification for it. In that special sense, there is no reason to be moral. This, I think, both Foot and Williams would have accepted.

248 See McDowell, “Might There Be External Reasons?” 98.
9 A Defence against Moral Error Theory

I Introduction

Moral error theory came into prominence in 1977 with J. L. Mackie’s *Ethics: Inventing Right and Wrong*. Moral error theorists like to point out that there are many error theoretic precursors to Mackie, starting with the sophist Antiphon and Platonic characters Callicles and Thrasymachus, and including Hobbes, Mandeville, Hume, Nietzsche, Russell, Wittgenstein, Camus, as well as the lesser known Richard Robinson and Axel Hägerström. Nonetheless, in this chapter most of our attention will be devoted to Mackie and to two recent error theorists, Jonas Olson, who both critiques and develops Mackie’s position, and Richard Joyce.

Let us begin with a general characterization of moral error theory or theories. Moral error theories use the same argument schema to target moral discourse, but may differ in the details of the premises:

---

250 Jonas Olson, *Moral Error Theory: History, Critique, Defence* (Oxford: Oxford University Press, 2014), 61; Richard Joyce, *Myth of Morality*, ix. As a follower of Wittgenstein, I should point out that the main evidence for calling Wittgenstein an error theorist comes from a lecture Wittgenstein prepared and delivered between 1929 and 1930 but which was only published later, as for instance *Lecture on Ethics*, ed. Edoardo Zamuner, Ermelinda Valentina Di Lascio, and D. K. Levy (Hoboken, NJ: Wiley-Blackwell, 2014). I think error theorists mischaracterize Wittgenstein’s position at the time, which was mystical rather than error theoretic. In addition, Wittgenstein’s later philosophy (e.g. *Philosophical Investigations*) moves away from the idea that moral propositions are nonsensical, and away from the idea that moral truths are ineffable, by allowing for a greater variety of language games and related forms of meaningfulness than his earlier philosophy – i.e. the *Tractatus* and *A Lecture on Ethics* – recognized. (See Johnston, *Wittgenstein and Moral Philosophy*, Chapter IV.) In general then, Wittgenstein should not be described as an error theorist.

226
Moral discourse is centrally committed to thesis X.\textsuperscript{252} (The conceptual claim)
Thesis X is false. (The ontological claim)
Therefore, moral discourse is fundamentally flawed.

A moralist has three response options for defense against arguments that follow this schema. She can concede the ontological claim but challenge the conceptual claim – Joyce and Kirchin call this the “concessive” response.\textsuperscript{253} Alternatively, she can accept the conceptual claim but disagree with the ontological claim – Joyce and Kirchin call this the “head-on” response. Finally, she can reject both claims. In this chapter I aim to defend morality against Mackian/Olsonian and Joycean error theories by giving a head-on response to both, for the most part. (When I instead take the concessive route, I will make this clear.) That is, I will argue that although Mackie, Olson and Joyce usually correctly identify claims that are central conceptual commitments of moral discourse, the claims are true rather than false – at least, error theorists have given us no persuasive reasons for thinking otherwise.

Very briefly, here are the claims which Mackie (followed by Olson) and Joyce think moral discourse is committed to. Mackie thinks moral discourse is committed to the existence of non-institutional, non-conventional objective moral facts and values. He believes there are no such facts or values. Joyce thinks moral discourse is committed to the existence (to speak loosely) of non-institutional categorical imperatives.\textsuperscript{254} He believes no sense can be made of such imperatives, and thus that there are no such imperatives. We shall see that Mackie’s version of the conceptual claim is basically the same as Joyce’s version, though they offer different arguments for the ontological claim.

\textsuperscript{252} …or some theses $X_1$, $X_2$, …, $X_n$. See Richard Joyce and Simon Kirchin, eds. A World without Values: Essays on John Mackie’s Moral Error Theory, Vol. 114 (Springer Science & Business Media, 2010), xvi. I will simplify the presentation and critique of error theory in this chapter by speaking of only a single thesis X in the conceptual claim.

\textsuperscript{253} Joyce and Kirchin, World without Values, xvi.

\textsuperscript{254} Of course, imperatives do not strictly speaking “exist”. Here and elsewhere, I use “the existence of categorical imperatives” as shorthand for the idea that there are cases where it is correct to use a categorical imperative to describe what someone ought or ought not to do.
Mackie offered two arguments for his ontological claim that there are no objective moral facts: the argument from relativity and the argument from queerness. The former argues that the error theory is a better explanation of the diversity of moral opinion than the idea that there are objective moral facts which we disagree about. I will ignore the argument from relativity for three reasons. Firstly, Mackie thought this argument to be weaker than the argument from queerness.\(^{255}\) Secondly, other moral error theorists think that the argument is not decisive.\(^{256}\) Thirdly, if I were to respond to the argument, I would largely repeat comments on the relativity of reasons made in an earlier chapter. The interested reader can return to these comments.\(^{257}\)

Mackie’s argument from queerness has three strains – motivational, metaphysical, and epistemological – which I will explain and respond to in Section II. In short, Mackie supposes that (1) objective moral facts, if they existed, would be intrinsically motivating, but the idea of intrinsically motivating facts is irreducibly queer, (2) moral facts require the existence of relations or properties in the world that are irreducibly queer, and (3) moral facts are usually taken to be things we can know to be the case, but any notion of epistemological access to such facts is irreducibly queer. For Mackie, irreducible queerness entails non-existence.

Joyce has two arguments for his error theory – or strictly speaking, one argument in favour of his error theory, and one argument against the justified-ness of moral belief. The first I will call the nesting argument. Joyce argues that moral discourse involves non-institutional, non-conventional categorical imperatives, and that these imperatives are meant to supply us with reasons for action. However, any notion of what we have reason to do must nest within a general framework of practical rationality. Joyce argues,

\(^{255}\) Mackie calls the argument from queerness “more important” and “certainly more generally applicable” than the argument from relativity. See Mackie, *Ethics*, 38.
\(^{256}\) See for instance Olson, *Moral Error Theory*, 72-75.
\(^{257}\) See Chapter 7, Section VI. Briefly, the point I would make is that there is much less difference between cultures in the list of virtues they value and vices they deplore than there is in their beliefs about the details of how and when they think these virtues and vices are manifested. Cultures disagree about how to show respect to a dead body, but not (for the most part) over whether respect should be shown. Cultures disagree about when it is acceptable to kill, but not about the fact that killing is, in most cases, an expression of vice. The explanation of the differences in belief about the “how” and “when” are explained by the complexity of some moral questions, questions which we often need years of experience, deliberation, and wise counsel to know how to answer correctly and sensitively.
on the basis of Bernard Williams’ argument against the existence of external reasons, that the best account of the nature of practical rationality is practical instrumentalism. However, practical instrumentalism sees what we have reason to do as depending on our desires, meaning that there can be no such thing as non-institutional, non-conventional categorical imperatives. This argument is challenged in Section III.

The second of Joyce’s arguments is an evolutionary debunking argument. The argument aims to show that we have good reason to doubt that we have justification for believing in categorical imperatives and their related moral requirements. Joyce argues that the reason why we believe as we do is explained by natural selection, and that although recognizing this does not demonstrate that there are no moral requirements, it nonetheless ought to undermine our faith in the existence of moral requirements. This argument is tackled in Section IV.

My overall aim is not to provide arguments which show that the moral error theorist is wrong. I do not believe such arguments exist. My aim is simply to defend moral discourse against certain error theoretic arguments and show that the error theorists responsible have not given us good reason to question our participation in moral discourse.

