THEORETICAL AND PRACTICAL RATIONALITY: TOWARDS A UNIFIED ACCOUNT

by

Jonathan D. Payton

Submitted in partial fulfilment of the requirements
for the degree of Master of Arts

at

Dalhousie University
Halifax, Nova Scotia
August 2011

DALHOUSIE UNIVERSITY

DEPARTMENT OF PHILOSOPHY

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled "THEORETICAL AND PRACTICAL RATIONALITY: TOWARDS A UNIFIED ACCOUNT" by Jonathan D. Payton in partial fulfilment of the requirements for the degree of Master of Arts.

Dated: August 15, 2011

Supervisor: _____

Readers: _____

_____

DALHOUSIE UNIVERSITY

DATE: August 15, 2011

AUTHOR: Jonathan D. Payton

TITLE: Theoretical and Practical Rationality: Towards a Unified Account

DEPARTMENT OR SCHOOL: Department of Philosophy

DEGREE: MA CONVOCATION: October YEAR: 2011

_____

Signature of Author

TABLE OF CONTENTS

ABSTRACT

This work is dedicated to the development of a unified account of both theoretical and practical rationality. I adopt a particular view of evaluative properties, according to which entities are evaluated as good or bad according to how well they fulfill the constitutive functions of their kinds. I argue that the function of belief is to accurately represent reality, while the function of action is to satisfy the agent's desires. These functions fix the goodness- or success-conditions of belief and action. With these functions in place, I adopt a reliabilist conception of reasoning which evaluates reasoning processes by how well they allow us to achieve the constitutive aims of belief and action. Moreover, I argue that the process of determining which action will best satisfy our desires is a cognitive matter – non-cognitive states like desire do not actually provide the agent with reasons.

LIST OF SYMBOLS AND ABBREVIATIONS USED

Φ	Mental State/Action Variable
Ψ	Mental State/Action Variable
Ω	Mental State/Action Variable
$=$	…is identical to…
$\neq$	…is not identical to…
$\wedge$	Conjunction
$\vee$	Disjunction
$\supset$	Material Conditional
$\neg$	Negation
$\exists$	Existential Quantifier
$\forall$	Universal Quantifier
$\Box$	Necessity Operator
$\Diamond$	Possibility Operator
x	Entity Variable
y	Entity Variable
X	Property Variable
p	Proposition Variable
q	Proposition Variable
r	Proposition Variable
iff	…if and only if…
(M/E)	Means-End Principle
(M/E*)	Modified Means-End Principle
IOT	Ideal Observer Theory

ACKNOWLEDGMENTS

Special thanks have to go to my supervisor, Greg Scherkoske. He was a constant source of guidance and encouragement. He pushed me when I was on to something good, and brought me back down to Earth when the project seemed to be getting out of hand (I shudder to think of how the last few months might have gone if I tried to pull off my original project). He pointed me in the direction of useful source materials, and the thesis would not have reached the point that it is has if I hadn't had him there to help me get caught up – as much as is possible – on the last twenty or so years of research in these areas. And, of course, he provided stimulating objections along the way. Having to deal with these objections made the final product much better in the end, although I'm not at all convinced that I've been able to answer them to my satisfaction. However, Greg's assistance and encouragement did not begin with my thesis. Throughout my year at Dalhousie he was incredibly supportive, and helped to make this year even more productive than it otherwise might have been.

In addition, thanks go to my readers, Duncan MacIntosh and Kirstin Borgerson. They both proved to be attentive to the work, and supportive. Moreover, like Greg, each of them provided encouragement and support throughout my year at Dal. Duncan in particular showed great enthusiasm for my work throughout the year, an enthusiasm which motivated me to work even harder to make my papers and chapters the best that I could (or at least as best as I could make them before the due dates).

This thesis is the result, not only of the criticism and encouragement provided by my supervisor and readers, but of all those who gave me this sort of support throughout my time at Dalhousie. This department is incredibly cohesive, in that professors and students of all levels intermingle with one another, and create not only a stimulating academic environment, but a real sense of community. In particular, I want to thank, in alphabetical order: Darren Abramson; Adam Auch; Christina Behme; Steven Burns; Richmond Campbell; Samantha Copeland; Mike Hymers; Chike Jeffers; Heather McAteer; Peter Schotch; Meredith Schwartz; Kira Tomsons; Tom Vinci; and Marisa Webster. I only list them off so quickly to save space: I could go on for paragraphs at a time about each of these individuals.

Thanks also go to those professors at York University who supported me during my undergraduate degree, provided me with the opportunity to take advanced courses, invited me to co-write papers, and encouraged me to pursue philosophy at the graduate level. As with the community at Dal, I could on about each of these individuals, but I will simply name some names here: thank you to Lorraine Code, Michael Giudice, Muhammad Ali Khalidi, Alice MacLachlan, Robert Myers, Judy Pelham, Patrick J. J. Phillips, and Claudine Verheggen.

Finally, thanks go to my family and friends. First, my mother and father, Richard and Ruth Payton, and my brothers Geoffrey and Michael. I'm sure more than one of them had some reservations when I announced that I wanted to pursue a career in philosophy instead of becoming a lawyer, but they have been supportive and encouraging over the last few years. My friends Mike Crane and Ayesha Shah in particular have also been supportive in all areas of my life, and I might have gone crazy years ago without them.

CHAPTER 1: INTRODUCTION

Throughout this work, I will be attempting to lay the groundwork for a unified account of both theoretical and practical rationality. That is, I will be trying to develop an approach which can be used to explain, for those cases in which we have reasons for our beliefs, why it is that we have those reasons, and to explain, for those cases in which we have reasons for our actions, why it is that we have those reasons. The approach is unified, because it attempts to provide a single kind of explanation in each case. Rather than developing separate accounts of theoretical and practical rationality – even accounts which are tied together by structural analogies, for instance by treating desires in practical reasoning as behaving similarly to beliefs in theoretical reasoning – my aim is to show how we might provide a single account which can be applied to both beliefs and actions. My account, as presented here, is quite general. I will say very little about what particular reasons we actually have. I will be concerned, rather, with providing a general theory of when and why we have reasons for belief or action.

The account is based on certain assumptions that I will be making about the semantics of evaluative terms and the metaphysics of evaluative properties. The semantic thesis I adopt is derived from Peter Geach.[1] He draws a distinction between two different types of adjectives, which he calls 'predicative' and 'attributive'. In defining these terms, Geach writes, "I shall say that in a phrase 'an A B' ('A' being an adjective and 'B' being a noun) 'A' is a (logically) predicative adjective if the predication 'is an A B' splits up logically into a pair of predications 'is a B' and 'is A'; otherwise I shall say that 'A' is a (logically) attributive adjective."[2] For example, the adjective 'red car' is predicative,

[1] See Geach (1956).
[2] Ibid., p.300.

because it is represented in the predicate calculus by two predicates, 'is red' and 'is a car'. From '*x* is a red car' we can derive each of '*x* is red' and '*x* is car'. Likewise, we can recombine 'red' with other predicates that *x* satisfies to get new true sentences: '*x* is a red Cadillac', '*x* is a red machine', etc.

An attributive adjective does not have either of these features. Consider 'small'. From '*x* is a small elephant' we cannot derive '*x* is small'. Because of this, we cannot recombine the pseudo-predicate 'is small' with other predicates that *x* satisfies without risk of generating false sentences: '*x* is a small mammal', for instance, will be false; even a small elephant is pretty big for a mammal.

'Good', Geach thinks, is attributive: it does not detach itself from the adjective in which it is placed and recombine with other predicates. From '*x* is a good chess player' and '*x* is a tennis player' it does follow that *x* is good tennis player. 'Good' does not function logically as a first-order predicate. Rather, it functions as a predicate-modifier: it takes first-order predicates as inputs and yields new first-order predicates as outputs.[3] It is because 'is a good chess player' is logically simple – being a first-order predicate – that we cannot distill 'good' out of it.

The metaphysical thesis I assume derives from Judith Jarvis Thomson's recent book *Normativity.* She suggests two theses:

(1) Being a good K is being good *qua* K.

(2) There is such a property as being good *qua* K only if K is a goodness-fixing kind.[4]

[3] For a helpful discussion of predicate-modifiers and other elements of natural language along with their logical analogues, see Gamut (1991), ch. 4.
[4] Thomson (2008), pp.19-21.

However, she does not elaborate too much on what a goodness-fixing kind is. She points to intuitive distinctions ('toaster' and 'seeing-eye dog' are goodness-fixing kinds, but 'pebble' isn't), but doesn't say much beyond: goodness-fixing kinds are those kinds which contain, as part of their essence, a determination of what counts as a good instance of that kind.

I assume a functional account of goodness-fixing kinds: there is a property of being a good K *qua* K iff. the kind 'K' has, as part of its essence, a function. Artifacts are fairly clear examples: toasters are meant to produce toast of a certain level of quality, and so this is the function of each individual toaster. A toaster is good iff. it is sufficiently reliable at achieving the function of toasters. 'Good' doesn't mean 'perfect', so we don't require instances of the kind 'toaster' to *always* produce really excellent toast; a certain level of quality is sufficient for goodness *qua* toaster. This functional account is inspired by J. David Velleman's approach to theoretical and practical rationality.[5] He argues that we can construct accounts of the normative features of theoretical and practical reasoning – what makes for good or bad reasoning of these kinds – by determining what the constitutive aims of belief and desire are.

With this account in place, the project of my thesis is, in outline at least, fairly straightforward. Terms like 'rational' and 'reasonable' are normative or evaluative terms, and so must be explained in terms of some function. The obvious starting point for theoretical rationality is the function of belief: since theoretical reasoning is reasoning about one's beliefs – whether they are good or bad, properly or improperly held, etc. – we should try to find a function for the kind 'belief' which will explain why certain beliefs are rational or irrational. Likewise for practical rationality: since practical reasoning is

[5] See Velleman (1996) and Chapter 3 of this work.

reasoning about actions – whether they are good or bad, which action we should perform, etc. – we should try to find a function for the kind 'action' which will explain why certain actions are rational or irrational.

In Chapter 2 I defend the view that truth is the aim of belief against some objections to it that have recently been made by David Owens. I will argue that the functional account gives us an understanding of the slogan 'truth is the aim belief' that can help us to avoid objections like Owens'. He seems to assume an interpretation of the slogan according to which the aim of truth is something like an end that agents set for themselves. On my view, to say that truth is the aim of belief is not necessarily to say that truth must be the end that an agent has in coming to form a belief; rather, it is to say that the function of the kind 'belief' is to accurately represent reality. Beliefs can have this function independently of the aims and interests of believers, and as we will see, this point will come in handy in defending the view that I favor. In addition, I will argue that any non-evidential considerations that agents bring to bear on their theoretical reasoning – considerations such as how much time the agent has in which to form a belief, how important it is to her to have an answer to the question she is asking, etc. – are considerations that determine the rational structure of *inquiry*, rather than the rationality or irrationality of forming a particular belief at the end of inquiry. Non-evidential considerations help to determine how long it would be reasonable to engage in inquiry, what form the investigation ought to take, and how much evidence one should acquire before forming a belief. Once these features of the inquiry have been fixed, however, the only question which needs to be answered at the end of the inquiry, before one actually forms a belief, is whether or not one's evidence meets the contextual standard for a

rational belief. The actual process of belief-formation is governed only by evidential considerations.

In Chapter 3 I defend the view that the function of action is the satisfaction of the agent's desires. Before defending this view, I argue against two others. First, I argue against Velleman's account, according to which the aim of action is to achieve 'autonomy' or 'self-knowledge' (he treats these terms as interchangeable). I claim that the simple version of this account which he originally proposed in (Velleman (2006)) fails to place any substantive constraints on action: just about all actions will count as providing the agent with self-knowledge or autonomy, and so just about all actions will count as achieving the success-condition on action. I then go on to claim that the more complex version of the account which he develops in later papers faces a dilemma: either it fails to provide a substantive account of reasons, because anything that an agent *believes* to be a reason for action will thereby count as such a reason; or it sneaks a substantive account of reasons into the definition of the constitutive aim of action, thus failing to derive such an account from the constitutive aim. The second view I argue against is the 'guise of the good' view, which I claim will be unworkable given my assumptions about evaluative terms and properties: there will not be an available account of goodness which will apply to all things that agents might pursue, and if we say that the function of action is to cause or bring about a good *action*, then the view ends up being completely vacuous. Finally, I give a brief defense of my desire-based account of the aim of action. I argue that the view does not commit us to a Humean theory of motivation, and that it builds on a plausible explanation of when and why people have reasons for action, an explanation which has an easier time than other views of saying how it is that

we can recognize other people as having reasons we fail to have, or as failing to have reasons that we do have. (It should be noted that, although I do not commit myself to a Humean theory of motivation, I do give an account of how we are motivated by our reasons which I think is attractive in its own right, and which is properly 'continuous' with my account of what reasons for action consist in.)[6]

In Chapter 4 I turn to providing my account of theoretical and practical rationality. On my view, reasoning consists in the making of inferences which are meant to lead on to a certain kind of true belief. In the case of theoretical reasoning, we want to form a true belief about which proposition – of those we are considering in our context of inquiry – is the most likely to satisfy the aim of belief, that is, which one is most likely to be true. Likewise, in the case of practical reasoning, we want to form a true belief about which action – of those we are considering in our context of inquiry – is the most likely to satisfy the aim of action, that is, which one is most likely to satisfy our desires. I briefly develop a process-reliabilist view of proper belief-forming methods, drawing on the work of Alvin Goldman. This account is used to explain when it is that we have a reason to adopt the conclusion we have reached at the end of our inquiry, and hence to explain when we have a reason to form a belief or adopt a course of action. If we have reached our conclusion through reliable methods, then the belief we adopt or the action we perform on the basis of this conclusion will be reasonable. To complete the unified theory, I argue that, although my account of the aim of action is desire-based, desires themselves do not generate reasons. Rather, one's properly formed beliefs about one's desires are what provide reasons for action: a desire can fail to provide a reason for action if the agent does not have a properly formed belief that she has that desire; and an agent

[6] Thanks to Greg Scherkoske for raising this challenge.

can have a properly formed belief that she has that desire, and hence have a reason for action, even if she does not in fact have that desire. Thus, both theoretical and practical reasoning involve inferences between propositions, or beliefs. They are both cognitive enterprises, directed at forming beliefs about what instance will best satisfy the function of its kind.

I close Chapter 4 by arguing against Michael Smith's view of reasons, which I call the Ideal Observer Theory, or IOT. This theory comes in two versions. On the Example Model, I have a reason to Φ iff my fully rational counterpart in a counterfactual world Φs in my circumstances. On the Advice Model, I have a reason to Φ iff my fully rational counterpart in a counterfactual world would want me to Φ in my circumstances. On both models, my counterpart's being 'fully rational' consists in his being idealized in a certain way: he has all relevant information on my situation, he lacks any of the false beliefs that I might have which affect my decision, and he deliberates correctly. I argue that both versions of the IOT lead to false results: the IOT tells us that I have a reason to Φ in situations where intuitively I lack such a reason, and it tells us that I lack a reason to Φ in situations where intuitively I have such a reason. A large part of my argument consists in drawing analogies between theoretical and practical reasoning. I try to show that the IOT has no plausibility as an account of theoretical reasons, and motivate the claim that the same considerations which lead to this conclusion apply in the case of practical reasons as well.

Finally, in Chapter 5 – my concluding chapter – I sum up my thesis, and sketch some areas where further development is needed. I try to show what the issues are, and how I now think they might be profitably pursued.

CHAPTER 2: THE CONSTITUTIVE AIM OF BELIEF

In this chapter, I begin laying the groundwork for a unified account of theoretical and practical rationality, one based on the functional account of goodness-properties outlined in the Introduction. First, in Section 2.1, I argue that the function of the kind 'belief' is accurate representation of reality: I thus give a functional interpretation to the familiar claim that truth is the aim of belief. Section 2.3 deals with objections that David Owens has recently raised against the idea that truth is the aim of belief. Specifically: in Section 2.3.2 I respond to his claim that the truth-aim account of belief cannot explain the authority of epistemic norms, why we ought to care about them; while in Section 2.3.3 I respond to his claim that the truth-aim account of belief cannot explain the motivational force epistemic norms, how it is possible for us to be moved by them to the extent that we are. In both cases, in addition to presenting other difficulties for Owens' arguments, I will argue that the function account allows us to avoid problems that we would have a more difficult time dealing with on a more traditional construal of the truth-aim account of belief which attributes to agents an overriding interest in the goal 'obtaining true beliefs'.

2.1. TRUTH IS THE AIM OF BELIEF

On the function account of goodness-properties, beliefs can only be good or bad *qua* beliefs if the kind 'belief' has a function. If the notion of truth is going to provide the foundation for an account of the evaluative properties that beliefs possess – including the properties of being rational or irrational – then the function of belief must be *to be true*, that is, to accurately represent the part of reality with which it is concerned. What reason do we have, then, for thinking that accurate representation is the function of belief?

I think the most natural answer is that belief has the function of accurately representing the world: this idea is captured in the slogan 'Truth is the aim of belief'. If a belief is true, then it has succeeded in performing its function *qua* belief; if it is false, then it has failed to fulfill its function. Do we have any reason to think that this answer is the correct one? I think we do: namely, that any mental state Φ, directed at a proposition *p*, which we attribute to a person and which does not seem to be sensitive to truth or falsity of *p* thereby fails to count as a belief. Consider, for example, someone who maintains a mental state Φ*p* in the face of overwhelming evidence that *p* is false, but not by developing *ad hoc* hypotheses which would explain away the evidence; rather consider someone who actually treats evidence against *p* as irrelevant to the question of whether they ought to maintain the state Φ. They don't see evidence against the truth of *p* as providing any reason for them to re-evaluate their mental economy. The attitude this person has makes sense in the case of mental states like desiring: one can easily and without fault maintain the desire that a proposition be true in the face of evidence that it is false; and so this person can quite easily insist that such evidence is irrelevant to the question of whether they ought to retain their desire. The functional approach to the quality of mental states can easily explain why we can, without fault, retain our desires that certain propositions be true in the face of evidence that they are false: desires do not have as their function the accurate representation of reality. The point of a desire that *p* is not to track the truth or falsity of *p*, so evidence for the truth or falsity of *p* on its own does not speak either for or against holding the desire. By contrast, evidence for the truth or falsity of *p is* relevant to the question of whether one ought to believe that *p*. We can explain this by supposing that the function of belief is to accurately represent the world:

because the whole point of a belief is to represent the way the world is – rather than, say, the way the subject would like it to be – evidence that it fails to do so is evidence that one should abandon the belief.

We can consider a stronger example. Suppose someone maintains their mental state Φ towards proposition *p*, not simply in the face of evidence against *p*, but together with the explicit belief that *p* is false. This person does not see *p*'s falsity as a reason to re-evaluate their mental state. Again, this attitude makes sense in the case of desires: if charged with irrationality, the person can say 'Look, I'm not committed to *p* actually being the case; I just *want* it to be the case, and the fact that it isn't the case is no reason, in and of itself, to stop wanting it to be!' Moreover, this person would be completely right in saying this. The same attitude, however, doesn't make sense when applied to belief. Our hypothetical individual can't justify holding certain attitudes by saying 'Look, I'm not committed to *p* actually being the case; I just *believe* it to be the case, and the fact that it isn't the case is no reason, in and of itself, to stop believing it to be!' This defense doesn't make any sense, because believing that *p* actually *does* commit one to *p*'s being true. Believing just is believing-to-be-true, and that is because the function of belief is to accurately represent the world.

Remember, on the view being developed here, talk of 'the aim of belief' should be interpreted in functional terms: to say that belief aims at truth is to say that belief has as its function the accurate representation of reality. This will be important for the rest of this chapter – as well as for my treatment of the aim of action in Chapter 3 – because it implies that the aim of belief is, to borrow a phrase from J. David Velleman,

'subagential'.[7] What this means is that the aim of belief is not (necessarily) a consciously-held end of believers. For one's belief that *p* to have truth as its aim, it is not necessary that one actually have the goal of truth in mind when formulating it, or that one fail to have any other goals in mind. As we will see, this distinctive feature of my version of the truth-aim account will be of help in answering Owens' objections.

2.2. OWENS' OBJECTIONS TO TRUTH AS THE AIM OF BELIEF

Before moving on to show how we might develop an account of theoretical rationality on the basis of the semantic and metaphysical account of goodness that I have sketched – a project dealt with in more detail in Chapter 4 – I will consider some objections that have recently been raised against the idea that truth is the aim of belief. In particular, I will be dealing with arguments presented by David Owens, who claims not only that belief does not aim at truth, but that "believing is not purposive in any interesting sense."[8]

2.2.1. THE EXPLANATORY BURDEN OF THE TRUTH-AIM ACCOUNT

On Owens' view, the idea that belief aims at truth, if it is to be accepted, must solve four philosophical problems. These problems are:

(1) To explain why truth acts as the standard of correctness for beliefs;

(2) To explain why beliefs are governed by those epistemic norms which in fact govern them, rather than any other possible norms;

(3) To explain the rational authority of epistemic norms, why we ought to care about whether our beliefs satisfy the norms or not; and

[7] Velleman (1996), p.717.
[8] Owens (2003), p.283.

(4) To explain the motivational force of epistemic norms, how it is that we can be moved by them to adopt or revise beliefs just to the extent that we are.

It seems plausible that the truth-aim account of belief – and in particular, the functional interpretation of this account that I am advocating – can solve at least the first two problems. First, if the function of belief is accurate representation, if that is what beliefs are *for*, then we seem to have a simple explanation of why truth acts as the standard of correctness for beliefs: true beliefs fulfill the function of beliefs, while false beliefs do not fulfill this function. Second, if epistemic norms are a kind of instrumental norm – that is, they tell us what sorts of considerations make it probable that adopting a belief Φ will best enable us to achieve the aim of belief – then we can give an account of why certain norms apply and others don't: to take a simple example, the norm 'Believe that *p* when possession of that belief would increase your overall happiness' does not apply, because the fact that a belief would make you happy is no indication that it is true.

Owens accepts, for the sake of argument, that the truth-aim account can give answers to (1) and (2), although he expresses some reservations about the solution to (2). In particular, he thinks that the idea that belief aims at truth will not explain certain very specific norms, such as:

(i) A belief that *p* is reasonable only if there is some evidence for *p*;

(ii) A belief that *p* is reasonable only if there is more than just a little bit of evidence for *p*; and

(iii) A belief that *p* is reasonable only if there is more evidence for *p* than against *p*.[9]

[9] Ibid.

In addition, he suggests that the truth-aim account will have difficulty explaining why it is rational, in some cases, to believe a contradiction.[10] I will leave these issues to one side for now, and focus on his objections to the truth-aim theorist's claim to be able to solve problems (3) and (4). I will return to the issue of these specific epistemic norms and the case of contradictory beliefs in Section 2.4.

2.2.2. THE TRUTH-AIM ACCOUNT AND THE AUTHORITY OF EPISTEMIC NORMS

Beginning with (3), why does Owens think that the truth-aim theorist can't adequately explain the authority of epistemic norms? That is, why can't the truth-aim theorist explain why we ought to care about epistemic norms? Owens' argument comes in two parts. First is a 'parity of reasoning' argument: he claims that guessing, just as much as believing, aims at truth; and yet, the norms which govern when we ought to make a guess are not all explained just in terms of the aim of truth. In particular, our other interests, such as an interest in maximizing utility, also play a role in determining when we ought to make a guess. Therefore, we need some extra reason to think that the norms governing belief are explained *just* in terms of the aim of truth. Second, he argues that no such reason is available; there is no plausible account of how the norms governing truth could be determined solely by the aim of truth.

Taking the parity argument first, Owens offers the following abstract formulation of the truth-aim theory, which does not specify that the formulation only applies to beliefs:

[10] Ibid., p.288.

> Φ-ing that p aims at the truth if and only if someone who Φs that p does so with the purpose of Φ-ing that p only if p is true.[11]

He points out that this formulation doesn't just apply to beliefs; it also applies to the act of guessing. Guessing aims at the truth in the sense that, in making a guess, one is doing so with the aim of getting the right answer, that is, with the aim of guessing *p* only if *p* is true. Someone could always use the word 'I guess that *p*' without having this intention, of course, but that would not be a genuine act of guessing. (For instance, one might be simply impatient with the command 'Guess who I just ran into!' and just want to get the conversation moving, without any interest in stating the right answer.)

Not only does guessing aim at truth in the same way that belief does, Owens thinks, but this fact seems to explain why guessing is governed by evidential norms: the more evidence there is for *p*, the better my guess is, even if it turns out to be wrong; the less evidence there is for *p*, the worse my guess is, even if it turns out to be right. The reason considerations of evidence are important for evaluating guesses is that evidence is an indicator of truth, and in guessing one is trying to guess truly.

As with belief, Owens asks if the truth-aim account can explain, not just the fact that evidential norms apply to guessing, but the content of specific epistemic norms which govern just how much evidence is required for a guess to be rational in any given case. His answer is: no. When we consider the specific evidential norms which govern the rationality of guessing, Owens thinks, we must appeal to something which gets left out by the truth-aim account, namely all of the agent's *other* interests aside from an interest in guessing correctly:

---

[11] Ibid., p.289.

> "[How much evidence is required in a particular case] will be a function of two factors: first the aim of making only correct guesses which any guesser must have and second the subject's other desires and purposes. Not getting it wrong may be the only purpose constitutive of guessing but if that were the guesser's only purpose, he could achieve it simply by making no guess at all. We need to integrate error-avoidance with other aims which lead the guesser to guess and thus determine when the guesser has sufficient evidence to hazard a guess."[12]

Given the formulation of the truth-aim account provided, according to which the purpose a guesser has is to guess that *p* only if *p* is true, one could always satisfy the truth-aim by simply not hazarding a guess: if one doesn't make a guess, then *a fortiori* one doesn't make a false guess. This strategy would be most evident in cases where the guesser doesn't believe that they know the answer: their evidence for *p* does not, as far as they can tell, guarantee that *p* is true. However, it seems obvious that one can rationally make a guess on the basis of inconclusive evidence, and that in some cases it would be positively irrational to withhold from guessing. Owens presents some examples of just this sort of case. Consider, for instance, a case where I will win a large sum of money if I guess correctly. To make it simple, we can suppose that the question is a straightforward yes-or-no question: the answer is either 'yes' or 'no'. Even if I have no evidence either way, it is rational for me to make a guess, because doing so is the only way to win the prize money. If I don't make a guess, I thereby guarantee that I won't win the money, whereas by making a guess my chances thereby are improved (in this case, they have gone from 0 to 0.5).

What Owens thinks examples like this show is that "we treat the rationality of guessing just as we would the rationality of any other action: guessing that p is reasonable when aiming at the truth by means of a guess that p would maximize expected

[12] Ibid. p.291.

utility."[13] That is, although guessing aims at the truth, the *rationality* of a guess depends on more than just whether or not it is likely that the guess will achieve that aim. Even if the chances of a correct guess are low, guessing can still be made rational by the wider purposes that the agent has in making a guess, such as improving one's odds of winning the prize money.

When we guess, non-evidential considerations play a role in determining whether or not our guess is rational. Specifically, these non-evidential considerations determine how much evidence is required for a guess to be rational in one's current situation: if the possible benefits are fairly low, and one might actually lose something by guessing incorrectly, and one has plenty of time to think before guessing, then the amount of evidence required for rationality will be relatively high; if the possible benefits are high, and one will not lose anything by guessing incorrectly, and one has little time to think before guessing, then the amount of evidence required for rationality will be relatively low. What this shows is that the three evidential norms which Owens singles out as applying to belief:

(i) A belief that *p* is reasonable only if there is some evidence for *p*;

(ii) A belief that *p* is reasonable only if there is more than just a little bit of evidence for *p*; and

(iii) A belief that *p* is reasonable only if there is more evidence for *p* than against *p*.

simply don't apply to guessing. A variation in the possible gains can make it rational to make a guess on the basis of very little evidence, or even no evidence at all. Moreover, non-evidential considerations can make it rational to hazard a guess when one has equal

[13] Ibid., p.292.

evidence for each of two possible answers; withholding a guess in such a case can be irrational, whereas the act of withholding belief might be the rational thing to do.[14]

So, even though guessing satisfies the formulation of the truth-aim account, that aim does not, on its own, explain the norms which govern the rationality of particular guesses. The other interests that an agent has can contribute to the rationality or irrationality of a guess. Thus, by parity of reasoning, we should not expect the truth-aim account to explain the norms which govern the rationality of particular beliefs, unless we are given some reason to think that the truth-aim behaves differently for beliefs than it does for guesses. That is Owens' first argument.

The question now becomes: do we have any reason to think that the truth-aim account of belief can, on its own, explain the norms of theoretical rationality? Owens considers three ways in which we might try to reach this conclusion:

1. Believers only *care about truth.*

One might think, Owens says, that believers *only* have the aim of truth in mind when they come to form their beliefs; their other goals do not interact with the truth-aim in the way that they do in the case of guessing. That is, the other goals that I have – such as maximizing my expected utility – do not determine how much evidence is required for a belief to be rational in a given case. Owens thinks that this idea is clearly false. He writes,

> "Were avoiding error one's only objective in forming any particular belief, one could very easily achieve that objective simply by forming no beliefs at all. Reasonable people form beliefs and they do so on the basis of much less than conclusive evidence. This practice makes sense, on the purposive model

[14] Ibid. Owens only commits himself to the idea that the first two evidential norms for belief don't apply to guessing, withholding judgment as to whether the third norm applies to guessing or not.