II The Queerness Arguments

Mackie uses Plato’s Forms as an example of the kind of things objective values would need to be. In Mackie’s account of Plato’s philosophy, if something is objectively good, this fact “tells the person who knows this to pursue it and makes him pursue it.” 258

Likewise,

An objective good would be sought by anyone who was acquainted with it, not because of any contingent fact that this person, or every person, is so constituted

258 Mackie, Ethics, 40, emphasis added.
that he desires this end, but just because the end has to-be-pursuedness somehow built into it.\textsuperscript{259}

It seems that what Mackie has in mind here is that moral discourse is centrally committed to a relatively strong form of motivational internalism, i.e. that if someone accepts as a fact that he morally ought to φ, for some φ, then he is necessarily motivated to φ and he is motivated to such a degree that he will seek to φ.\textsuperscript{260} Mackie believes that the idea of necessarily motivating facts is queer, and thus that no such facts exist.

The most fitting response is one that combines a concessive with a head-on response. Here I will draw on the position on motivational internalism that I articulated in Chapter 3, “Motivational Internalism and Authenticity”. An authentic agent, one who has matured enough to understand what morality involves and who has accepted morality’s guiding role in her own life, \textit{will} necessarily be motivated by a judgment that she morally ought to do some particular thing. This is the head-on aspect, in that I accept the conceptual claim that moral discourse is committed to there being moral facts which necessarily motivate \textit{authentic} agents, but reject the ontological claim that no such facts exist. There is no guarantee that the authentic agent will actually do what she believes she morally ought to do, given the possibility of akrasia, but she will be motivated by the thought that she morally ought to do it. There is nothing queer or mysterious about morality’s ability to sway her motivations. This is simply a result of her having the character she does.

If on the other hand we are speaking of an immature, inauthentic agent, there is no guarantee that he will be motivated to act according to his moral judgments. This is because he has not matured enough to disentangle in his thinking the obligations that he accepts as binding him from the obligations that society is attempting to impose on him.

\textsuperscript{259} Ibid., 40, emphasis added.
\textsuperscript{260} This interpretation of Mackie has been disputed in Olson, \textit{Moral Error Theory}, 106-109. It may be difficult to know what view to attribute to Mackie here if Olson is correct that “Mackie failed to distinguish clearly between motivation and normativity and in particular between the claim that moral facts are \textit{intrinsically motivating} and the claim that they are \textit{objectively prescriptive}.” See page 117. For Richard Joyce’s take on the connection between motivational internalism and error theory, see Joyce, \textit{Myth of Morality}, 17-28. Joyce concludes that current attempts to argue for an error theory on the basis of motivational internalism only lead to an impasse.
but which he does not accept. If we are speaking of this type of agent, the motivational internalist thesis is false. This is the concessive aspect of my response, for I accept the ontological claim that there are no moral facts which necessarily motivate an *inauthentic* agent, but reject the conceptual claim that moral discourse is committed to there being such facts.

In short, amongst those who accept authentic internalism (or something like it) as an account of the relations between moral belief and moral motivation, moral discourse is not centrally committed to anything queer in relation to motivation.

Let us move to the next argument. The core of Mackie’s argument from metaphysical queerness is encapsulated in this statement:

> 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.\(^{261}\)

One must add two more premises to turn this into a full error theoretic argument: firstly, moral discourse is committed to the existence of objective values, and secondly, entities (etc.) which are very strange (in the way that objective values are reputedly strange) do not exist.

Intuitively, the argument from metaphysical queerness is closely related to the argument from epistemological queerness, for if one’s attempt to explain what moral facts are and how moral facts can exist is judged to be trading in queerness and nonsense, one’s attempts to explain what it is to know such facts will likely fall to the same charge. Likewise, if one can satisfactorily explain what moral facts are and what it is for them to be the case (or argue that no explanation is needed), there should be no problem in principle with explaining how we might, if lucky, come to know that some particular moral facts are the case (or, failing that, no problem with arguing that no explanation is needed). Therefore, my response to the metaphysical queerness argument is equally a

response to the epistemological queerness argument. These arguments stand or fall together.

To make progress, it will help to consider in turn the key terms and phrases from the Mackie quote above: “objective”, “very strange”, “utterly different” and “in the universe”. We can begin with “objective”. What Mackie thinks objective values would need to be, whether moral values or other sorts of values, is entities that are objectively prescriptive, which is to say entities that, in Olson’s terminology, “entail irreducibly normative reasons” or equivalently entities that, in Mackie’s terminology, are “intrinsically action-guiding.” Irreducibly normative reasons are reasons that are not contingent on the agent’s desires. The common way for philosophers to talk about such reasons is to talk of categorical imperatives. Likewise, intrinsically action-guiding entities are entities which, by their nature, provide such reasons, and thereby create categorical imperatives.

Interestingly, error theorists such as Mackie, Olson, and Joyce do accept the existence of reasons that are independent of an agent’s desires and likewise do accept the existence of categorical imperatives, but only when the reasons or imperatives are institutional or conventional. For example, Olson states that there is a desire-independent reason for writers in English to avoid splitting infinitives, since the rules of English grammar forbid it. This is not the best example Olson could have chosen, since to occasionally split an infinitive or two (as I just did) is widely accepted nowadays, but the sort of thing Olson has in mind is clear.

However, I would dispute the inclusion of this kind of conventional reason under the umbrella of irreducibly normative reasons and this kind of conventional imperative under the umbrella of categorical imperatives. The error theorists seem to be confusing (a) an aspect of the way we typically express ourselves with (b) what we actually mean.

262 Olson, Moral Error Theory, 117.
263 Mackie, Ethics, 32.
264 Olson, Moral Error Theory, 119.
265 See footnote 254.
266 Ibid., 121; Joyce, Myth of Morality, 39-42. Also see Foot, “Morality as a System.”
267 Olson, Moral Error Theory, 121.
by what we say. It is true that we usually do not mention any conditions when we express conventional or institutional imperatives. For instance, when teaching someone to play chess we say:

*You mustn’t move your rooks diagonally!*

We do not usually find it necessary to say:

*You mustn’t move a rook diagonally, that is, if you want to play chess as it’s conventionally played.*

Nonetheless, this is what we mean. The conditions are simply left out according to the Gricean maxim of quantity. What makes a genuine categorical imperative categorical is whether it provides a reason to act or not act in some way that is independent of our desires. Plainly, the imperative to avoid moving rooks diagonally or to avoid splitting infinitives (etc.) is not categorical in this sense. The reasons involved depend for their existence on the desire of the person concerned to play chess as it’s normally played and the desire to write standard, grammatical English.

I suspect the only cases where conventional or institutional rules provide real categorical imperatives are where they have the backing of *moral* obligations. For instance, the laws of the land are institutional rules that create desire-independent reasons to act accordingly, but only because there is a moral obligation to follow the law lying beyond or beneath the law, (except of course where the law is unjust or unduly restrictive to the degree that there is moral justification for civil disobedience).