> of belief, only if believers have some objective in forming beliefs other than the mere avoidance of false beliefs."[15]

According to Owens, my aim in forming a belief that *p* is not simply to believe *p* only if *p* is true: if it were, then I could satisfy that aim by simply remaining agnostic unless and until I possessed conclusive evidence. But, on a plausible reading of the passage just quoted (see note 10), Owens thinks, quite rightly in my view, that it can be perfectly rational to form a judgment on the basis of inconclusive evidence, and one's interests other than truth can play a significant role in determining when one is rational in acting this way. So it can't be true that believers only care about truth when forming their beliefs, because their other interests play a role in determining when one is rational in forming a belief (the details of this role for non-evidential considerations will be considered below).

2. Believers care more about truth than guessers do.

If we accept Owens' claim that our interests in things other than truth play a role in determining the rationality of belief, can we explain the difference in evidential norms between belief and guessing by claiming that, in the act of forming a belief or gathering evidence with the purpose of doing so, believers care more about truth, assign it more importance in their judgments about how to proceed, than guessers do? Owens thinks it is unlikely that this tactic will work, since the difference in norms still applies even where the believer does not really care about the issue at hand. For instance, suppose I am asked

[15] Ibid., pp.293-294. Note that here Owens may be departing from the view he develops in *Reason without Freedom*. There, he argues that one can believe oneself to have a rational belief only if one believes oneself to have conclusive evidence. If the claim above, that reasonable people form beliefs on the basis of *in*conclusive evidence, is read as saying that they form such beliefs while thinking of themselves as only having inconclusive evidence, then this marks a shift away from his earlier view. Owens' argument for the claim that one can only believe that one has a reasonable belief if one thinks that one has conclusive evidence is discussed in Section 2.3.3.

to guess how many planets there are in the nearby solar system. I don't much care how many planets there are, and in all likelihood I would forget the answer rather quickly if it were told to me. However, there is a large sum of money at stake; I am in a position to make a great profit by guessing correctly. In such a case, it seems I have good reason to hazard a guess, even if I have no evidence to go on in selecting one from a range of possible answers I could give. However, if we change the example so that I am being asked to form a *belief* rather than to make a guess, suddenly the money seems irrelevant. I cannot rationally get myself to form a belief simply by considering the fact that doing so would be of some benefit to me. But in this case, I don't care about the truth of the matter any more than I do in the guessing case, and yet the norms are still different. The difference of norms cannot be explained by a difference in how much importance guessers and believers give to truth.[16]

3. Distinguishing epistemic and non-epistemic interests.

Finally, Owens considers the option of distinguishing between a subject's epistemic and non-epistemic interests, and insisting that the former kind can play a role in determining the rationality of belief which the latter kind cannot. Believers still have interests other than truth, but these interests are not relevant to the formation of beliefs. Owens replies that, even if we could pull off this sort of division – a prospect of which he is skeptical – it would leave us with a puzzle about the authority and motivational force of epistemic norms. If our epistemic interests are isolated in this way, where they cannot be subjected to balancing and trade-offs with one's other interests when one is forming beliefs, then what reason do we have to guide our beliefs by them, and how are epistemic norms able to move us in the way they do? "Why should we be held to norms which take no account

[16] Ibid., p.295.

of many of the things which matter most to us? And how, in any case, can we be moved by such norms?"[17]

It seems to me that the right way to go here is to adopt a modified version of the first option that Owens considers, that believers *only* have truth as their aim. Owens is quite right that, if we give this option the agential interpretation, so that we end up claiming that believers only have one goal when they are investigating a certain question, and that goal is simply to form only true beliefs on the matter, then the idea has little to no plausibility. It is not true that believers only have the truth as their goal. They have other interests which guide their investigation, most obviously those interests that determine what questions they choose to investigate. I care about certain questions, but not about others, and among those questions to which I would like to know the answer, there are some that I care about more than others. Moreover, how the investigation will proceed depends on my cognitive resources – Should I get a degree in mathematics and do some original research on the question of how to treat infinity in a mathematical context, or should I simply read some secondary materials and defer to the experts? If I have *x* amount of time available for research, how much time should I devote to investigating each of the questions that interest me? Crucial information for answering such questions is information about how much I care about the question at hand, and how I can best incorporate my investigation into my everyday life. (After all, not all of my time can be devoted to research. Presumably I have some other goals I would like to achieve in my lifetime.) I need to be concerned with more than simply truth.

However, on the functional view, the aim of belief is subagential: belief does not aim at truth in the sense that the believer represents truth to themselves as one of their

[17] Ibid., p.295.

goals, which they can then proceed to balance among all of their other goals. Rather, accurate representation is the *function* of belief, and this function is not simply one goal among an agent's many goals. What the truth-aim theorist says in not that believers only have truth as their goal when they are deciding what to believe: rather, the truth-aim theorist says that belief, as a matter of its nature, has accurate representation as its function, and that this function is what determines the rationality of forming a belief.

In order for this response to be effective, however, I need to say something about Owens' distinction between the rationality of inquiry and the rationality of forming a given belief once the process of inquiry has ended. He writes:

> "What is in question…is the rationality of belief, *not* the rationality of the activities of evidence-gathering, or of evidence-storage and retrieval, or of thinking about the value of evidence before you. I shall call these activities *inquiry*. Inquiry often precedes belief formation and is motivated by all sorts of non-evidential considerations: only if I am interested in whether OJ is guilty will I go to the trouble of reading the newspapers, thinking about the evidence reported, trying to memorise [sic] it and so forth. This paper focuses on the role non-evidential considerations play *after* we have ensured (perhaps by means of inquiry) that a certain quantity of evidence on the matter is before us. Suppose the evidence favours [sic] OJ's guilt. Still the question remains: does it favour [sic] OJ's guilt enough to make it reasonable for me to make up my mind on the matter and form the belief that OJ is guilty? I maintain that this question cannot be answered without invoking non-evidential considerations (the importance of the issue etc.). And this question is not equivalent to the question: should I go on gathering evidence or thinking about OJ's guilt?"[18]

The function account could potentially resolve the main objection we are dealing with, that the truth-aim theory cannot explain the content and authority of epistemic norms, if we could show that non-evidential considerations *only* govern the process of inquiry, and not the process of belief-formation. If we need to appeal to non-evidential facts to justify our beliefs once the process of inquiry is ended, then it is not true that the rationality of

[18] Ibid., p.288.

belief is governed solely by concerns with truth: beliefs could also be made rational by how much we care about the issue, how much time we have to invest in an investigation, etc. However, if non-evidential considerations only govern the norms of inquiry, then we could maintain that, once the inquiry is finished, all that must be appealed to in order to justify a belief is one's evidence, or the reliability of one's methods, or the like.

Owens phrases the problem of determining the role of non-evidential facts as the problem of determining whether the question 'Do I have enough evidence?' is equivalent to the question 'Should I continue with my inquiry?' Owens thinks that there are some clear examples where an answer to one question is not an answer to the other: "There are all sorts of reasons why I might stop inquiring into OJ's guilt without forming a belief (I got bored) or continue my inquiry despite having formed a belief (I need to convince someone else)."[19] However, in order to evaluate the issue at hand, we should restrict ourselves to cases where I am inquiring into the question whether *p* is true because I want to form a belief on the matter, and where the inquiry has come to an end, not because I am bored and have lost interest, but because it is time to make a decision whether to form a belief on the matter, and if so, what I should believe. I think that by focusing on such cases we can be clear about what role non-evidential facts play in determining the rationality of forming a belief given a certain amount of evidence: we will not be distracted by cases where the inquiry is designed to convince someone else, rather than to help me form a belief, or cases where the inquiry is brought to a premature close because I have simply lost interest. If we focus on these cases, however, the question 'Should I continue my inquiry?' does not arise: we are assuming that the inquiry has actually been brought to a close. If we want to determine whether non-evidential considerations

[19] Ibid.

determine the rationality of belief-formation or just the rationality of inquiry, we need a different way of asking the question.

I have been assuming that Owens is correct that the amount of evidence required for a belief to be rational will vary from context to context, and that the contextual features which account for this variance are largely, if not solely, non-evidential features, such as my talent for certain kinds of inquiry, how much importance I have placed on the answer to my question, etc. Once we adopt this position, however, it seems to me that the question of whether I have enough evidence to justify a belief is in a way ambiguous: it could be a question about how much evidence is actually required in this context for a belief to be rational; or, it could be a question about whether my evidence satisfies the criterion – which is assumed to be already given – for having enough evidence for a belief to be rational in this context. The latter question asks whether one's evidence meets the relevant standard, while the former asks what the relevant standard *is*. Notice that the question about what the relevant standard is will need to be answered by appeal to non-evidential facts, since those facts determine the standard, while the question about whether I have enough evidence to meet the standard can be answered *merely* by appeal to evidential facts.

With this distinction in place, there is a strategy available to the truth-aim theorist: namely, to argue that the standard of evidence in a given context, which is determined by non-evidential considerations, co-varies with the norms governing one's inquiry. That is, facts about how much one cares about the issue, how much time one can devote to it, etc., determine what kinds of inquiry would be rational – how long it should take, what resources one should look into – and it is these facts about the rationality of inquiry

which determine the relevant standard of evidence. Whether one meets the standard of evidence, however, is not a question about the rationality of one's inquiry.

Consider how Owens' OJ example appears when we adopt this strategy. It is true that the amount of evidence required to justify belief in OJ's guilt varies from person to person based on non-evidential facts, but these facts also determine what sort of inquiry it is rational for one to pursue. If I don't much care about celebrity news, then it is rational for me to believe the conclusion that the court comes to; the courts may not be infallible, but I don't have enough interest in the issue to warrant further investigation. That might change, however. Suppose years down the road I have become a legal scholar, and I am writing a book that would require an in-depth treatment of the Simpson trial. In such a case, the kind of inquiry rational required of me will be quite different: it would be epistemically irresponsible to simply rely on the court's decision. Rather, I ought to do further reading, examining *why* the court reached the conclusion that they did, and examining the evidence for and against OJ's guilt. The non-evidential facts which determine how much evidence is required *also* determine the kind of inquiry required. The question 'How much evidence is required for belief to be rational?' is closely connected to the question of what kind of inquiry it would be rational to pursue: how much one cares, what intellectual talents one has, etc., determine how one ought to go about answering the question, and how much evidence one should have before forming a belief. Such non-evidential considerations, however, do not need to be appealed to when the time comes to determine whether one ought to form a belief: one need only appeal at that point to the evidence one has available.

If the arguments of the last few pages work, then we can answer Owens' objection to the truth-aim theory. We can adopt a variant on the position that believers only have truth as their aim when forming beliefs: it is not that believers only care about truth; rather, belief has accurate representation as its function, and this function operates at the sub-agential level. Moreover, the other concerns that Owens appeals to are concerns which govern *inquiry*, not belief-formation. The amount of evidence required for one's belief to be rational is determined by the kind of inquiry that one ought to pursue; whether one actually meets the standard, by contrast, is not a fact about the kind of inquiry one ought to pursue. It is a question that can be answered simply by pointing out how much (or how little) evidence one has. It is these purely evidential facts which govern the rationality of belief-formation at the end of inquiry; non-evidential facts govern the rationality of inquiry itself. This distinction also allows us to account for the fact that rational believers will not simply choose to remain agnostic in order to satisfy Owens' formulation of the truth-aim hypothesis. The interests which govern inquiry allow us to adopt beliefs on the basis of inconclusive evidence, and so it will be rational to risk a false belief rather than remain agnostic.

2.2.3. THE TRUTH-AIM ACCOUNT AND THE MOTIVATIONAL FORCE OF EPISTEMIC NORMS

Moving on, why does Owens think that the truth-aim theory can't adequately answer problem (4)? That is, why can't the truth-aim theory explain the motivational force of epistemic norms, the fact that people tend to be motivated to adopt or revise their beliefs in accordance with their evidence?

To begin with, we should note that Owens interprets the claim that the truth-aim account can explain the motivational force of epistemic norms in a very particular way. He writes, "It is possible to violate, deliberately and self-consciously, many of the norms which govern our mental life. But philosophers who hold that belief aims at the truth have inferred from this that a believer can't self-consciously disregard indications of the truth (i.e., evidence) in forming a belief."[20] In support of this claim, Owens quotes Bernard Williams, who writes, "If in full consciousness I could will to acquire a belief irrespective of its truth, it is unclear that before the event I could seriously think of it as a belief, i.e., as something purporting to represent the truth."[21] That is, if one adopts a mental state Φ towards the proposition *p* without any regard for whether *p* is true, then it is difficult to think of Φ as a genuine belief; it is difficult to interpret such a person as *believing* that *p*, rather than hoping that *p*, desiring that *p*, etc. Presumably, this interpretive constraint is explained by the fact that belief has accurate representation as its function: since beliefs by their very nature are directed at truth, any mental state Φ(*p*) which is insensitive to considerations which are relevant to *p*'s truth (in the sense described in Section 2.2) cannot be a belief.

So far, so good. But Owens think that this kind of explanation cannot be given if 'truth is the aim of belief' only means that belief has truth as its standard of correctness. Assertion, he says, aims at truth in just this sense, and yet it would be superstitious to imagine that this fact somehow deprives people of the ability to self-consciously make groundless assertions.[22]

---

[20] Ibid., p.296.
[21] Williams (1973), p.148.
[22] Owens (2003), p.296.

In order to account for the motivational force of epistemic norms, the slogan 'belief aims at truth' must, he thinks, mean that "belief's standard of correctness [namely, truth] is embodied in the very purpose of the belief-former."[23] To clarify this idea, he considers the activities of guessing and of shooting at a goal, as in a game of hockey. In such cases, he says, the standard of correctness is embodied in the purpose of the agent, because disregard for the satisfaction of that standard is inconsistent with performance of the activity. Here, Owens seems to be thinking in terms of agential, rather than sub-agential, aims: if I am not aiming to get the puck into the goal, in the sense of consciously attempting to get the puck into the goal, then I am not really shooting at the goal; if I am not aiming to get the right answer, in the sense of consciously attempting to get the right answer, then I am not really guessing.[24]

Moreover, Owens says, we cannot show that belief aims at truth simply by showing that we cannot get ourselves to form beliefs on the basis of purely non-evidential considerations, such as the fact that believing *p* would make us feel a lot better than we presently do. We would only be justified in thinking that belief aims at the truth in the sense described above if "we can also get ourselves to form beliefs, as we can get ourselves to make guesses, by reflecting on what the best way to achieve the aim of belief would be."[25] Owens seems to mean that, if truth is the aim of belief, then we ought to be able to get ourselves to believe *p* by considering the evidence we have for the truth of *p*. The guesser makes a guess when they decide that – taking into any other purposes she

---

[23] Ibid., p.297.
[24] A more comical example of this phenomenon can be found in Douglas Adams' book *Life, the Universe, and Everything*, in which the character Arthur Dent learns the trick behind being able to fly. All you have to do, he discovers, is throw yourself at the ground and miss.
[25] Owens (2003), p.298.

might have which determine the amount of evidence required for a guess to be rational – they are sufficiently likely to guess correctly.