Having made this point about the nature of true categorical imperatives, I should state clearly that I believe that most moralists *are* committed to the existence of moral imperatives that are sometimes, though not always, true categorical imperatives, and equally, committed to there being at least some moral reasons for one’s doing or not
doing something independent of one’s desire – reasons that are more than mere institutional or social convention. I accept, that is, this conceptual claim.\textsuperscript{268}

I accept the objectivity of moral facts or values where objectivity is understood in the manner above, but there is a different sense of objectivity which we should be careful not to confuse with the above. This is objectivity in the sense of detachment. A scientific investigation of a subject matter sometimes requires that the investigator set aside as irrelevant and unhelpful any personal preferences or practical concerns she may have, as well as any value judgments she believes to be true, at least while the investigation proceeds, so as not to let these preferences or ends or values interfere with her engagement in the investigation. After the investigation is complete the results can be put to some purpose if she so wishes, but it should not matter to the success of the investigation that the investigator had some purpose or respects some value.\textsuperscript{269} Someone indifferent to that purpose and value should be able to come to the same result. However, the moral realm cannot be investigated in this manner, and is not objective in this sense. To take up a completely detached stance is to take up a stance from which moral truth becomes invisible, because moral truth concerns what our ends and preferences should be, and what is valuable. Only from an involved, committed stance can moral facts even show up as facts. To paraphrase a remark I made in Chapter 7, and to make the two sorts of objectivity clear, moral facts require a certain sort of psychological makeup and evaluative orientation to see, but not a certain sort of psychological makeup or evaluative orientation to exist.

With this clear, I can say again that I accept the conceptual claim that moral values are meant to be objective. It is the ontological claim which is the problem. To understand Mackie’s basis for the ontological claim, let us turn to the next two key phrases, “very strange” and “utterly different”. The question is what it means to say that categorical imperatives (or the facts, entities, qualities or relations, etc. that ground them) really are “very strange” and “utterly different” to other imperatives (or to the facts, entities, entities, qualities or relations, etc. that ground them).

\textsuperscript{268} Some philosophers deny the conceptual claim, but I do not have space in this chapter to argue against them. See for instance Stephen Finlay, “The Error in the Error Theory,” Australasian Journal of Philosophy 86, no. 3 (2008): 347-369.

\textsuperscript{269} ...other than, of course, the aim of getting at the empirical facts, and the instrumental value of empirical truth as opposed to error (which is valuable no matter what one’s ends are, so long as one has ends).
qualities or relations, etc. that ground them). One sense of strangeness is simply that of
difference, of being *sui generis*. Categorical imperatives are, we will admit, a different
kettle of fish from hypothetical imperatives. Irreducibly normative reasons are very
different from desire-dependant reasons. But so what? The platypus has no trouble
existing even though it is quite unlike any other species. Mathematical facts are quite
unlike empirical facts, moral facts, aesthetic facts, and legal facts, but this does not
undermine mathematics or mathematical existence.\(^{270}\)

We can also admit that when we are in a certain frame of mind moral claims can seem
uncanny or baffling, just as the fact that we will one day die, the beauty of the
Mandelbrot set, or the very existence of the universe, can seem uncanny or baffling.
Stare at a face for long enough, say a word repeatedly, and eventually it will strike you
as odd that it has the properties it does or that it exists at all. But again, this proves
nothing about the existence or non-existence of such things.

There is a third and deeper sense in which morality is queer, in that for our most basic
statements about moral value there is nothing to be said in their justification because
they are *grammatically basic*, as was argued in Chapter 8. The claim that cruelty is
wrong is just such a claim. It is *unaccountably true* – i.e. we cannot explain *why* it is
true, we cannot justify our belief that it is true – and in that sense it is queer compared to
the majority of cases where we can give a justification for believing as we do. Once
again, this queerness is not problematic, but rather a feature of our reaching the point at
which our spade of justification and explanation is finally turned.

We have not yet, I think, got close enough to what Mackie is attacking as “queer”, “very
strange” and “utterly different.” We get closest when we consider two targets: what
Mackie takes to be a queer, puzzling relation of supervenience of the moral on the
natural, and entities which are, though not really queer, posited unnecessarily from
ontological profligacy. We begin with puzzling supervenience.

\(^{270}\) For similar comments see Lee Shepski, “The Vanishing Argument from Queerness,” *Australasian
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 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’?\textsuperscript{271}

Mackie finds the supervenient relation between moral facts and natural facts queer. Let us call this relation the moral-to-natural relation, or MNR. He believes we are owed an explanation or account of the nature of the MNR, without which moral facts are suspect. His argument must be something like this:

(1) Unless we can make sense of the relation MNR between moral facts and related natural facts, moral facts are suspect.
(2) We can make sense of logical entailment and semantic necessity relations in other cases, but MNR cannot be a relation of logical entailment or semantic necessity.
(3) We can make sense of co-occurrence as a relation in other cases, but MNR cannot merely be this kind of relation (since this relation is not consequential).
(4) We can make sense of causal relations between natural facts and (socially guided) reactions to those facts, but MNR is not this kind of relation and moral facts are not meant to be this kind of relata.\textsuperscript{272}
(5) No other understandable account of MNR is correct.
(6) Therefore, moral facts are suspect.

My objection is to the very idea that an account of the MNR relation is needed. Explanations and accounts must stop at some point, at a point where shared understanding is assumed (or explanation fails), on pain of regress. The case of moral

\textsuperscript{271} Mackie, \textit{Ethics}, 41.

\textsuperscript{272} I include this premise because of a claim Mackie makes to support his suspicion of moral epistemology. “How much simpler and more comprehensible the situation would be if we could replace the moral quality with some sort of subjective response which could be causally related to the detection of the natural features on which the supposed quality is said to be consequential.” See Mackie, \textit{Ethics}, 41.
relations is not unique in this respect. As Lee Shepski observes, Mackie is seemingly content with the relation of logical entailment, whereby “some particular argument’s being deductively valid supervenes on its having a certain form.” Yet if we were asked to give an account of this relation – why it is the case that certain forms of argument are valid while others are not – we “might fumble for an illuminating answer, but in the end we would probably not be able to do more than rephrase ourselves and multiply examples.”

Most of us do not find the MNR relation problematic. We understand perfectly well that an action can be wrong because it is a case of “causing pain just for fun,” or lying under oath, stealing another’s property, or adultery, and we do not worry ourselves about explaining what we take to be obvious. Certainly, some MNR relations admit of a little explanation. For instance, we might explain that lying under oath and adultery are both wrong because they are cases of breaking a promise. Perhaps (and only perhaps) we think breaking a promise is wrong, when it is wrong, because it causes another pain. At some point though, these explanations must stop at some class of actions whose wrongness we are unable to explain. This is not problematic. This is the nature of explanation.

We cannot explain why certain very basic classes of action are wrong. Can we, at least, explain the nature of the necessity in the supervenient relation MNR? I think we can say a little, though not much. We can say that the relation is synthetic rather than analytic. The relation between the concept of “wrong” and the concepts of “pain” and “just for fun” is not a conceptual relation, and not one that would make possible an analytic deduction of the relevant claim. Synthetic necessity need not be queer. Olson gives an example of a familiar synthetic necessity: an equiangular triangle is necessarily an equilateral triangle. Of course, it is possible to give an explanation of this relation on the basis of Euclid’s axioms and common notions, but the axioms and notions themselves are not likewise explainable. It would be impossible to explain, for example, this necessity in Euclid: Things that coincide with one another are equal to one another.

274 Ibid., 384.
275 Olson, Moral Error Theory, 93.
Just as we are not able to explain the nature of the necessity of this notion except to say it is mathematical necessity, we cannot explain the nature of the necessity of the MNR except to say that it is moral necessity.

To conclude, there is nothing especially queer about the supervenience relation between moral facts and their natural grounds or criteria. Therefore, let us consider the final possible justification for the charge of queerness: ontological profligacy.

There is an advantage to interpreting what Mackie is objecting to as what has been called “ontological profligacy.” 276 This interpretation allows the argument from queerness to non-existence to run comparatively smoothly, in that there is, prima facie, a plausible connection between a claim that we do (or do not) need to posit the existence of an entity and the claim that the entity does (or does not) exist. However, if this is indeed what Mackie has in mind, I think it would be better to talk of an argument from profligacy than an argument from queerness.

Mackie thinks we can do without positing the existence of moral value, since we can explain everything we need to explain without it by citing facts about our innate and learned patterns of response to “moral” and “immoral” situations.

How much simpler and more comprehensible the [explanatory] situation would be if we could replace the moral quality with some sort of subjective response which could be causally related to the detection of the natural features on which the supposed quality is said to be consequential. 277

Here is an analogy. There was once a time when Norse mythology was taken seriously, and at that time thunder would have been explained as being caused by the god Thor and his actions and temperament. In those days it was necessary to posit the existence of Thor to explain storm phenomena, but there was no other need to posit his existence –

277 Mackie, Ethics, 41.
such as first person encounters with the god, for no one had met any gods. However, with the advent of modern meteorology and electrical theory, we have an alternative and superior explanation of lightning and thunder, and consequently we no longer need to posit the existence of the god Thor. Since the only reason to posit Thor’s existence has disappeared, we are justified in saying that Thor does not exist.

Does the same sort of thing hold true for moral value? This depends on what needs to be explained, or predicted, etc., and whether it can be explained, predicted, etc. without citing moral values. Unfortunately, the questions “what needs to be explained?” and “what counts as a full or satisfying explanation?” do not have answers the moral error theorist and moralist can agree on. For her part, the moralist will claim that there are things which cannot be explained without citing moral value, but these things are themselves characterized in terms of moral value, and are not explained under that description without invoking moral value. The error theorist, for his part, will disagree.

For instance, suppose that the king of a country has perpetrated an abuse against its lower class citizens. As observers, an error theorist and I might agree that one way to explain the actions of the king is to point out that he finds it politically expedient to neglect the needs of these citizens. We can also agree that the king is not very much concerned with the welfare of these citizens in comparison to concerns of expediency. Thirdly, we might agree that the king was raised in a culture which pays little respect to its lower members, and that this was where he picked up his attitude. So far then, we agree. However, as a moralist I will want to add something more to yield a truly satisfying explanation, something that the error theorist will not admit: the king is immoral, shameless, selfish, vicious (or suffers from some other vice or vices) and this is why he did an immoral, shameless, selfish, vicious thing. This is why the king allowed thoughts of expediency to trump thoughts of concern, and this is why he acted immorally. The error theorist only sees motivated behavior in need of motive-based and environment-based explanation, and is satisfied with a story involving a Humean desire-belief pair and, further back, a certain sort of upbringing. I see immoral behavior in need of moral character-based explanation which both adds to and corrects the oversimplification of the error theorist’s explanation.
The mistake then, is to imagine a neutral notion of explanatory need, from which a neutral notion of ontological profligacy and ontological parsimony could be characterized. There isn’t one, and thus there is no shared basis for deciding what ontological commitments we are entitled to.

Talk of ontological commitment and ontological parsimony naturally leads one to think of Quine. Quine’s fundamental ontology was famously sparse, comprising physical objects and sets. I am inclined to think the typical error theorist’s fundamental ontology is rather similar. For instance, in the earlier quote, Mackie thought objective values would be “utterly different from anything else in the universe.” When a philosopher tends to talk about what is in the universe, and does not distinguish this from talk of what is in the world, this often indicates an ontology similar to Quine’s. This is a scientistic ontology, because it prefers and privileges the types of existence that science and mathematics deal in over other everyday senses of existence.

It is true that natural science does not need to, and cannot, use talk of moral properties or relations in its theories or claims. The physiologist who explains a damaged liver cites alcohol as the cause, not the weak moral character of the alcoholic whose body she is examining. However, nothing stops the physiologist from stepping out of her role, and making a moral claim, and the fact that the claim is non-scientific (but not unscientific) does not mean it is any less true.

Believers in moral truths and categorical imperatives can be made to look ridiculous if we neglect the difference between saying we live within a world and live within the natural universe. How (says the scientistic error theorist) does one look at a murder and see the badness there? What is there to see during the event apart from the movement of various limbs and objects, and what is there to see after but a motionless body and blood? Do we see signs of goodness and badness in an arrangement of atoms or along the electromagnetic spectrum? Such skeptical questions express a scientistic

---

278 This distinction is stressed by Martin Heidegger in Being and Time, 60-1.
perspective, but they do not point to a convincing argument that we can or should do without belief in moral values and categorical imperatives.

To conclude, there are various interpretations of what precisely Mackie finds queer about objective moral value claims, such as that they invoke queer supervenience or are ontologically profligate, but the only kinds of queerness that the moralist needs to admit are the unproblematic sorts of queerness.

We have reached an impasse which even Olson recognizes. As he says,

> the issue [of whether irreducible normative reasons are queer] is at a bedrock metaphysical level. It is difficult for error theorists to convince those who find nothing queer about irreducible normativity. And vice versa, of course. So the stubborn response from the non-naturalist seems to leave her and the error theorist in a stalemate, staring incredulously at each other.\(^{279}\)

This is a fair assessment of the situation. If we are to have an argument that there is something fundamentally wrong with moral value, moral requirements, irreducibly normative reasons, and genuine categorical imperatives, we will need to look elsewhere than Mackie. We turn then to Joyce and his nesting argument, which attempts to show that there is good reason to reject talk of objective moral requirements.

### III The Nesting Argument

In *The Myth of Morality*, Richard Joyce argues for a moral error theory. He presents his core argument for the error theoretic position as follows:

---

\(^{279}\) Olson, *Moral Error Theory*, 136. Olson attempts to move past this impasse by refuting companion in guilt arguments which use the existence of other sorts of irreducibly normative reasons to show that moral reasons exist, but his attempt is unpersuasive. He claims, for instance, that whether or not we should adhere to modus ponens in our thinking depends on whether or not we want to occupy the “role” of reasoners, (ibid., 137). This is plainly incorrect. We can resist this view of rationality as we resist error theory.
1. If x morally ought to φ, then x ought to φ regardless of what his desires and interests are.
2. If x morally ought to φ, then x has a reason for φ-ing.
3. Therefore, if x morally ought to φ, then x can have a reason for φ-ing regardless of what his desires and interests are.
4. But there is no sense to be made of such reasons.
5. Therefore, x is never under a moral obligation.\(^{280}\)

Joyce’s book can be divided into three sections. In the first section (Chapters 1 and 2) Joyce argues for premises 1-3, the conceptual part of the argument. That is, these three premises articulate what Joyce believes is involved (inter alia) in something being a *moral* ought as opposed to, say, a prudential ought. In the second section (Chapters 3 through 6, with support from §2.5), Joyce argues for premise 4, which is the ontological part of the argument, because when Joyce says that no sense can be made of “such reasons” he means it to follow that “such reasons” do not exist. The third section (Chapters 7 and 8) considers what we are to do with morality if the core argument is sound, and concludes in favour of moral fictionalism.

With regard to the first section and premises 1-3, I believe Joyce has correctly captured part of the logic or grammar or conceptual structure – call it what you will – of morality. With regard to the second section, I believe that it is open to Joyce to claim premise 4, but not to *argue* for it, because its truth or falsity is beyond argument. Consequently, I will try to show that his arguments in favour of 4 carry no logical weight with the believer in morality, and that his attempt to *argue* for 4 shows he has ignored important parts of the logic (etc.) of morality. The third part of Joyce’s project, moral fictionalism, does not concern me, for it only becomes an interesting thesis if one accepts Joyce’s core argument, which I do not.\(^{281}\)


\(^{281}\) There is a tension between Joyce’s fictionalism and the ontological profligacy reading of Mackie’s metaphysical queerness argument. It is not clear that one can believe that we need to retain moral discourse in fictionalised form and at the same time claim that we have no need of positing the existence of moral oughts. One cannot say that we need to retain moral judgments, and then say that we have no need of the objects of those judgments.
Let us begin with the conceptual part. Joyce argues that belief in a moral obligation is necessarily belief in categorical imperatives that are independent of motivation and interests, and on one reading I agree. However, the meaning of “interests” is potentially ambiguous, given that from a moral point of view there is a sense in which it is always in one’s best interest to be moral. But, so long as Joyce means to limit interests to those things that even the moral error theorist recognizes as interests (perhaps health and survival, pleasure, procreation, satisfaction of desire and completion of plans), I believe he is correct.