By contrast, Owens thinks, believers don't adopt a belief that *p* when it seems sufficiently likely that they will thereby believe what is true. He goes on to explain *why* believers don't do this, and here it is worth quoting him at length:

> "That is not because a believer has only one purpose: to avoid error. As already noted, this objective could be achieved easily enough by believing nothing at all, by declining to shoot at the target. Believers with limited cognitive resources wish to form a view on large a [sic] number of matters and so the formation of a particular belief will be reasonable only if it is part of a general policy of belief formation which is itself reasonable. Such policy [sic] must be sensitive to various non-evidential considerations. As James points out, believers must strike a balance between the agonies of agnosticism and the risks of error; evidence alone cannot strike this balance for them…"[26]

The rationality of a belief is governed, not simply by the amount of evidence there is either for or against a proposition, but by non-evidential considerations, such as the amount of time one has available to spend on investigating the matter, whether or not there are practical benefits to forming any belief, whether true or false, with regard to the question at hand, etc. Owens thinks that if belief really aimed at the truth – and here he seems to adopting the agential, rather than the sub-agential, view of what aiming at the truth involves – then the kind of sensitivity to non-evidential considerations that is required by rational belief-formers "would involve integrating the goal of having [a] belief only if it is true with the subject's wider purposes in forming beliefs."[27] Now, Owens thinks that we *do* ordinarily integrate the goal of having only true beliefs with the other goals we have in forming beliefs, and thus that we *are* sensitive to non-evidential considerations when we form beliefs, but that

[26] Ibid., p.298.
[27] Ibid.

> "this sensitivity to non-evidential reasons for belief[28] is not the sort of thing which is registered efficaciously in the subject's deliberations about what he ought to believe. Our subject can't get himself to believe by reflecting that he has sufficient evidence to form a belief, given his wider purposes in forming beliefs. When we bystanders consider whether it is reasonable for our subject to form a belief on this matter, we will attend to the importance of the issue, the limited cognitive resources which he has to devote to it and so forth. But reflection on these matters will not move the believer himself to belief, as it moves him to guess…"[29]

According to Owens, belief-formation ought to be, and for the most part is, sensitive to non-evidential considerations. However, although these facts are relevant to the rationality of forming a belief, reflection on them does not motivate one to believe, "[o]r, at least, [one's] rationality alone does not guarantee that such reflections will have any influence on [one's] beliefs."[30] Belief formation is not under our control in the way that guessing is, according to Owens. While we can determine the best means to our ends and guide our behavior so that it accords with those means – as in guessing, or stamp collecting – we cannot get ourselves to believe a proposition by considering that doing so is the best means to achieving truth.

The first thing to notice about this line of argument, and something which I have highlighted in my exposition, is that Owens is assuming the agential view of what it means for belief to aim at truth. Hopefully, I have done enough so far in this chapter to undermine our confidence that the truth-aim account ought to be given the agential

---

[28] It should be noted that the idea that non-evidential considerations are genuinely reasons for belief is crucial to Owens' rejection of the thesis he calls *Reflective Motivation*: "if *R* is a *prima facie* reason to believe that *p*, reflection on *R* provides the rational subject with a motive to believe that *p*." (Owens (2000), p.20.) Owens argues that since non-evidential considerations are reasons for belief, and yet it is not rational for subjects to be motivated by consideration of them to form beliefs, *Reflective Motivation* is false.

[29] Owens (2003), pp.298-299.

[30] Ibid., p.300.

interpretation. Does the argument succeed if we interpret it so as to apply against the functional view I have offered?

It seems that Owens' line of thought can be broken down into two separate arguments:

Argument 1

(1) If belief has an aim, then an agent will come to believe *p* if it seems to the agent sufficiently likely that by doing so they will achieve the aim of belief.

Therefore,

(2) If truth is the aim of belief, then an agent will come to believe *p* if it seems to the agent sufficiently likely that by doing so they will believe what is true.

(3) It is not true that an agent will come to believe *p* if it seems to the agent sufficiently likely that by doing so they will believe what is true.

Therefore,

(4) Truth is not the aim of belief.

The first problem here is that Premise (1) – and hence Premise (2), which is a particular instance of (1) – is false as stated. It is not true that if belief has an aim, then an agent will form a belief that *p* if doing so seems to them sufficiently likely to achieve that aim. At most, what follows from the hypothesis that belief has an aim is that an agent is being *irrational* if they fail to form a belief that seems to them sufficiently likely to achieve the aim of belief.

This first difficulty extends into a second, consequent difficulty, namely that it threatens to make Premise (3) a *non sequitur*. Of course, Premise (3) is ambiguous

between an existential and a universal interpretation. Assume, first, the existential interpretation, so that (3) reads: *some* agents will not come to believe *p* if it seems sufficiently likely to them that by doing so they will believe what is true. On this interpretation the premise is pretty clearly a *non sequitur*: the truth-aim account implies, at most, that it is rational for agents to form beliefs when it seems sufficiently to them that by doing so they will believe what is true; the fact that some agents do not behave this way does nothing to undermine this normative claim.

Now assume the universal interpretation, so that (3) reads: *no* agents will come to believe *p* if it seems sufficiently likely to them that by doing so they will believe what is true. The issue is more complex on this interpretation, since we might very easily question the idea that belief aims at truth if the belief that '*p* is sufficiently likely to be true to warrant belief' *never* motivated belief in *p*. However, it is not at all clear that Owens has presented us with an argument for this premise. He has pointed out that facts about one's evidence *on their own* cannot motivate belief, since they do not determine how much time one has available to consider the question, how important it is to have an answer, etc. These points can all be accepted by the truth-aim theorist, I have argued, once the crucial distinction between inquiry and belief has been acknowledged. As I said in the previous section, we are interested in what we have reason to believe, if anything, once the inquiry has been closed. The very idea, which is used in Owens' argument, of a belief being 'sufficiently likely to be true' is context-sensitive: what counts as a sufficient level of evidence to warrant belief depends on non-evidential considerations of the kind which govern the process of inquiry. So Premise (3) needs to be read, if it is to be relevant, as 'No agents will come to believe *p* if it seems to them that, given the

constraints which govern sufficiency of evidence in the agent's context, it is sufficiently likely that by doing so they will believe what is true', and that is just plain false. It certainly seems that we *can* come to believe something by reflecting on the fact that we have sufficient evidence for it, considering the non-evidential facts which govern the required levels of evidence. If I genuinely believe that the amount of evidence I have is good enough to warrant belief in *p*, then I might very well come to believe *p* on that basis.

Owens' first argument falls afoul of considerations regarding non-evidential facts which play a role in deliberation and belief-formation. His second argument, however, explicitly addresses such facts:

Argument 2

(1) A rational belief-forming policy must be sensitive to non-evidential considerations. That is, the norms governing the required amount of evidence for a rational belief must be sensitive to contextual features, such as the agent's cognitive limitations, what they are interested in knowing, etc.

(2) If belief has an aim, then this sensitivity to non-evidential considerations on the part of rational agents must take the form of explicitly using non-evidential considerations in one's reasoning about what we ought to believe.

(3) Rational agents do not explicitly use such considerations in their reasoning about what they ought to believe.

Therefore,

(4) Belief does not have an aim.

Therefore,

(5) Belief does not aim at truth.

It should be noted that Premise (3) seems to be ambiguous between two different readings:

(3a) Rational agents do not use non-evidential considerations in isolation from evidential considerations in their reasoning about what they ought to believe. For instance, they do not form beliefs *solely* because having those beliefs will make them happier.

(3b) Rational agents do not form beliefs on the basis of norms which explicitly make reference to non-evidential considerations.

Now, (3a) certainly seems true. If I am told that I will receive a one million dollar prize if I form the belief that *p*, it is not irrational of me to fail to be motivated to believe *p* on that basis. The fact that having a certain belief will benefit me does not make it rational to hold that belief, since it does not make that belief more likely to be true. However, there is no reason to think that this fact is in any way damaging to the truth-aim theorist, since there is no reason to think that the truth-aim theory is committed to the claim that we *can* be rationally motivated to form a belief *purely* on the basis of non-evidential considerations.

Evaluating (3b) is more difficult. In the previous section, I argued that non-evidential considerations only govern inquiry and the amount of evidence required within one's context for a belief to be rational, and that they cannot be appealed to in justifying one's belief after the inquiry has been closed. That is, non-evidential considerations determine the truth of statements of the form '*x* amount of evidence is required in this context for belief to be rational', but only evidential considerations can be appealed to in

determining whether one meets the standard. In this sense, it seems that rational agents *do* use norms which make reference to inconclusive evidence, since certain kinds of inquiry will allow that one does not need to have conclusive evidence in order for one's belief to be rational. It seems that I can reason to myself that, given the limited time I have to investigate a particular question which interests me, I have gathered enough evidence to make it reasonable to believe *p* rather than either believing not-*p* or withholding judgment altogether, even though the evidence is inconclusive. For example, suppose I want to know whether the theory of general relativity is true. I could, of course, dedicate the next decade or so to getting an education in theoretical physics, and dedicate several decades after that to my own research in the area. However, I might not have the talent for mathematics which would be required for such a plan of inquiry to be reasonable, and I might not care enough about the answer to the question to dedicate that much time to it. Instead, I could simply defer to the scientific community. Such deference would require some research, of course – reading some books on the subject directed at non-specialists would be a good start – but not nearly the amount that would be required by the first plan of inquiry. I might reason to myself as follows: 'Given my limited cognitive resources, and given the fact that I'm not so passionate about general relativity that it would make sense to dedicate the rest of my life to investigating it, I should simply defer to what the experts say, as I do with many other questions. Once I have done *x* amount of research into figuring out what the general consensus within the scientific community is, I have done enough research to warrant forming a belief on the matter and moving on to other questions which interest me'. This looks like a perfectly rational way of forming beliefs, contrary to what (3b) tells us.

Why might Owens think that (3b) is true? The answer to that question can be found in his recent book *Reason without Freedom*. There he argues that, in order to think to myself that I have a rational belief that *p*, I must think that I have *conclusive* grounds for believing *p*. Leaving aside questions of whether Owens can hold this view consistently with what he says in "Does Belief Have an Aim?", why should we accept the claim that thinking I have a rational belief requires thinking that I have conclusive grounds? Owens argues for the claim as follows:

(6) To think that I have a rational belief that *p* is to think that I know that *p*.

(7) To think that *x* knows that *p* is to think that *x* has a conclusive ground for *p*.

Therefore,

(8) To think that I have a rational belief that *p* is to think that I have a conclusive ground for *p*.[31]

Premise (7) is defended on the basis of considerations arising from Gettier cases. On Owens' view, the problem with the traditional justified-true-belief account of knowledge is that it conceives the justification in question as possibly being inconclusive: it is possible to have a justified true belief that *p* where one's evidence is consistent with not-*p*. Following Dretske, Owens proposes to solve the problem by saying that knowledge requires, not just the kind of justification which is consistent with one's belief being false, but *conclusive* justification, the kind which arises from what Owens calls a 'knowledge-constituting' state, such as seeing that *p* or learning that *p*. "To perceive or learn of *p*", Owens writes, "is to be justified in believing that *p* in a way which really does suffice for knowledge of *p*: it is to have a conclusive ground for *p*. Such a ground would be simply

[31] Owens (2000), p.46.

unavailable unless *p* were true: we can hardly perceive or learn of something which isn't so."[32] If my justification for believing *p* is conclusive in this way, then it is not susceptible to the peculiar feature of Gettier cases, wherein one's justification falls short of making it conclusive that *p*, but one's belief turns out to be true because of some kind of intervening bit of luck.[33]

Owens' analysis might be right, and so Premise (7) might be true. The point I want to make is that Premise (6) is given no clear defense, and in fact seems to be false. To see why, consider the third-person equivalent of (6):

(6*) To think that *x* has a rational belief in *p* is to think that *x* knows that *p*.

(6*) is clearly false, since if it were true, we could not consistently think of ourselves as being engaged in a rational disagreement with anyone. If thinking that my interlocutor's belief is rational requires thinking that he has knowledge, then it also requires thinking that his belief is true, since knowledge implies truth. But if I commit myself to the truth

---

[32] Ibid., p.43.

[33] On the luck-based diagnosis of Gettier cases, see Zagzebski (1994). She writes: "[There is] a general rule for the generation of Gettier cases. It really does not matter how the *particular* element of knowledge in addition to true belief [the 'justification' or 'warrant' element in the JTB analysis] is analyzed. As long as there is a small degree of independence between this other element and the truth, we can construct Gettier cases by using the following procedure: start with a case of justified (or warranted) false belief. Make the element of justification (warrant) strong enough for knowledge, but make the belief false. The falsity of the belief will not be due to any systematically describable element in the situation, for if it were, such a feature could be used in the analysis of the components of knowledge other than true belief, and then truth would be entailed by the other components of knowledge, contrary to the hypothesis. The falsity of the belief is therefore due to some element of luck. Now emend the case by adding another element of luck, only this time an element which makes the belief true after all. The second element must be independent of the element of warrant so that the degree of warrant is unchanged. The situation might be described as one element of luck counteracting another." (pp. 209-210.)

For an argument that the feature described by Zagzebski as 'one element of luck counteracting another' can actually be compatible with having knowledge, see Sosa (2007) pp.80-86.

of his belief, then I cannot *disagree* with him without being inconsistent. The question is, why should be accept (6) if we reject (6*)? After all, if I were to say 'It is literally unthinkable that any of my rational beliefs should turn out to be false', it seems that I ought to be dissuaded of this view by considering that it is *not* unthinkable of anyone else that one of their rational beliefs is false. How could it be that while it is perfectly consistent to think that someone else has a rational false belief, it is somehow impossible for me to think that *I* might have a rational false belief? What is so special about me, that my rational beliefs are all, simply in virtue of being rational, also true? The answer is: nothing. There is nothing special about me which would make this the case, and that is precisely why it is perfectly consistent for me to think that I have a rational belief which is false, and therefore it is consistent for me think that I have a rational belief which is not knowledge, which it wouldn't be if (6) was true.

It might be objected that (6) is true while (6*) is false for reasons similar to those which make it the case that I cannot say

(9) The cat is on the mat, but I don't believe it;

but I can say

(10) The cat is on the mat, but *x* doesn't believe it.

If I claim that the cat is on the mat, then I commit myself to believing that the cat is on the mat, and likewise if I claim to believe that the cat is on the mat I commit myself to its being the case that the cat is on the mat. (9) is not an inconsistent proposition, since it could be true, but it cannot be consistently thought or uttered, since such a thought or utterance would involve an inconsistent set of commitments. (10) does not have this feature, since claiming that the cat is on the mat does not commit me to the claim that *x*

believes the cat is on the mat, and likewise claiming that *x* believes the cat is on the mat does not commit me to the claim that the cat is on the mat. To attribute to someone else a belief that *p* is not to commit oneself to *p*, but simply to describe that person's mental state, whereas attributing such a belief to oneself *does* commit one to *p*; self-ascription of beliefs is not simply a kind of description of one's mental state, but I kind of commitment to the content of those beliefs.

Extending this point to (6) and (6*), it might be argued that, even though it is logically possible for me to have a rational belief which is not knowledge, I cannot consistently believe or claim that this possibility is actualized. The trouble here is getting clear on how such an argument would work. Suppose we present the argument like this:

(11) If I attribute to myself a belief that *p*, then I commit myself to the truth of *p*.

(12) If I commit myself to the truth of *p*, then if I commit myself to the claim that I have grounds for believing *p*, I commit myself to the claim that I have *conclusive* grounds for believing *p*.

(13) If I commit myself to the claim that I have conclusive grounds for believing *p*, then I commit myself to the claim that I know that *p*.

Therefore,

(14) If I attribute to myself a belief that *p*, then if I commit myself to the claim that I have grounds for believing *p*, I commit myself to the claim that I know that *p*.

Note that (14) is simply a more logically complex version of (6). Crucially, the argument is supposed to rely on the falsehood of the third-person equivalents of (11) and (12):

(11*) If I attribute to *x* a belief that *p*, then I commit myself to the truth of *p*.

(12*) If I commit myself to the truth of *p*, then I commit myself to the claim that if *x* has grounds for believing *p*, I commit myself to the claim that *x* has *conclusive* grounds for believing *p*.

If (11*) and (12*) are false, then I cannot argue from them via

(13*) If I commit myself to the claim that *x* has conclusive grounds for believing *p*, then I commit myself to the claim that *x* knows that *p*;

which is true, to

(14*) If I attribute to *x* a belief that *p*, then I commit myself to the claim that *x* knows that *p*.

The falsehood of (11*) is analogous to the falsehood of (10): attributing beliefs to another does not involve committing oneself to the truth of those beliefs. This is what prevents us from running the argument from (11*) to (14*). According to this argument, the reason I cannot think that my rational beliefs are false while I *can* think that others' rational beliefs are false is that the former kind of thought, but not the latter kind, commits me to the truth of the belief. If the inference from (11) to (14) works, then thinking that I have a rational belief will also commit me to thinking that I have knowledge.

But is it true that if I attribute to myself a belief that *p* then I commit myself to the truth of *p*, where this commitment is strong enough to allow (12) to be a valid inference? That is, does my commitment to *p* entail that if I think I have reason to believe *p* then I must think that those reasons are *conclusive*? I don't think that simply believing *p* commits us to anything this strong. The reason why can be made clearest, I think, by

using a possible-worlds account of belief, along the lines developed by, among others, David Lewis. Lewis writes:

> "the content of someone's system of belief about the world (encompassing both belief that qualifies as knowledge and belief that fails to qualify) is given by his class of *doxastically accessible* worlds. World W is one of those iff. he believes nothing, either explicitly or implicitly, to rule out he hypothesis that W is the world where he lives. Whatever is true at some…doxastically accessible world is…doxastically possible for him. It might be true, for all he knows or for all he believes. He does not know or believe it to be false. Whatever is true throughout the…doxastically accessible worlds is…doxastically necessary; which is to say that he…believes it, perhaps explicitly or perhaps only implicitly."[34]

The content of a belief Φ is given by the set of possible worlds which that belief excludes from a description of ways the world might be. If a belief state eliminates from consideration all the worlds in which *p* is false, then that belief commits one to truth of *p*.[35] Interestingly, the passage quoted above seems to *identify* belief that *p* with exclusion of all the not-*p* worlds. However, Lewis recognizes that this isn't quite right, and goes on to account for beliefs which *don't* exclude the possibilities inconsistent with those beliefs:

> "[W]e must…provide for partial belief. Being a doxastic alternative is not an all-or-nothing matter, rather it must admit of degree. The simplest picture, idealized to be sure, replaces the sharp-edged class of doxastic alternatives by a subjective probability distribution. Thus you may give 90 per cent of your credence to the hypothesis that you are one or another of the possible

[34] Lewis (1986), p.27.

[35] On Lewis' view of things, the set of worlds from which a believer can begin to exclude possibilities is fixed by the abilities of the subject to represent those worlds to him- or herself. He puts this idea to use in his rejection of Frank Jackson's hypothesis of phenomenal information: "When someone doesn't know what it's like to have an experience, where are the alternative open possibilities? I cannot present to myself in thought a range of alternative possibilities about what it might be like to taste Vegemite. That is because I cannot imagine either what it *is* like to taste Vegemite, or any alternative way that it *might* be like but in fact isn't…I can't even pose the question that phenomenal information is supposed to answer: is it this way or that? It seems that the alternative possibilities must be unthinkable beforehand; and afterwards too, except for the one that turns out to be actualized." (Lewis (1988), p.281) For an interesting discussion of this strategy for dealing with Jackson's argument see Stalnaker (2008), pp.33-35.

> individuals in *this* class, but reserve the remaining 10 per cent for the hypothesis that you are one of the members of *that* class instead."[36]

That is, when I consider the various possibilities – which include not only the ways the world might be, but where I might be located in it – I may not be able to completely eliminate certain possibilities. However, I may give differing levels of credence to different possibilities. This is a familiar enough phenomenon: a high level of credence in a proposition can properly be described as a belief, although the subject has not eliminated all of the possibilities in which the proposition is false.

In a sense, if I place a high degree of credence in a proposition, I commit myself to its truth. I can say to myself: 'I can't rule out the possibility that *p* is false, but I believe that it is true'. However, if my belief is of this sort, then (12) is completely implausible: in holding such a belief, I need not commit myself to the claim that my evidence for this proposition is *conclusive*. I can allow that I have formed the belief on the basis of inconclusive evidence, and yet retain the belief all the same, without thereby being irrational. At best, (12) will apply to only two varieties of belief: (a) beliefs to the effect that, given everything else that one believes, all the doxastically possible worlds are worlds in which *p* is true (for instance, where one sees that *p* follows logically from the other elements in one's belief set); and (b) beliefs to the effect that, regardless of what else one believes, all the doxastically possible worlds are worlds in which *p* is true (for instance, where one believes *p* to be necessarily true). Owens might choose to restrict his account of rationality to only beliefs of these kinds, but there is little reason to think that by doing so he can provide an adequate account of the rationality of belief. Beliefs which do not have maximal degrees of credence are a common phenomenon, and are often used

[36] Lewis (1986), p.30.

to guide our actions: we might choose to do Φ, for instance, because we believe that doing so will bring about a certain result, even though we don't believe this in anything like the strong sense described by (a) and (b) above. The rationality of such an action will be determined, in part, by how rational it is to have that belief *to that degree*. Moreover, the notion of degrees of belief plays an invaluable role when we try to determine the role that logic plays in the rationality of our beliefs.[37]

If it is true that we ought not to restrict ourselves to beliefs which actually *rule out* the possibility of their falsehood, then (12) is false. (12) might be true for some beliefs, but it is certainly not a general principle. Without (12), however, we cannot derive the claim that by committing myself to the truth of *p* I commit myself to the claim that I *know* that *p*. Therefore, the argument that we can hold (6) without (6*) fails.

To sum up this section: Owens presents two arguments against the claim that the truth-aim theory can explain the motivational force of epistemic norms. The first argument faced two difficulties: (a) that it assumed that the truth-aim theorist is committed to claiming that all agents, at all times, will come to believe *p* when it seems sufficiently likely to them that by doing so they will believe the truth, when in fact, the truth-aim theory is only committed to the claim that if agents *don't* behave this way, they are for that reason behaving irrationally; and (b) that its third premise, if re-phrased so as to be relevant to the position that the truth-aim theorist is actually committed to, is false: sometimes we *are* able to get ourselves to be believe certain propositions because they seem to us sufficiently likely to be true. The second argument, on its most plausible reading, depends on the claim that agents can never rationally form a belief by reasoning according to principles that make reference to non-evidential considerations. I have

[37] See, for instance, Field (2009).

suggested that the only defense of this claim open to Owens is the idea that agents can never rationally form a belief by reasoning to themselves that they have enough inconclusive evidence to warrant adopting a belief, even if such a belief is not guaranteed to be true. This idea has, I hope, been shown to be false: although it might be true (I won't take a stand on this issue here) that one can't rationally hold a belief with a maximal degree of credence while explicitly thinking that one's evidence for that belief is inconclusive, the claim will not extend to beliefs with less than maximal degrees of credence. Given how frequent such beliefs are, the principle is false in, at the very least, a large number of cases.

CHAPTER 3: THE CONSTITUTIVE AIM OF ACTION

Having argued that truth is the aim of belief, in the sense that belief has as its constitutive function the accurate representation of reality, I now turn to develop the analogous view in the case of practical reason. The idea that truth is the aim of belief is meant to ground an account of theoretical reason, because theoretical reasoning is reasoning about what to believe, and the constitutive function of belief provides an account of what makes a belief good *qua* belief. In order to carry out the project in the practical sphere, we need an account of what makes an action good *qua* action, because practical reasoning is reasoning about what to do. In Section 3.1 I examine J. David Velleman's view that autonomy or self-knowledge is the constitutive aim of action: the good act *qua* act is the one which allows us to have a certain kind of knowledge about ourselves. I argue that this view cannot be made to work: a simple version places too few constraints on successful action, while a more complex version ends up assuming an account of reasons for action, rather than serving to ground such an account. In Section 3.2 I examine the 'guise of the good' view, and argue that it is unworkable on the assumptions about goodness that I am working with: the idea that the function of action is to bring about the good cannot ground an account of successful action. In Section 3.3 I present my alternative view: the function of action is to satisfy the agent's desires. I briefly defend the claim that this account is both plausible and capable of satisfying the goals of an account of the constitutive aim of action: namely, the four aims presented by Owens for an account of the constitutive aim of belief, transposed into the practical sphere. However, the fourth aim, to explain how it is that we can be motivated by practical norms, will need to be discussed more fully in the next chapter where I present my view of how desires figure

into our reasons for action. Finally, in Section 3.4 I deal with some initial objections to my view of the aim of action.

3.1. J. DAVID VELLEMAN: AUTONOMY AS THE CONSTITUTIVE AIM OF ACTION

3.1.1. VELLEMAN'S MOTIVATION AND STRATEGY

The motivation for Velleman's view comes primarily from a dilemma which he sees as arising from two plausible suppositions about what it is to have a reason to act in a certain way:

> (1) "[R]easons for someone to do something must be considerations that would sway him toward doing it if he entertained them rationally"[38]
>
> (2) "[T]he only considerations capable of swaying someone toward an action are those which represent it as a way of attaining something he wants, or would want once apprised of its attainability."[39]

(1) leaves open the possibility that someone can remained unmoved by considerations which are, in fact, reasons for him to act. That is, it leaves open the possibility that some fact, *R*, genuinely counts as a reason for an agent *A* to do act Φ, but *A* remains unresponsive to *R*. All that is required for *R* to be reason to Φ is that *A* be moved by it upon considering it *rationally*. (2), however, places a constraint on what sorts of facts can move an agent. It says that an agent can only be moved to act on a fact *R* if *R* represents Φ as way of obtaining something *A* desires, has as a goal, etc. If that is true, then *a fortiori* the same restriction applies to the kinds of facts that can motivate an agent who considers them rationally.

[38] Velleman (1996), p.694.
[39] Ibid.

There are two traditional ways of accommodating (1) and (2) – the internalist way and the externalist way – and each one yields a different account of what reasons agents have. On the internalist view, represented by Bernard Williams,[40] the reference to 'rational entertainment' in (1) gets interpreted with reference to the subject's desires or goals. Specifically, to entertain some fact *R* rationally means something like: to entertain it with full understanding of what one's goals, desires, etc. actually are, and how *R* relates to them. Williams formulates this idea in his postscript to "Internal and External Reasons" as follows: "*A* has a reason to Φ only if there is a *sound deliberative route* from *A*'s subjective motivational set…to *A*'s Φ-ing."[41] So, although it is not necessary that, if *R* is a reason for *A* at time *t*, then *R* must be capable of swaying *A* to act at *t*, it is necessary that, if *R* is a reason for *A* at time *t*, then *R* must be capable of swaying *A* to act at *t* if at *t A* understands the contents of *A*'s subjective motivational set and can reason correctly from these contents to the proposition that Φ-ing would satisfy some element in that set. The externalist view does not interpret 'rational entertainment' in this way. On the externalist view, the failure to have certain elements in one's subjective motivational set can, in and of itself, constitute a failure of practical rationality, one which disables one from rationally entertaining some fact *R* which constitutes a reason for action.

These two different ways of accommodating (1) and (2) thus lead to two different views of what counts as a reason for an agent to act. On the internalist view, being practically rational just consists in understanding what is in one's subjective motivational set, being able to see how certain facts about the world relate to the attainment of those goals and desires, and being moved to act by such considerations. As long as one satisfies

[40] Williams (1979).
[41] Williams (2001), p.91, emphasis in the original.

these criteria, any fact which fails to move one to action thereby fails to count as a reason. By contrast, the externalist view requires, not just that one be aware of what is in one's subjective motivational set, but also that certain things be members of the set (and also, most likely, that certain things *not* be members of the set). If one does not satisfy these further criteria, then a fact which fails to move one to action might still constitute a reason, since it might move one to action if one were practically rational, that is, if one had certain dispositions to respond to facts of these kinds which are required for full practical rationality.

The problem that Velleman has with the externalist view of reasons is that it brings with it an explanatory burden that internalism is able to avoid, and which it is very difficult to meet. The internalist model does not make any pronouncements on what counts as a reason in anything more than an agent-relative way. That is, it defines reasons wholly in terms of the subjective motivational sets of agents, and since these sets vary from agent to agent, internalism is not committed to there being any facts which will count as reasons for action for all agents, or for all agents in a particular context. [42] Externalism *is* committed to this, however, and thus imposes upon itself a burden to explain why a particular fact counts as a reason for all agents, independently of what they desire or what they have as their goals.

The attempt to take up this burden leads us, Velleman thinks, to the question of what the aim of practical reasoning is supposed to be. If we can specify what we are

[42] This is not to deny that there may be general inference rules which apply to all agents. For example, the rule 'If Ψ is a member of my subjective motivational set, and Φ-ing will allow me to achieve Ψ, then I have reason to Φ'. The point is simply that particular facts, like 'By going to the theater down the street I can see the new Lars von Trier film' will not count as reasons for action for everyone, since not everyone will want to see the new Lars von Trier film. For a stronger argument, that we can treat the generalized means-ends principle as a categorical imperative, see Drier (2001).

inquiring after when we ask what we ought to do, then that specification might entail that there are certain features of an action Φ which automatically make it the case that all agents have a reason to Φ in certain contexts. Such a specification is difficult to come by, however. The goal of practical reasoning cannot simply be said to be knowledge of 'what one ought to do', since that gives us no guidance as to what the right answer to the question is, nor about how we are to engage in the practice of finding the answer. Such a specification, as Velleman says, is like being told that the object of a competitive game is to win, or that the object of a hunt is the quarry. Both of these statements are true, but entirely uninformative. In order to have a clear idea of how one ought to go about achieving the goal of an activity or enterprise, one needs a substantive idea of what that goal is.[43] Velleman is skeptical of any attempt to derive this substantive idea from an analysis of concepts like 'reason' and 'rationality', but it isn't immediately clear how else it is to be derived.