Unfortunately, Joyce prefers not to take any stand on the conceptual relation between desires and interests, as he thinks moralists are committed to categorical imperatives that apply irrespective of our ends, i.e. our desires and interests, whether or not our interests are understood in terms of desire satisfaction or not. I do not think this is so. Most ethicists in the Aristotelian tradition, for example, believe that our ultimate end is eudaimonia, a morally or ethically infused type of flourishing. This end is the ground for determining what it is right or wrong for us to do on particular occasions and this end determines what is truly in our interest. With this conception of ends and interests, what we have reason to do is never independent of our ends and interests, and premise 1 turns out false.

I think the best thing to do here is to see that Joyce has tried to argue too much, or at best been less than clear, and allow him to retreat (or stick) to a more modest position where our interests include only our health, survival, pleasure, and so on, but not our moral well-being or ethical flourishing. If we do this, premises 1 to 3 become acceptable, at least in my opinion.

Let us now look at the nesting argument Joyce offers in support of premise 4, that there is no sense to be made of an agent having reasons that are independent of her desires

---

283 Premises 1 to 3 would not be accepted by everyone who rejects error theories. Some philosophers argue that morality does not involve non-institutional categorical imperatives at all, and thus that the conceptual component of Joyce’s argument is in error – see for instance Stephen Finlay, “The Error in the Error Theory,” *Australasian Journal of Philosophy* 86, no. 3 (2008): 347-369. I disagree with Finlay and with the concessive responses to the error theory in general, but leave aside this matter in this chapter because my target is error theories rather than flawed responses to error theories.
and interests, and thus in support of the conclusion that there are no moral reasons. A good way into this argument is to see how Joyce responds to a perceived threat from Rudolf Carnap’s notion of internal and external questions. Here is an extended quote from Joyce.

This distinction between what is accepted from within an institution, and “stepping out” of that institution and appraising it from an exterior perspective, is close to Carnap’s distinction between internal and external questions. Certain “linguistic frameworks” (as Carnap calls them) bring with them new terms and ways of talking: accepting the language of “things” licenses making assertions like “The shirt is in the cupboard”; accepting mathematics allows one to say “There is a prime number greater than one hundred”; accepting the language of propositions permits saying “Chicago is large is a true proposition,” etc. Internal to the framework in question, confirming or disconfirming the truth of these propositions is a trivial matter. But traditionally philosophers have interested themselves in the external question – the issue of the adequacy of the framework itself “Do objects exist?”, “Does the world exist?”, “Are there numbers?”, “Are there propositions?”, etc. Carnap’s argument is that the external question, as it has been typically construed, does not make sense. From a perspective that accepts mathematics, the answer to the question “Do numbers exist?” is just trivially “Yes.” From a perspective which has not accepted mathematics, Carnap thinks, the only sensible way of construing the question is not as a theoretical question, but as a practical one: “Shall I accept the framework of mathematics?”, and this pragmatic question is to be answered by consideration of the efficiency, the fruitfulness, the usefulness, etc., of the adoption. But the (traditional) philosopher’s questions – “But is mathematics true?”, “Are there really numbers?” – are pseudo-questions. By turning traditional philosophical questions into practical questions of the form “Shall I adopt...?”, Carnap is offering a noncognitive analysis of metaphysics. Since I am claiming that we can critically inspect morality from an external perspective – that we can ask whether there are any non-institutional reasons accompanying moral injunctions – and that such questioning would not amount to a “Shall we adopt...?” query, Carnap’s position represents a threat.

Joyce challenges the claim that one cannot ask the question whether moral requirements really exist from outside morality except in the pragmatic sense. He recognizes that this may be the case for some other types of existence claims (for instance, mathematical

---

285 Joyce, Myth of Morality, 45-46.
existence claims), but affirms the possibility that for other types of claims the external
question can be asked and answered from within an external framework which the inner
framework nests within.

One of Joyce’s examples of nesting is Maori tapu claims – claims, roughly speaking,
about what is forbidden or taboo. Joyce’s thought is that the tapu claim framework is
nested within a causal claim framework. If the causal claims associated with something
being tapu are found to be false, then that thing will be found not to be tapu. This can be
affirmed by anyone who accepts the causal framework – roughly, the framework of
science – and thus it is possible for someone operating outside the tapu framework to
reject the tapu framework for empirically substantive rather than pragmatic reasons.

Joyce then proceeds to argue that the same thing can be done with respect to the moral
framework. The external framework that functions as science does for tapu-ness is, he
thinks, practical rationality, for “the wider non-moral notion of having a reason comes
from practical rationality”. Joyce’s vision of practical rationality is non-Humean
practical instrumentalism, according to which one only has reason to do something if it
serves one’s desires or interests. (By contrast, Joyce takes Humean practical
instrumentalism to be the claim that one only has reason to do something if one believes
it will satisfy one’s present desires.) If Joyce can show that practical instrumentalism
(whether Humean or not) is the correct account of rationality, he believes he will have
shown that morality can be criticized from an external standpoint, and that from that
standpoint it will be obvious that moral requirements do not exist.

Now let me respond to this argument. The issue turns on what notion of practical
rationality it is correct to nest moral reason claims within. To settle this issue, we must
first distinguish between a grammatical and a substantive notion of practical rationality.
By grammatical notion, I mean that notion of rationality which one must understand or
at least abide by in order to talk meaningfully and comprehensibly about such things as
what actions are rational and what one’s reasons for acting are. This is a very thin

286 Ibid., 1-5, and 47-49.
287 Ibid., 48.
288 Ibid., Chapter 3.
It requires, for instance, that one endorse claims (or truisms) that follow these schemas:

a. If x has overriding reason to $\varphi$ (for some $\varphi$), then $\varphi$-ing is the rational thing for her to do.

b. If x has better reason to $\varphi$ than to $\psi$ (for some $\varphi$ and $\psi$), $\varphi$-ing rather than $\psi$-ing is the rational thing to do, if those are the only options.

c. If x accepts that there is overriding reason to $\varphi$, but intentionally chooses not to $\varphi$, x is being irrational.

d. If x has reason to $\varphi$ all things considered, then x has overriding reason to $\varphi$.

e. If x ought to $\varphi$ (or ought to have $\varphi$-ed), then x should $\varphi$ (or should have $\varphi$-ed).

f. If R was x’s reason for $\varphi$-ing, then R was why x $\varphi$-ed.\(^{289}\)

g. Barring self-deception and failure of memory, if x has $\varphi$-ed and he believes his reason for $\varphi$-ing at the time was R, then R was his reason for $\varphi$-ing, and he $\varphi$-ed for that reason.\(^{290}\)

Someone who denies claims that follow these schemas simply does not understand the language well enough to know what reasons and practical rationality are, perhaps because she speaks another language or she has read too much philosophy and forgotten how the terms involved are actually used. These claims are not meant to be substantive – they are those claims about the nature of reasons that supporters of morality and moral error theorists like Joyce can affirm together. However, if any of the above claims is not of this sort, then I withdraw it. Similarly, if any of the claims needs refinement to deal with counterexamples before it can be universally endorsed, then let it be so refined.

The grammatical framework of practical rationality is one which the moral framework can and does nest within, in the same way that all quantitative scientific theories, true or

\(^{289}\) Of course, one can have a reason for performing an action and yet do it for some other reason, for no reason at all, or not do it at all. However, “my reason (at the time) for $\varphi$-ing was R” is not the same as “the reason I had (at the time) for $\varphi$-ing was R”. The former entails both that I took that reason to heart and that I acted on it. The latter does not.

\(^{290}\) See Chapter 4, Section II.
false, nest within the framework of mathematics. However, the grammatical practical-rationality framework is not one from which morality can be criticised unless believing in moral reasons necessitates affirming something contrary to the external framework, which it plainly does not, because a moralist understands and affirms the truisms above. In the same way, one cannot offer a mathematical criticism (as opposed to an evidence-based criticism) of a scientific theory unless it contains mathematical errors. So long as our theories are mathematically sound, there is no space for mathematical criticism of one physical theory in favour of another. This fact explains why the external framework Joyce wishes to use to criticise morality is something substantive: non-Humean practical instrumentalism.

However, although morality is indeed at odds with practical instrumentalism, there is no reason to believe that it is nested within it. In fact there is every reason to think that it is not. We can see this quite simply: a moralist will refuse to affirm that she only has reason to do what serves her desires and interests (narrowly conceived). Therefore she is not committed to the practical-instrumentalist frame, and her moral framework is not nested within that frame.

Joyce’s criticism of morality is in reality a substantive rejection of morality from outside morality that a believer in morality is perfectly entitled to reject in turn. If this is the nature of Joyce’s argument then, with respect to moral reasons, it is misleading of him to claim that “there is no sense to be made of such reasons”, as he does in premise 4, as if the moralist were confused and unclear. Rather, Joyce’s argument amounts to nothing more than a denial of the existence of moral reasons and an endorsement of a substantive view of practical rationality which – quelle surprise – has no place for morality. The latter account of practical rationality does not give us an independent reason to deny the existence of moral claims. It is simply the same denial expressed differently and more specifically, and a believer in morality does well to reject the claim in both forms.

To avoid this conclusion, Joyce needs some further argument in support of his nesting thesis and premise 4. He is sensitive to the fact that some will find fault with his account
of normative reasons on the grounds that it is not the only one possible, that “there are alternative versions available”.

He recognizes that his account leaves no place for morality, and he admits that an alternative account with a place for morality that also retained the virtues of his account would be preferable. He therefore wishes to show that no such account exists. He follows two “strategic avenues” in attempting this. The first consists of showing that “non-instrumentalist rationality necessarily commits an error.” Let us call this the Williams strategy, as his argument is adapted from one in Bernard Williams’ article “Internal and External Reasons”. Interestingly Joyce admits that the Williams strategy is the only argument he knows of that has such a conclusion. The second strategy, which we will consider in Section IV, he calls the genetic strategy.

As I explained in Chapter 7, Williams argued against the existence of external reasons, which roughly speaking are reasons that are independent of an agent’s desires. Joyce is unhappy with the scope of the argument since, contra Williams, he feels that institutional reasons are external reasons and genuinely existent. However, by restricting the scope of Williams’s argument to showing that there are no *non-institutional* external reasons Joyce believes he has the argument he requires. So, what argument does Joyce have in mind? He says:

Williams’s argument against external reasons...revolves around their inability to motivate those to whom they are ascribed.

In Chapter 7, I gave my rebuttal to this argument. I called the argument the possible explanation objection. I will repeat what I said there to characterize the argument, and give a brief summary of my response. Williams argues that an agent’s reasons must be able to explain and motivate her actions if they are to be her reasons, but so-called external reasons cannot do so without being unveiled as, or depending on, internal reasons that are doing the real explaining and motivating work. Thus such so-called external reasons are either not reasons, or are internal reasons. I suggested that what

---

292 Ibid., 101.
293 Joyce, *Myth of Morality*, 104.
Williams believed in was a link to explanation requirement, or LER, which he thought no external reason could satisfy, namely:

If an agent has a reason R to φ, then it must be possible for her to φ for reason R, and if she does so (that is, if she φs for reason R), the fact that she has this reason to φ must be able to figure in an explanation of her action.\(^{294}\)

In response, I argued that there are different ways to read “possible” in the LER. Two of these ways of reading “possible” yield versions of the LER that could be acceptable (at least when discussing reasons related to moral obligations), and these ways allow for the existence of external moral reasons. First of all, we can understand what is possible for an agent as what she can do now or come (in time, with training, in the right circumstances, with maturity, after conversion etc.) to do. Secondly, one can understand what is possible for an agent to do as what it is appropriate or acceptable to expect or demand of an agent. The details of these interpretations, which one is preferable, and their possible interconnection, will not be discussed here. What was shown in Chapter 7 was that the existence of these interpretations means the LER and the possible explanation objection are not obviously problematic for believers in external reasons.

As I said earlier, Joyce’s internal reasons argument is the only one he knows of for the conclusion that the correct account of practical rationality is instrumentalist. Without such an argument, he cannot justify his contention that moral reason claims must be nested within an instrumentalist version of rationality. Likewise, he lacks justification for saying that there is no sense to be made of reasons that are independent of desire. Thus premise 4 is undefended.

**IV The Evolutionary Argument**

In Chapter 6 of *The Myth of Morality*, titled “Morality and Evolution”, Joyce considers the possibility that an evolutionary debunking argument (or EDA) may undermine our

\(^{294}\) See Chapter 7, Section IV.
confidence in the truth of our moral beliefs. This is a topic to which he has returned more than once.295

The idea of a project of debunking morality covers a number of related but separable projects, so it is fortunate that Joyce is aware of the variety and clear about the nature and limits of what he hopes to argue. He does not claim that an EDA will show that all moral beliefs are false. Instead, he claims that an EDA should make us doubt whether we have justification for our moral beliefs. That these two claims are distinct should be obvious from the fact that a belief can be unjustified while still true. A belief may be true and yet unjustified because it is formed by an unreliable process such as consulting horoscopes or trusting a habitual liar.

Joyce presents his evolutionary debunking argument across a number of pages.296 I believe what follows is a fair representation of that argument. Let B_m refer either to the general belief that there are moral requirements or to some specific belief in a moral requirement, such as the belief that one ought not to kill innocents.

P1: The explanation for the widespread belief that B_m is that this belief is the product of natural selection.

P2: In a world where there were moral requirements, belief that B_m would be widespread, because that belief would be the product of natural selection, and in a world where there were no moral requirements, belief that B_m would be widespread, because that belief would be the product of natural selection.

C1: For all those who believe B_m, the explanation why they believe B_m is independent of B_m’s being true.

P3: If the explanation why we believe a belief is independent of its being true, then our belief is unjustified.

C2: The belief B_m is unjustified.

296 Joyce, Myth of Morality, Chapter 6, especially §6.4.
I have attempted to present Joyce’s argument as fairly as possible. In doing so, I have tried to ensure the argument presented is valid and faithful in content and spirit to Joyce’s writings. Assuming I have succeeded, we can turn to considering the truth of the premises. I do not wish to raise any objections to P3. This is superficially plausible, and certainly far less controversial than P1 and P2.

Let us begin with P1. The scientists who make use of explanations in terms of natural selection (let us call these explanations NSEs) do not give NSEs of beliefs, but of phenotypes including morphological features, metabolic processes, and instinctive behaviour patterns (such as reflexes). Joyce will admit this, but argue that there is no reason why NSEs cannot be extended, at least sometimes, to cover beliefs as well. However, there is very good reason to object to this move.