This brings us to Velleman's dilemma: he thinks that if we adopt externalism, we incur a heavy burden of justification. We need to provide an account of what one has reason to do which will explain why certain features of actions constitute reasons for doing them whether or not a particular agent cares about, or is responsive to, those features. We can alleviate this burden by adopting internalism, but then we end up relativizing reasons to particular agents: there won't be any features of actions which constitute reasons doing them for all agents. Of course, not everyone will see this consequence of internalism as problematic. (It will become clear, when I present my own view of the constitutive aim of action, that I do not see this relativity as a problem.) But Velleman is concerned to avoid the consequence. He wants to provide an account of

[43] Velleman (1996), p.700.

reasons that is agent-neutral, in the sense that there will be some features of particular actions which constitute reasons to do those actions for all agents, at least in some contexts.

Velleman's response to this dilemma is to find a middle way between internalism and externalism. His strategy is to find some goal which all agents have of necessity: a goal that all agents possess simply in virtue of being agents. If there is such a goal, he thinks, then we can have the benefits of both internalism and externalism without either of their drawbacks. That is, we can alleviate ourselves of the need to provide an account of facts which count as reasons independently of agents' goals, but without relativizing reasons to particular agents, since reasons will be explained in terms of a goal that *all* agents have. We can thus have an agent-neutral account of reasons along with a plausible explanation of how the facts referenced in the theory can count as reasons for agents to act.

We need to be careful how we formulate Velleman's strategy. A brisk reading of "The Possibility of Practical Reason" can give the impression that he thinks a failure to treat a certain fact *R* as a reason for action shows, not only that one is acting irrationality, but that one is not, in fact, an agent.[44] But it is unlikely that there is any such feature. To see this, consider the case of theoretical reasoning. A single failure to respond to evidential considerations will not rule *A* out as a believer: it simply makes *A*'s belief

[44] See, in particular, p.706, a section of which I quote below. When he refers to 'an inclination to believe what is true on a matter', Velleman can be read as referring to a very particular inclination: the inclination, *here*, *now*, to believe what seems true on *this* matter. However, he seems to actually have in mind a more general inclination: the inclination to believe what seems true on *any* matter, at *any* place, at *any* time. It is thus a very general disposition, more akin to water-solubility than to 'the inclination to dissolve *here*, *now*, in *this* particular sample of water', which an object might lack while still being water-soluble.

irrational. Being a believer does not require one to fully and adequately respond to relevant evidence at all times. At most, it requires that one do so sufficiently often that it makes the most sense of *A*'s behavior to interpret him as capable of holding beliefs. In the case of particular beliefs, a single failure to respond appropriately to the evidence is consistent with taking *A*'s mental state to be a genuine belief. However, if the mental state persists for long enough without being responsive to evidence against *p* in the way in which it is appropriate for beliefs to do – where this responsiveness can include the formulation of *ad hoc* hypotheses, since such a response signals that *A* is at least treating the evidence *as* evidence – then we have reason to interpret the mental state as something other than a belief, such as a hope or a desire.

What matters for Velleman's strategy is not particular failures to respond to reasons for action, but the lack of any disposition to respond to them. Velleman makes this point clear when discussing the parallel position in the case of reasons for belief:

> "Perhaps we should ask whether the absence of [an *inclination* to treat certain considerations as reasons for belief] would undermine the existence of reasons for belief or would alternatively undermine the believer's claim to rationality. The answer to this question would determine whether reasons for belief were internal or external reasons. If someone weren't *inclined* to believe what seemed true, would sign of truth in a proposition no longer count as reason for him to believe it? Or would he no longer qualify as a rational believer?...Both, I think…if someone isn't *inclined* to believe what seems true on a topic, he is no longer subject to reasons for believing things about it; but is no longer subject to reasons for belief about it…because he is no longer a believer about it at all, and a fortiori no longer a rational believer."[45]

Likewise, what we are looking for in the practical case is some feature of particular actions[46] which has the following characteristic: if a purported agent *A* remains

[45] Ibid., p.706, emphasis added.
[46] Or rather, a certain kind of action, as I will explain over the next few pages in discussing Velleman's idea of 'full-blooded' action.

unresponsive to this feature – that is, they fail to treat it as a reason for action – in a sufficiently high number of cases, then we have reason to reinterpret *A*'s behavior so as not to count *A* as an agent.

The sort of strategy just outlined is, of course, very similar to the one I have adopted in this work, and my strategy is derived mainly from Velleman. As we saw in Chapter 2, my view of theoretical rationality is derived from what I take to be the constitutive aim of action (where I interpret the idea of a constitutive aim in functional terms): since belief has accurate representation as its function, the success condition for belief *qua* belief is truth. When reasoning about what we ought to believe, we are trying to figure out what is true. Velleman makes this argument as well.[47] He then goes on to apply the idea to practical reasoning: if we can determine the goal of theoretical reasoning by reference to the constitutive aim of belief, then we ought to be able to determine the goal of practical reasoning by reference to the constitutive aim of action. Once we know what the constitutive aim of action is, we will know what features of actions allow them to satisfy this aim, and it is *these* features to which someone must be responsive in order to count as an agent. So what, on Velleman's view, is the constitutive aim of action? He has presented two related views on this matter, which I will deal with separately. Both of Velleman's accounts refer to the aim of action as 'autonomy', but his later view consists in supplementing the earlier account of what autonomy consists in. I will call the earlier view the 'simple' view and the later view the 'sophisticated' view (although this is not to suggest that the 'simple' view is without complexity).

[47] Velleman (1996), pp.707-714.

3.1.2. THE AUTONOMY VIEW 1: THE SIMPLE VIEW

According to Velleman, the constitutive aim of action cannot be simply one end among others that an agent possesses. Rather than being an end or goal within an agent's motivational set, the aim of action must be 'subagential': it must be something which action, as a kind, possesses independently of the particular aims that agents have. While I can't claim to understand his argument for this position,[48] I will adopt it for the same reason that I adopted the position that the aim of belief is subagential, namely, that the goodness-conditions for the particulars of a certain kind is the function of the kind. If the aim of action is going to determine the goodness-conditions for action, then this aim must be a function, and functions can exist independently of the goals or desires of agents. The function of action is what it is whether or not an agent represents it to herself as one of her ends or goals.

In addition to this restriction on the possible answers to the question 'What is the aim of action', Velleman proposes that we focus our attention on what he calls 'full-blooded actions'. This is because we are concerned with the nature of action, and this concern can become side-tracked if we pay too much attention to reflexes or unconscious behaviors: "The fundamental question in the philosophy of action is not how imperfect an exercise of agency can be while still qualifying as an action. The question is the nature of

[48] He writes: "An end is something conceived by an agent as a potential object of his actions. It is therefore something that one cannot have unless one already is an agent, in a position to act, and so it cannot be something that one must already have in order to occupy that position. If action is to be constituted by an aim, that aim must be, so to speak, subactional or subagential – something that a subject of mere behavior can have, and by having which he can become an agent, as his behavior becomes an action." (Ibid., p.717.)

The major difficulty I have with this passage is that it slides between treating the constitutive aim of action as something that *action*, as a kind, possesses, and treating it as something that certain kinds of being – namely, subjects of 'mere behavior', such as non-rational animals – possess.

agency itself, and agency, like any capacity, fully reveals its nature only when fully exercised."[49] So what, according to Velleman, is a full-blooded action? He proposes that what differentiates full-blooded actions from mere behaviors is that the former are under one's conscious control. Full-blooded actions are autonomous, in the sense that they are composed of two states – (a) being conscious of one's behavior and (b) controlling the behavior – where the former state causes the latter. That is, it is by being conscious of what we are doing that we exert control over our behavior.[50]

The idea of consciousness of behavior controlling that behavior can be difficult to unpack. Velleman attempts to clarify it by introducing the concept of 'directive knowledge', which is contrasted with 'receptive knowledge'. Receptive knowledge is gained in a particular way, by accepting a proposition "in response to its being true".[51] This usually occurs when one is moved to accept a proposition *p* by considering the evidence for and against *p*. Directive knowledge, by contrast, is gained by accepting a proposition "in such a way as to make it true."[52] That is, knowledge is directive when one's mental state ensures that the proposition is true, not simply in the ordinary sense in which knowledge is a factive state and hence '*x* knows that *p*' entails *p*, but rather in the sense that one's mental state is what causes the fact to which *p* must correspond if it is to be true to obtain. By adopting certain beliefs, we thereby cause the world to be as it must for those beliefs to be true, and the causal connection at play here allows these beliefs to count as knowledge.[53]

---

[49] Ibid., p.714.
[50] Ibid., p.720.
[51] Ibid., p.721.
[52] Ibid.
[53] The causal role of the belief in generating the truth of *p* allows that state to satisfy Sosa's Safety condition: in all nearby possible worlds, if one holds the belief that *p*, then

According to Velleman, our beliefs about our own behavior can play the kind of causal role required for directive knowledge. What exactly the content of these beliefs is supposed to be raises certain issues. One might think that it is one's actual current behavior which makes one's beliefs about it true: since one's current behavior can be described by a present-tense proposition like 'I am Φ-ing now', it might be thought that it is *this* proposition which is believed to be true, and hence caused to be true. However, the causal process running from the belief to the behavior presumably takes time, and hence we immediately run into difficulties over how the belief could have been true at the time is was first held. If I form the belief that 'I am Φ-ing now' at time *t*, then 'now' refers to *t*, and 'I am Φ-ing now' will only be true if I am Φ-ing at *t*. If the causal process does not conclude with my Φ-ing until *t*+1, then it will not be true at *t* that I am Φ-ing, and hence 'I am Φ-ing now' will not be true then. Presumably at *t*+1 I will be able to observe my behavior and form the new belief that 'I am Φ-ing now', but this will be a case of receptive knowledge, not directive knowledge. At *t*+1 my behavior is what causes my consciousness of it, not the other way around.

Fortunately, in later work Velleman makes it explicit that the beliefs which play the requisite causal role are future-tensed: they take the form 'I will Φ', perhaps with some particular time being specified.[54] If I form the belief 'I will Φ' at *t*, this belief begins a causal chain that terminates with me Φ-ing at a later time, call it *t*+1. Assuming determinism, it is true at *t* that I will Φ, so my belief counts as an instance of directive

---

p is true. For discussion of the Safety condition and how it differs from the Sensitivity condition prominent in the work of Dretske, Nozick and DeRose, see Sosa (1999). For a recent argument to the effect that Safety is not a necessary condition for knowledge, see Sosa (2007), pp.80-86.

[54] Velleman (2004), p.277.

knowledge. If one goes on to act so as to make beliefs of the form 'I will Φ' true, and those actions are caused by those beliefs, one's behavior is under one's control.[55]

Now, Velleman claims that one cannot exert this kind of control over one's actions without aiming to – one can't accidently act self-consciously – and that this fact gives us reason to think that self-knowledge, or autonomy, is the constitutive aim of action.[56] Presumably, he intends this not to mean that one has conscious control as an end, but that action as a kind has as its aim the production of a state of being in control of one's behavior. In my functional terms, this would mean: the function of action is to put one in such a state that one's beliefs about what one will do come out true. If one's action is in accordance with what one believed one would do, and that belief has played the appropriate role in producing the action, then the action counts as successful or good.

3.1.3. OBJECTIONS TO THE SIMPLE VIEW

Objection 1: Autonomy and Agent-Neutral Reasons

Recall Velleman's issue with internalism: on the internalist views of reasons, a feature *F* of an act Φ counts as a reason for an agent *A* to Φ only if there is some element of *A*'s subjective motivational set which will be satisfied by an action with the feature *F*. Since the elements of the subjective motivational set – as the internalist traditionally construes it – will vary from agent to agent, there is no guarantee that there will be actions which all agents have reason to do in all contexts. It might turn out that there are no such actions. Whether or not one finds this result troubling, Velleman's goal is to avoid it: his goal is find some aim that action has which could be used to guide practical reasoning,

[55] Velleman (1996), pp.721-722.
[56] Ibid., p.719.

and which would show that all agents necessarily have a reason to behave certain ways in certain contexts.

The simple view of autonomy or self-knowledge will not ground an agent-neutral account of reasons. On the simple view, autonomy consists in doing what you believe you will do, and doing it because you have that belief: the belief is a kind of self-fulfilling prediction.[57] However, just as there are not necessarily any elements which must be present in all agents' subjective motivational set, so there are not necessarily any beliefs of the form 'I will Φ' that all agents share (with the possible exception of beliefs about unconscious or reflexive behaviors, such as 'I will continue to breathe while I sleep'). In a particular circumstance where I am faced with a choice between two potential courses of action, Φ and Ψ, what does Velleman's account tell me I ought to do? It tells me to do what will satisfy the constitutive aim of action, which is to act so as to make my beliefs about what I will do come out true. So, if I believe that I will Φ, then I have reason to Φ. If you are placed in the same circumstance with the belief that you will Ψ, then you have reason to Ψ. The same point holds in all circumstances: what an agent has reason to do depends on what they believe they will do, and these beliefs are not necessarily going to be the same for all people. Just as with the traditional internalist view of reasons, there is no guarantee that all agents will have reason to do the same thing in any situation.

Objection 2: (Almost) Nothing Fails to Meet the Aim of Autonomy

Recall that Velleman has restricted his attention to 'full-blooded' actions: he is concerned with finding a constitutive aim only for actions that count as full-blooded; reflexive and unconscious actions are to be ignored. Velleman defines full-blooded actions as those which are under an agent's conscious control, and he defines actions which are under an

[57] Thanks to Greg Scherkoske for this helpful phrase.

agent's conscious control as actions done in accordance with an agent's belief about what they will do, and done because the agent has that belief. With these two definitions in place, we can generate the first premise in my argument:

(1) Full-blooded action is Φ-ing through one's belief that one will Φ.

Velleman uses the word 'autonomy' to designate the constitutive aim of action, and says that an action is autonomous if it is done in accordance with one's belief about what one will do, and done because one holds that belief. From this definition we get the second premise:

(2) Autonomous action is Φ-ing through one's belief that one will Φ.

By straightforward reasoning, it follows that

(3) Full-blooded action is autonomous action.