Although Joyce does not talk of reflexes – the closest he gets is using the term “hardwired” – it will be very helpful to use the term in articulating what is wrong with P1. Beliefs should not be confused with reflexes. Indeed, beliefs cannot be reflexes. This is a grammatical remark, rather than a substantive claim. One can criticize (or praise) someone for holding a particular belief, and one does so by judging whether or not he has exercised epistemic virtues in coming to hold that belief. One does not, in general, criticize (or praise) someone for exhibiting a particular reflex. This shows that beliefs belong in the logical domain of rationality, and this is a domain in which we are held responsible for what we believe and do. By contrast, no one is held responsible for his reflexes, for the most part.

Crucially, if some piece of verbal behaviour that looks like an expression of belief is revealed to be a reflex over which the speaker has no rational control, then it is not in fact the expression of a belief. Similarly, if we accept that someone holds a certain belief, we cannot say that she believes it by reflex.

Here are some reflexes: the scratch reflex, the blink reflex, the photic sneeze reflex, the withdrawal reflex, the suckling reflex, and the shivering reflex. Let us grant, for sake of
argument, that other reflexes exist with a more important relationship to morality. For instance, let us hypothesize that there is a disgust reflex (involving gagging, withdrawal, etc.) to the thought of sexual relations with a relative (and certain other sexual behaviours), and a sympathy reflex (involving a tendency to mimic facial expressions of pain and pleasure seen in others and to experience sympathetic feelings), and so on. Call these reflexes *morally relevant reflexes*.

Now, I can admit that there are NSEs for reflexes, including morally relevant reflexes should they exist. Also, I can admit that in all probability a human whose genetic abnormality meant he was born without any morally relevant reflexes would be highly unlikely or unable to understand or accept the existence of moral requirements as an adult. Nonetheless, this would not mean that there is an evolutionary explanation for belief in moral requirements.

It is possible for an NSE of a reflex to play a partial role in explaining someone’s belief, but there will always need to be a second element that goes beyond selective pressures to make the explanation function as a correct explanation. This second element will refer to the epistemic virtues or vices of the believer, for if there is no such second element, the belief-like behaviour in question will be revealed as mere reflex, and not a belief at all, in accordance with the comments above. So, for instance, suppose that there were an evolutionarily selected reflex to feel disgust at homosexual activity, and suppose that Bob holds the belief that homosexuality is immoral and homosexuals should be shunned, or punished, etc. If we say that the explanation for Bob’s belief is that the disgust reflex was naturally selected for, we will not have explained Bob’s *belief* unless we also add that Bob is so lacking in epistemic virtue that he is willing to form beliefs about what is immoral solely on the basis of reflexes like disgust. If we suppose that Bob *has* done his duty to consider the importance of his reflex, and concluded *on some other basis* that the right belief to hold regarding the object of the reflex is the one suiting that reflex, i.e. that “Homosexuality is immoral” is the right belief to hold with respect to homosexuality, the NSE will no longer be the core explanation for his belief. He will believe that homosexuality is immoral for *some other reason.*
That is, NSE’s of moral belief only work as explanations if we add that the believers involved are thinkers who have not done their epistemic duty with respect to the right to hold the belief in question. If the thinkers have done their duty, the NSE is not the explanation for their belief. At most, NSEs are suggestive of a possible attitude or belief, and explain the prevalence of that belief among the intellectually lazy. Certainly something like this may lie behind racist and sexist beliefs - that is, behind prejudice.

Earlier I said that we are not responsible for our reflexes, “for the most part”. I added this phrase because many if not all reflexes can be overcome or blocked. The blinking reflex can be stopped by surgically removing the eyelids, or damping the associated muscles, and in general one could overcome many individual reflexes if one were willing to go so far as to use surgical or pharmaceutical methods. Similarly, there are some reflexes over which we may with time, effort, and training come to exert a measure of conscious control. Therefore, we could in principle be held responsible for exhibiting a reflex if it could be overcome in the ways above, and there were good reason to expend the time and effort to overcome it.

However, this caveat is of no help to Joyce – he cannot claim that moral (or any other) beliefs are reflexes over which we have a measure of control and that these beliefs can be explained by NSEs. If a person has failed to overcome a reflex she could and should have, that will mean that there is no pure NSE for her indulging in the reflex. Just as with belief, the explanation of her behaviour will involve reference to her failure to fight that reflex as she should have, assuming she should have.

What the above goes to show is that there can be no NSE for beliefs, because beliefs are things over which we have rational control. There are NSEs for such things as reflexes, but beliefs cannot be reduced to reflexes (or the other things) for which NSEs exist.

How might Joyce respond to the above criticisms? One way to rescue P1 which error theorists like Joyce would not accept would be to re-characterize the nature of morality as a matter of reflexes rather than beliefs, yielding P1’.
P1’: The explanation for the widespread moral reflexes (of sympathy and disgust towards, e.g., another’s pain and murder respectively) is that they are the product of natural selection.

Error theorists cannot accept this. Error theorists argue that moral beliefs are both unjustified and false, and if morality is a matter of reflex rather than belief, one cannot be unjustified or in error for exhibiting a moral reflex. As such, a necessary part of Joyce’s error-theoretic position is a rejection of moral non-cognitivism.297

Is there any other way to rescue P1? Joyce could take up my earlier comment that NSEs of belief need a second element, and introduce one, by changing P1 to P1’’.

P1’’: The explanation for the widespread belief that Bm is (1) it is the product of a natural selected reflex or tendency that (2) believers fail to adequately weigh the value of, and thus fail to rise above.

This move is possible for Joyce, but if he takes it, he obviously owes us an argument to show that believers in Bm have failed to do their epistemic duty. But worse, seeing this, one realizes that the larger evolutionary debunking argument is a mere distraction from the real issue. If Joyce has an argument for P1’’, then he already has an argument that belief in Bm is unjustified supporting that premise, and since the claim that Bm is unjustified is the very conclusion of the evolutionary debunking argument (EDA), the overall argument is circular.

I can see no other way for Joyce to rescue P1, and therefore the argument fails. However, it is worthwhile examining P2, since it is also flawed. Here is the claim:

P2: In a world where there were moral requirements, belief that Bm would be widespread, because that belief would be the product of natural selection, and in a world where there were no moral requirements, belief that Bm

297 Joyce, Myth of Morality, §1.2.
would be widespread, because that belief would be the product of natural selection.

Premise 2, as I have written it, expands on premise 1, and therefore the criticisms directed at P1 apply to it. That is, P2 like P1 includes the mistaken idea that beliefs can be explained as resulting from natural selection without reference to the exercise of epistemic virtues and vices in the formation of those beliefs. I will not dwell further on this point. Instead, I am interested in the second half of the claim, which requires of us that we imagine a world in which there are no moral requirements and yet one in which moral belief is widespread because of natural selection.

The issue that concerns me is that we may be being asked to imagine something incoherent. Joyce is aware that this challenge can be raised, and in a passage responding to a challenge from Robert Nozick attempts to face it.