If the constitutive aim of action is autonomy, then every full-blooded action has the function of contributing to one's autonomy, or self-knowledge. But if (3) is true, then every full-blooded action, simply in virtue of being full-blooded, counts as autonomous action. That is, every full-blooded action satisfies the constitutive aim of action. If the constitutive aim is meant to establish the success-conditions of action, from which we can derive practical norms, then this account is useless. Every action will count as successful, and so no full-blooded action can be ruled out by any practical norms we could derive from Velleman's account. The reason we can use truth to formulate norms of theoretical reasoning is that not all beliefs are true: a belief can fail to satisfy the success-condition, so the condition can be used to mark a division between successful and unsuccessful beliefs. If all beliefs were true by definition, then the aim of truth could not be used to mark the required division. Likewise with action: if all actions are autonomous by

definition, then the aim of autonomy cannot mark a division between successful and unsuccessful actions.

Kieran Setiya has recently made this point quite forcefully, arguing that the demands imposed by Velleman's account are far too easy to meet:

> "One falls short of these standards only when there is a mis-match between the content of one's intention, including its explanatory content, and what one actually does. So, for instance, I may not be adjusting the brightness of my television screen, even though I intend to be, because I am wrong about which button does what. Or I may be mistaken about my motives: I mean to be working hard because I enjoy it, but what really drives me is the fear of failure. So long as I do not go wrong in either of these ways, however, my action makes the explanatory content of my intention true, and thus provides me with knowledge of myself. If this is enough to satisfy the standards of practical reason, they are satisfied whenever we act in the way that we intend to act."[58]

Obviously, these restrictions – one must actually achieve what one intends, and be correct about what motivates one's action – impose virtually no limitations on what will count as a successful act, and may in fact be even less restrictive than a traditional internalist position. In any case, I take it as evident that not all actions which satisfy one's beliefs about what one will do, and which are the result of that belief, thereby account as good actions. One can go wrong in one's actions in far more ways than those Setiya points out, and the first kind of failure – not actually succeeding in doing what one thinks one is doing – does not automatically render one's action irrational.

So far we have found only two kinds of action which will fail to satisfy the aim of autonomy. I might fail to achieve what I intend to do, and I might be mistaken about what motivates me (aside from the extra motivation that Velleman thinks is provided by my belief about what I will do). But Velleman's more sophisticated account recognizes a third kind of failure, one which promises to impose more substantive constraints on what

[58] Setiya (2007), p.111.

counts as a successful action. I now turn to an examination of the more sophisticated view.

3.1.4. THE AUTONOMY VIEW 2: THE SOPHISTICATED VIEW

According to the simplified view, one's beliefs about what one will do simply describe the act in question, without making any reference to why it is being done. However, Velleman insists – drawing on his Kantian inspiration – that the reasons one is acting in the way that one is are crucial to our evaluation of the action as either rational or irrational. The pursuit of self-knowledge is not just the pursuit of knowing what one is doing, but knowing what one is doing *and* why one is doing it. As Velleman puts it, we aim to perform actions under descriptions which make sense to us. To give an idea of how this view works, he considers an example discussed by Jonathan Dancy: suppose Dancy goes shopping for a new pair of shoes, despite the fact that the shoes he has are perfectly fine. He has no need to buy a pair of shoes. Because of this, Velleman thinks, Dancy will be unable to make sense of his own action; why on earth is he buying a pair of shoes that he doesn't need? If he needed new shoes, that would go a long way to explaining his action; without that need there is no explanation. Velleman argues, on the basis of this example, that reasons for or against a certain action are grounded in facts about what actions would make sense:

> "Reasons against doing something, on my view, are considerations in light of which doing that thing would be unintelligible or hard to understand. For example, Dancy's recognition that he does not need new shoes makes it harder rather than easier to understand why he is shopping for shoes. The description 'buying shoes I don't need' therefore incorporates a reason against buying shoes rather than a reason for buying them."[59]

[59] Velleman (2004), p.278.

Of course, the idea of an action 'making sense' is closely related to the idea of an action being rational, so Velleman takes pains to ensure that we have an understanding of 'making sense' which does not presuppose an account of rationality, but rather which can help to develop such an account:

> "[T]here is a definition of 'making sense' under which it is a term of practical rationality. What makes sense for someone to do, by this definition, is whatever he has reason for doing. The statement that reasons for an action are the considerations in light of which the action would make sense can therefore sound like a tautology. But I do not mean to speak tautologically…When I speak of 'making sense', I am borrowing the phrase from the domain of theoretical reason, where it is used to characterize phenomena as susceptible to explanation and understanding. What makes sense to someone, theoretically speaking, is what he can explain. This is what I mean when I say that reasons for doing something are considerations in light of which it would make sense. I mean that they are considerations that would provide the subject with an explanatory grasp of the behavior for which they are reasons."[60]

The appeal to theoretical reasoning suggests that the kind of explanation at play here is causal. That would presumably make sense of the shoe-shopping case: if Dancy doesn't need new shoes, and he knows this, then he will probably be at a loss to find any causal explanation for what he is doing; he won't know what is motivating his action, and so he won't be able to provide a causal explanation of it. The point here is that the explanation must appeal to a motive that the agent has antecedently to the formation of the belief 'I will Φ' and to the performance of the action.[61] The belief 'I will Φ' is a sort of prediction.[62] When forming the belief, I examine the motives that I have and predict what I will do in the circumstances; if it turns out that I do what I thought I would do, and the motive which causes me to do it is what I predicted, then I have achieved self-knowledge.

---

[60] Velleman (2000), p.26.
[61] Velleman (2004), p.280.
[62] Dancy (2004), p.240.

An initial problem for this view is that we can easily modify the shoe-shopping case so as to allow Dancy to achieve self-knowledge through his compulsive action. In Velleman's version, Dancy does not know about his compulsion: all he knows is that he doesn't need the shoes he is buying. It is this lack of knowledge which makes the action inexplicable: Dancy cannot explain why he is buying the shoes. However, suppose that Dancy comes to learn that he has a compulsion to buy new shoes, even when he doesn't need them. That is, he is aware of an element in his motivational set which causes him to want to buy new shoes, and in fact causes him to act so as to satisfy this desire. In such a case, it seems that Dancy would be perfectly capable of predicting his shoe-buying behavior: he would actually expect himself to end up buying a new pair of shoes, despite the fact that he doesn't need them. He can thus describe his action to himself in such a way as to provide a perfectly acceptable causal explanation of his action, and hence achieve the aim of self-knowledge by buying the shoes. Presumably, however, Velleman does not want to count such a compulsive act as achieving the aim of action.

Velleman might have had these sorts of cases in mind when introducing his distinction between motives and reasons, a distinction which he insists does not map onto the motivating reason/justifying reason distinction. In making this insistence, he is not hostile to the distinction between facts which give a causal explanation of one's behavior on the one hand and facts which justify one's behavior on the other. Rather, he is hostile to the assimilation of the first kind of explanation into 'reasons' explanation: he doesn't want to classify motives as a kind of reason, and hence causal explanation as a kind of reasons explanation. The reason he resists this assimilation is that "[a]n agent can act on or out of motives without acting for any reason at all, because being motivated does not

in itself entail responding to anything as a justification."[63] For Velleman, being able to rationalize one's behavior is not simply a matter of being able to find a causal source for one's action; it is also necessary that the action resulted from a source that the agent sees as justifying his or her action.

Making sense of one's action is still a kind of causal explanation, but the kinds of explanation that will make sense are limited by what motives one has which one sees as justifying one's action. To achieve self-knowledge is to do Φ where one believes that one will Φ, this belief causes the action, and 'Φ' includes a description of the causal sources of the action in one's motivational set, where those sources are seen by the agent as justifying the action. When one forms the belief 'I will Φ', Φ-ing will allow one to achieve self-knowledge or autonomy *only if* the belief was formed on the basis solely of those elements in the agent's motivational set which the agent sees as justifying Φ-ing. If Φ-ing satisfies the description 'the action I predict I will do if I act only on elements in my motivational set which justify the actions they motivate', then Φ-ing will satisfy the constitutive aim of action. Returning to the shoe-shopping case, Dancy might succeed in buying a new pair of shoes, and he might know that this action is motivated by a strange compulsion he has to buy new shoes, but this is not enough for his action to be autonomous on the more sophisticated account, because that account imposes the extra requirement that Dancy sees the compulsion as making sense of his action.

3.1.5. OBJECTIONS TO THE SOPHISTICATED VIEW

Recall that the point of positing a constitutive aim of action is to explain why certain norms apply to action, and why certain considerations count as reasons. In order to fulfill

[63] Ibid., p.279.

its role, an account of the aim of action cannot *assume* a theory of reasons, but rather must be used to derive such a theory.

As we saw, Velleman places a constraint on self-knowledge, namely that one can make sense of one's action in light of the explanation of why one is doing it. Velleman needs to provide an account of 'making sense' or 'being intelligible' which does not reduce to the terms of practical rationality. That is, we can't simply say that Dancy's lack of need for the shoes is a reason against buying them, and that *this* is what makes his action unintelligible. That would be to import an account of reasons for action – in this case it seems like an instrumental sort of account – into the account of the aim of action, rather than deriving it from that account. If the aim of action is to achieve self-knowledge, then the definition of self-knowledge had better not make reference to reasons for action.

Velleman initially tries to avoid this problem by adopting a conception of 'making sense' from the sphere of theoretical rationality. As we saw, this conception appears to be a causal-explanatory conception: I can make sense of my action by knowing what is causing me to do it. However, Velleman ends up bringing a more practical conception of 'making sense' back into his theory by distinguishing between motives and reasons. On the sophisticated view, one only achieves self-knowledge if one sees the causes of one's behavior as justifying it; in this sense, compulsive actions, weakness of will, and action on the basis of desires that one sees as base, do not count as autonomous. The trouble is that now Velleman risks running into the exact problem that the theoretical account of making sense was meant to avoid: "What makes sense for someone to do, by this definition, is whatever he has reason for doing. The statement that reasons for an action

are the considerations in light of which the action would make sense can therefore sound like a tautology."[64] If 'making sense' of one's action requires appeal to actual reasons, then reasons for action are being assumed in the definition of the aim of action, rather than being derived from it: the aim of action amounts to 'knowing that one is acting for a reason'.

The way out of this problem for Velleman, as I see it, is to point out that the theory, if properly stated, does not make reference to the agent's actual reasons: it only makes reference to those considerations he *sees* as justifying his behavior. The theory says: *A* achieves self-knowledge if *A* (a) believes he is Φ-ing; (b) is aware that the cause of his Φ-ing is some element *E* in his motivational set; and (c) believes that *E* justifies his Φ-ing. This theory does not make use of the notion of justification or reasons any more than the theory according to which art is whatever the population is willing to refer to by the word 'art'.

The problem here is that the theory might commit Velleman to the view that anything an agent believes is a reason for action thereby counts as a reason for action. Consider: he accepts that "reasons for an action are the considerations in light of which the action would make sense".[65] If we don't want to actually build into the theory a full-blooded account of what makes sense – an account which will have to appeal to what reasons the agent has – we can limit ourselves to talking about what reasons the agent *believes* he has. It is not necessary that an agent *have* a reason for action in order for his action to make sense, only that he believes he has a reason. If we try to adopt this theory, however, anything an agent sees as a reason automatically counts as a reason. If he

[64] Velleman (2004), p.280.
[65] Ibid.

believes he is acting on a motive which gives him a reason, then that will allow him to make sense of his action, on the loose sense of that notion we are currently assuming. Since reasons for action just are those considerations which allow an agent to make sense of his action, it follows that anything the agent sees as a reason will thereby count as a reason. There is no difference between the reasons one has and the reasons one only thinks one has.

Velleman might try to avoid this problem by drawing a distinction between actions that really make sense to the agent, and those that only seem to make sense. Dancy might think that his compulsion makes sense of his action, but he is wrong: he is under the illusion that his action makes sense. While this distinction can help Velleman avoid the objection, if he takes this line he will owe us an account of the difference between actions that *really* make sense and those that only seem to make sense. The obvious way to achieve this would be to say that considerations which aren't reasons don't genuinely make sense of action: Dancy's compulsion does not justify his buying shoes that he doesn't need, so it does not really make sense of his behavior. But of course, now we are back to an account of making sense which assumes an account of reasons for action.

I conclude that autonomy or self-knowledge cannot be the constitutive aim of action, at least not if that aim is going to play the role Velleman and I require of it in the theory of practical rationality. The simple view places too few constraints on autonomous action, while the sophisticated view faces a dilemma: either it incorporates a strong criterion definition making sense which builds into the aim of action an implicit theory of

reasons; or it incorporates a weak definition of making sense which allows anything an agent believes to be a reason to thereby count as a reason.

3.2. AIMING AT THE GOOD?

If the constitutive aim of action is going to do for practical reason what the constitutive aim of belief does for theoretical reason, then it must fill the following roles:

(1) Explain why action has the success-conditions that it does;

(2) Explain why the practical norms have the content that they do;

(3) Explain the authority of practical norms: why we should care about them;

(4) Explain the motivational force of practical norms: how we can be moved by them.

Any account of the constitutive aim of action can satisfy (1) and (3), provided it takes the functional form I have been advocating. If it is the function of action to do Φ, then success in achieving Φ is the goodness-condition of action; if it is the function of action to do Φ, then we ought to care about practical norms because they help us achieve what we are trying to achieve (where this is understood non-intentionally) simply in virtue of acting. Whether an account satisfies (2) will be a more contentious matter, since the practical norms are themselves contentious. Provided that we do not go too obviously wrong in our account of what reasons we have, our theory can be considered to have satisfied (2). This is where Velleman's autonomy account failed: on the simpler construal virtually nothing was ruled out as an unsuccessful action, while the more complex account was only able to rule out more kinds of action by employing an uninformative account of reasons for action.

When we adopt the attributive view of normative properties together with the functional account of goodness, as I have been advocating, we are also in a position to see why another account of the function of action – the 'guise of the good' account – has difficulty satisfying (2). The traditional version of the guise of the good account, according to which the good is an end that an agent explicitly considers in practical reasoning can be set aside; we are here concerned with whether it can serve the functional account.[66] On this view, action has as its function the production of the 'good', whatever that turns out to be. Can the notion that action aims at the good be used for setting the standards of practical reason, and do so in a way that is in principle available to our reflection on what we have reason to do?

There are two problems with the suggestion: first, since 'good' is attributive, there is no such thing as 'the good' which action can aim at. Rather, there are many varieties of goodness, as Railton points out: "Action requires representing one's choice in a positive evaluative light, but which? There are many varieties of goodness: good for oneself, good for one's kith and kin, morally good, aesthetically good, and so on."[67] Since 'good' has this attributive character, there is no clear way to order these different kinds of goods, to say which one ought to pursued at the expense of the others. It might be suggested that even if none of these kinds of goodness can be considered better than the others in the sense of having the property 'goodness' to a greater degree, still, agents often to seem to have a subjective ordering of goods.[68] The difficulty here is that even if agents do

[66] For discussion of the more traditional account, see Velleman (1996), pp.716-717, and Setiya (2007), pp.86-98. I will return to Setiya's discussion when presenting my account of the nature of practical reason.
[67] Railton (1997), p.303.
[68] This ordering must be defined in terms of beliefs about what is better than what, if it is not going to collapse into the desire-based view that I defend. That is, the subjective

perform this sort of ordering, they would be wrong to do so on this view of evaluative terms, since the ordering does not correspond to any facts about goodness. Now suppose, not only that this view of evaluative terms is correct, but that there is an agent who *knows* that it is correct, and modifies her practices accordingly. She no longer orders these kinds of goodness in the way that other agents do. Such an agent would not be able to decide what to do by trying to achieve 'the good'. That is, an agent who was able to accurately represent to herself the features which would be relevant to her decision would, in virtue of that fact alone, be unable to make any sort of decision.