When we imagine the world at which this belief is false – at which there are no moral requirements at all – and maintain that humans will nevertheless hold the belief for evolutionary reasons, Nozick will … need to claim that the thought experiment is incoherent, on the grounds that moral requirements necessarily exist. But such a substantive theoretical claim is far from self-evident.298

Obviously, Joyce believes that there are no moral requirements. Others believe that there are. Let us set aside the question of whether the actual world is as Joyce sees it or not, and ask a different question. When I try to imagine a world in which there are no moral requirements, how do I know that I am imagining it correctly? What needs to be the case for me to know whether I am imagining the world with or without moral requirements? This is important because Joyce requires of me that I see that in a world without moral requirements, natural selection will still lead people to hold moral beliefs. However, not every world I imagine without moral requirements will allow this. For instance, I can imagine a world in which there are plants and insects, but where there are

no higher forms of life because the conditions are such that these forms cannot evolve. In this world there are no moral requirements, in the sense that no creature is morally required to do anything, because plants and insects are not subject to moral requirements. However, plainly, this is also a world in which humans do not (“evolve to”) hold moral beliefs because there are no humans to hold them.

Joyce must wish me to imagine something else – of course, he thinks it will do to imagine our world, but that is no help for me if I think this world is one in which moral requirements exist. I need to know what I am to imagine as different between the moral and amoral worlds and what I am to imagine the same, at least as it bears on the question of what humans will evolve to do and, supposedly, believe.

Suppose for example I imagine a world in which there are creatures that are very like humans in many respects, but these creatures do not die or suffer. Now, it seems plausible that in such a world these “humans” will not be under any moral requirements, because morality is intrinsically focused on issues related to human well-being and human suffering. In a world where the “humans” cannot die or suffer, they cannot fare well by being free of death or suffering either, and there can be no injunctions to benefit or harm other “humans” as a result. However, what role can natural selection play with respect to these “humans”? It seems highly implausible that an immortal being could evolve in any way at all, let alone evolve moral beliefs, because there are no selection pressures. So plainly this too is not the world Joyce wishes me to imagine.

To speak bluntly, it is possible that the only worlds I can coherently imagine in which there are no moral requirements are also the worlds in which humans would not “evolve” moral beliefs. If Joyce wishes to argue otherwise, he will need to give a clear account of what he means us to imagine.

Unsurprisingly, Joyce gives no such account. Similarly, he repeatedly speaks of the failure of moral beliefs to track the truth, but never explains what he thinks it would be like for moral beliefs to track the truth so that we can understand what would be involved in checking whether things stand in that way or not.
As a last ditch attempt to save his thought experiment in response to a concern over its conceivability, Joyce might demand an account of the nature of moral correspondence from the moralist, and claim that *this* is what he means to deny. That is, he might ask believers in moral requirements to explain the way they conceive of moral claims tracking the truth, and then say that he means to deny that our world is a world in which such a correspondence occurs or makes sense.

However, this will not work. Firstly, moral claims are not true in virtue of their correspondence to states of the world, at least not in the same way that empirical claims are true in virtue of correspondence to features of the *universe*, which would admit of explanation in terms of experimental testing and confirmation. We cannot give this kind of account of moral correspondence any more than we can give such an account of mathematical correspondence. I believe that “3+6=9”, and I can explain what I mean by it (to a child, say, by showing her how to count it out), I can justify it (to an undergraduate mathematics student, in terms of Peano’s axioms), but it would be a misunderstanding of the nature of mathematics to think that “3+6=9” is true because it corresponds to something in the nature of the universe, and a mistake to think that an account of our knowledge of this correspondence can be given in terms of empirical confirmation, say in the repeated combining and counting of sets of physical objects. If we must talk of “3+6=9” being true in terms of correspondence to fact (and we needn’t), it is correspondence to a *mathematical* fact in the *world of mathematics*.299 Similarly, what John Searle has termed *social facts* (e.g. “The cappuccino costs $3.50”, “Reagan was the President of the USA”) do not correspond to facts of the universe.300 If we must talk about them corresponding to anything, it is to facts of the social world. (Of course, these are not necessary truths as mathematical truth are.)

299 Paul Johnston, for one, thinks it is a mistake to see the correctness of moral claims as being a matter of correspondence. See Paul Johnston, *The Contradictions of Modern Moral Philosophy: Ethics After Wittgenstein*, (London and New York: Routledge, 1999). I take Johnston’s point, but think that distinguishing correspondence with worldly facts (facts about the world) from correspondence with empirical facts (facts about the universe) is another, perhaps superior, way to mark what is wrong with some forms of moral realism and their denials.

Then there is a second problem. For facts that are necessarily true, one cannot take an account of their truth and use it to explain what it would mean to deny their truth. Though I can explain the truth of “3+6=9” on the basis of Peano’s axioms and a few definitions, this does not mean I can give a parallel account of what it would be like for “3+6=10” to be true. “3+6=10” is false in virtue of its being a denial of a mathematical truth, not in virtue of its being a comprehensible description of how things might have been but happen not to be. As Joyce’s quoted response to Nozick makes plain, Joyce is aware that if one believes that “moral requirements necessarily exist” (as mathematical truths necessarily exist) one can maintain that his “thought experiment is incoherent.” He finds the premise “far from self-evident.” However, I find it self-evident that a world in which “humans” are not under moral requirements is simply a world in which there are no humans. It is incoherent to imagine an amoral human world.

At this point, the moralist and error theorist have reached an insuperable impasse. Since my goal in this chapter has been to show that error theoretic arguments need not trouble the believer in morality, and my goal has not been to show that the error theory is demonstrably false, a genuine impasse is an acceptable place to end the discussion.

V Conclusion

We have examined three arguments in support of the moral error theory, the motivational queerness argument, the metaphysical/epistemological queerness argument, and the nesting argument, and one argument against there being justified moral belief, the evolutionary debunking argument. In each case the arguments have been found to be unpersuasive. The moral error theoretic position is a substantive position on ethical matters that constitutes a logically coherent option, but the error theorists considered in this chapter have failed to show that it is more than this, that it is a position which is rationally superior to belief in genuine moral requirements.
10 Conclusion

We have examined a number of claims that centre on the relations between morality, reason, and motivation. Many of them, it has been argued, are conditionally permissible. Examples, in brief, are:

Moral judgments entail motivation. (Chapters 2-4)
Our reasons for action are desire-belief pairs. (Chapter 5)
Our desires and beliefs explain why we act. (Chapter 5)
Reasons for action are causes of action. (Chapter 6)
Reasons for action must be potential explanations of action. (Chapter 7)
There is a reason to be moral. (Chapters 7-8)

Other claims, it has been argued, are misleading, or simply false, because they ignore or covertly distort the meaning of our words and the grammar of our language games:

Moral judgments are not beliefs. (Chapters 2-3)
Desires and beliefs are brain states. (Chapter 6)
Immoral actions are always irrational. (Chapter 8)
There is no sense to the idea of moral facts or requirements. (Chapter 9)
Our moral beliefs are the product of evolution. (Chapter 9)

Finally, other claims are substantive. To accept one of them, or deny one of them, is to take a stand on what matters in life. Philosophy cannot tell us which of them are true:

---

301 See Chapter 1, Section IV.
302 These claims have been put into short form. For a better, less ambiguous, characterization of the claims, see the associated chapters.
Our motivations determine what reasons there are for us to act. (Chapter 7)
There is a reason to be moral. (Chapters 7-8)\textsuperscript{303}
Immoral action is unreasonable. (Chapter 8)
There are no moral facts or requirements. (Chapter 9)

Which of these we take to be true is revealed in how we live our lives.

\textsuperscript{303} This claim is both conditionally permissible, and (in its permissible version) substantive. There are no reasons to be moral if one means reasons from beyond morality for respecting morality’s intrinsic value and acting on that basis. This is a grammatical truth. However, many of us believe there are reasons to be moral in a different, and substantive, sense: acting morally is the right thing to do, since morality is of intrinsic value.
Bibliography


Hacker, P. M. S. “Davidson on the Ontology and Logical Form of Belief.” *Philosophy* 73, no. 283 (1998), 81-96