It might be objected that this sort of response raises a problem for the truth-aim account offered in Chapter 2. Just as there are many kinds of goodness, the objection goes, there are also many kinds of truth (scientific truth, philosophical truth, etc.), so if the 'many kinds of good' objection raises trouble for the guise of the good account of action, then the 'many kinds of truth' objection raises trouble for the truth-aim account of belief. But the difference between these cases, I think, is that 'true' is not an attributive adjective. The predicate 'is true' is not a function from first-order predicates to first-order predicates: there are truths of different varieties only in the sense that truth is to be found in many different intellectual domains; 'truth' picks out the same property, however, in all of these cases. The property constitutive of truth – on my view, correspondence to reality – does not change from enterprise to enterprise, in the way that the properties constitutive of goodness change from kind to kind.

Moving on to the second problem with the guise of the good account: suppose we try to avoid the first problem by singling out the goodness of actions *qua* actions as the

---

ordering cannot be simply that agents have greater desires for, or give more motivational importance to, certain kinds of goods.

aim of action, rather than moral goodness, prudential goodness, etc. There is a view of this sort which I think is right – I actually adopt it in Chapter 4 – according to which when we are engaged in practical reasoning, we are trying to determine what would be the good act *qua* act to perform. However, things become problematic if we try to say that the function of action is to produce a good act. The problem is that the expression 'good act', as used in the description of action's function, can only be given a functional interpretation on my view. With that in mind, the suggestion we are considering amounts to the following: the function of action is to do/produce the action which best fulfills the function of action. If this formula is true, it is so only trivially: it gives us no information whatsoever as to what the function of action is. The formula, taken as it is, seems to be capable of describing any account of the function of action; the formula is an abstract schema that applies to any determinate account of what a good action is, and so does not count as a determinate account itself.

If these arguments work, then the 'guise of the good' account is either false or entirely non-committal. The false interpretation leaves the determination of the good entirely open-ended – 'good' covers prudential goods, moral goods, aesthetic goods, and more – and therefore, assuming the attributive view of 'good' that I have been assuming throughout this work, gives us no way of comparing and deciding between different potential goods. The non-committal interpretation fixes the aim of action as good action, and thereby remains entirely schematic. Any account of the constitutive aim of action will count as a 'guise of the good' view on this interpretation.

3.3. SATISFACTION OF DESIRE AS THE FUNCTION OF ACTION

If neither autonomy nor the good are the constitutive aim of action, then what is? My proposal is that the satisfaction of desire is the aim of action: action, as a kind, has as its function the satisfaction of the agent's desires. Thus, an action counts as good *qua* action if it succeeds in satisfying some desire of the agent's. This suggestion has, I think, some immediate plausibility. We are creatures who take positive and negative attitudes to the world: we can dislike the way things are, and desire that they be otherwise; we can set ourselves certain goals for bringing about change in our lives; etc. Being able to act allows us to change the world so as to accord with our desires. On my account, this is the *function* of action, the achievement of which is the standard by which actions are judged.

Of course, the term 'desire', as is often the case in discussions like this, is a short form for the members of what Williams calls the agent's 'subjective motivational set', which can include "dispositions of evaluation, patterns of emotional reaction, personal loyalties, and various projects, as they may be called, embodying commitments of the agent."[69] The term 'desire' is useful in this context, because it emphasizes the non-cognitive aspect of these attitudes, their motivational force, and the point that, as I suggested earlier, evaluations of various objects or ends are not placements within a hierarchy of degrees of one property, 'goodness'.[70]

The first thing that should be noted about this view is that it does not commit us to the claim that all actions are motivated immediately only by a belief (or set of beliefs) and a desire. The view can allow that intentions, as well, can motivate actions. So, for

[69] Williams (1979), p.81.
[70] See n.20. Williams actually makes the same suggestion in "Internal and External Reasons", where he writes that the activity of attaching more weight to certain of ones ends and less weight to others "does not imply that there is some one commodity of which they provide varying amounts." (Williams (1979), p.80.)

example, if I have formulated the intention to cook spaghetti for dinner, this intention can motivate me to go through all the necessary steps, even if the desire for spaghetti has long since past.

The second thing that should be noted is a point made by Setiya, that the sort of view on offer here does not commit us to the Humean claim that beliefs and desires are distinct essences, that there can be no such state as a 'besire'. It is perfectly consistent with my view that intentions form a variety of hybrid state, having the representative quality of belief – namely, representing to oneself the proposition 'I will Φ' as being true – and the motivational quality of desire – motivating one to make it the case that 'I will Φ' is true. Setiya argues for this conception of intention as a way to avoid problems with Velleman's account of how beliefs with the content 'I will Φ' relate to intentions, and I am happy to follow him in this.[71]

In fact, this view seems to avoid one of the major problems with 'besires', namely, relating the belief-like aspect of the besire to one content while relating the desire-like aspect to another content. For example, on McDowell's view, it seems, the virtuous agent is in a state which a Humean would try to define as composed of two separate mental states, such as:

(A) Believing that *A* is a shy, sensitive person

[71] He writes: "If Velleman is right, we should expect to be moved to act in ways that confirm our beliefs about ourselves, in general – and this does not seem to be the case. I may believe that I am about to trip over a doorstep, or that I am the sort of person who is bad at keeping secrets, so that a tendency towards self-knowledge would be satisfied by tripping, or breaking a confidence, without being *motivated* to do these things, in light of my belief." (Setiya (2007), p.109, emphasis in the original.) In making this point, Setiya references Bratman (1991).

(B) Being motivated to try to include *A* in the conversation, but to do so gently, and not to raise points of discussion that will make *A* uncomfortable.[72]

McDowell wants to treat (A) and (B), not as separate mental states which exist independently of one another in the agent's mental economy, but as features of a single, unified mental state.[73] The problem with this idea is that (A) and (B) seem to describe different propositional attitudes, since they are attitudes directed at different propositions. (A) describes a belief-attitude directed at the proposition that *A* is a shy, sensitive person, while (B) describes a desire-like-attitude directed at the proposition that the agent behaves in a certain way. Even if we accept that being in a state which matches (A) requires being in a state that matches (B), we have reason to think that these two states are nonetheless distinct. Ordinarily, propositional attitudes are differentiated in part by the propositions to which they are directed: I cannot have one belief directed towards two propositions, although I might have a belief directed towards their conjunction; I cannot have one desire directed towards two propositions, although I might have a desire directed towards their conjunction. If this method of differentiation is correct, then the compound state that McDowell posits could only be accommodated by treating it as directed at the conjunction '*A* is a shy, sensitive person, and I try to include *A* in the conversation in a certain way'. But if there is a single state here, why is it that the belief-like feature is only directed at the first conjunct, while the desire-like feature is only directed at the second conjunct? After all, believing that *A* is shy and sensitive doesn't

---

[72] See McDowell (1978).

[73] He suggests that anyone who is not motivated in something like the way described in (B) does not actually understand what it means to say that someone is shy, and hence cannot be in the state described in (A) (although this criterion of understanding might need to be modified in the light of externalist and anti-individualist accounts belief content). See McDowell (1978), §4 and pp.85-86.

require that I want *A* to be shy and sensitive, and wanting to behave a certain way towards *A* doesn't require that I believe that I am so behaving, or will do so. This difference in the belief-like features and the desire-like features of the state seem best accounted for by saying that we actually have two states here: a belief directed at one proposition, and a desire directed at another.

To take another example, Michael Smith presents the besire theorist[74] as holding the following view of moral judgments:

> "The besire that Φ-ing is right is appropriately described as being in a state that must fit the world because it tends to go out of existence when the subject is confronted with a perception with the content that Φ-ing is not right…Moreover…the besire that Φ-ing is right is also appropriately described as being such that the world must fit with it. For a subject's having the besire that Φ-ing is right disposes her to Φ."[75]

The problem here is that the belief-like aspect is directed at the proposition that Φ-ing is right. If that is the case, then the desire-like aspect is not actually directed at the same proposition, since it is directed at the proposition that the subject Φs; it is a desire that 'I do Φ' be the case, and which motivates the subject to Φ. If the two aspects of the state *were* directed at the same proposition, then the besire that Φ-ing is right would incorporate a desire-like state directed at the proposition that Φ-ing is right, and would motivate the subject to bring it about that Φ-ing is right. That is obviously not how moral judgments motivate us. Believing that giving to charity is right doesn't motivate me to bring it about that giving to charity is right; it motivates me to give to charity. Intentions, as Setiya thinks of them, do not have this problem that other sorts of besires do, since the

[74] In particular, he is interested in J.E.J. Altham's discussion of the idea that moral judgments have this character; see Altham (1986).
[75] Smith (1994), pp.118-119.

belief-aspect and the desire-aspect are directed towards the same proposition, the proposition 'I will Φ'.

Third, and again with indebtedness to Setiya, we should note that the desire-based account does not commit us to the view that practical reasoning must always take the form of explicitly taking the operative goal as an end, and reasoning from this together with a belief to a new desire or goal. We should reject this view, since it is possible for one to engage in instrumental reasoning without explicitly considering the goal one is advancing; it is enough for the goal or desire to play a background role. For instance, I need not explicitly reason as follows:

(1) I desire to own the new book by Mark Schroeder.

(2) By clicking the 'Buy' button, I will thereby come to own the new book by Mark Schroeder.

Therefore,

(3) I have a reason to click the 'Buy' button.

I need only consider (2) and thereby come to believe (3). This inference is a rational one provided I do actually have the desire mentioned in (1).[76] I think this is the right way to go, since it has a clear parallel in theoretical reasoning. I need not reason explicitly as follows:

(4) Global skepticism is false.

(5) If theory *T* is true, then global skepticism is false.

Therefore,

(6) Theory *T* is false.

[76] Setiya (2007), p.100. Setiya references Pettit and Smith (1990).

I need only consider (2) and thereby come to believe (3). This inference is a rational one provided I do actually believe (1).[77] It seems that both beliefs can desires need play only a background role in good reasoning.

With these points out of the way, we turn to the motivation behind the desire-based view. Why should we think that the function of action is to satisfy the agent's desires, goals, etc.? Recall the four roles that the constitutive aim of action must fill:

(1) Explain why action has the goodness-conditions that it does.

(2) Explain why the practical norms have the content that they do.

(3) Explain the authority of practical norms: why we should care about them.

(4) Explain the motivational force of practical norms: how we can be moved by them.

As I said earlier, *any* function account can claim to satisfy (1) and (3), in the same way that the truth-aim account satisfies them in the case of belief. Moreover, (2) will be, almost as a matter of necessity, a contentious issue. Just what the norms governing action are is a matter of some contention, and no doubt some will remain unconvinced by an account which fails to yield the result that certain norms actually apply to action. However, as long as the account gets the obvious cases right, the more contentious ones should not be allowed to put too much pressure on it. We have already seen that the autonomy and goodness accounts fail to satisfy (2). Neither account was able to provide any substantive norms of practical reasoning. The account I am offering, by contrast, *does* provide substantive norms. At the very least it allows us to account for instrumental

[77] In both of these examples, I have left to one side issues about how the conclusions of one's inferences become justified. In particular, I have left to one side the question of whether good inferences generate justification, or merely transmit it from premises to conclusion. I think we can assume that the beliefs expressed in (1) and (4) are both rationally held, and leave these questions to later chapters.

norms: norms which guide agents to perform those actions which will satisfy the members of their motivational sets.

In addition, I think the account can satisfy (4), although the details of this point will have to wait until the next chapter, where I present my view of how desires function in practical reasoning. The basic idea, however, is that practical norms can motivate us through our general disposition to want to perform those acts which we believe will satisfy the elements in our subjective motivational sets.

3.4. OBJECTIONS TO THE DESIRE-BASED VIEW

3.4.1. THE HUMEAN THEORY OF MOTIVATION

Setiya has recently argued that a desire-based account is committed to the Humean theory of motivation, and that this theory is false. His argument is based on what he calls the Difference Principle:

> *The Difference Principle:* If *F*'s are a kind of *G*, and being a *good F* is not simply a matter of being an *F* that is a *good G*, there must be something in the distinctive nature of *F*'s to explain or illuminate the difference.[78]

I have myself assumed this principle, since it is informed by the attributive view of evaluative predicates, although I have placed a restriction on what sorts of facts about the nature of *F*'s are relevant, namely, facts about the constitutive function of *F*'s. The question is: why should we take the desire-based account and the Difference Principle to imply the Humean theory of motivation? I think Setiya's argument depends on his focus, not on the nature of *action*, but on the nature of *dispositions of practical thought*. The overall argument of his book is that good dispositions of practical thought are just good

[78] Setiya (2007), p.83.

character traits, so his goal is to argue against theories which posit something distinctive about the nature of dispositions of practical thought. If the desire-based account took this form – that all dispositions of practical thought are dispositions to be moved to action by at least one operative desire – then it would seem that we were committed to the Humean view. If all dispositions of practical thought had a desire as input, then in order to satisfy the Difference Principle and show that such dispositions are not simply good character traits, we would need to adopt "the part of the Humean theory that requires the motivation of action and desire always to be traced in part to a prior desire, where it is not assumed that desires are categorically distinct from beliefs."[79]

The account I am offering, however, makes no such commitment regarding the nature of dispositions of practical thought. On my view, satisfaction of desire is the function of action, and this function places constraints on what counts as a *good*, or *effective* disposition of practical thought, not on what counts as such a disposition *simpliciter*. To see this, consider the truth-aim account of belief from Chapter 1: the account does not imply that all dispositions of theoretical thought are dispositions to believe on the basis of evidential considerations, or on the basis of the inputs of reliable methods, etc.; it only implies that *good* dispositions must have these sorts of features. The very possibility of a disposition to believe propositions because their truth would make one happy is not ruled out (although the subject in question would need to possess *other* dispositions – such as the disposition to treat countervailing evidence as relevant to the quality of the belief – in order for this disposition to be thought of as a disposition to form *beliefs*). Likewise with the practical case: the very possibility of being motivated to act (or perhaps, to form an intention) on the sole basis of a belief, one which is not relevant

[79] Ibid., pp.101-102.

to the possibility of satisfying one's desires via the act in question, is not ruled out. But such a disposition will not count as a *good* disposition of practical thought, since it is not properly connected to the elements in the agent's subjective motivational set.

3.4.2. DESIRES AND THE VARIANCE OF REASON ACROSS PERSONS

Even *this* commitment might seem too strong, however. Why should we think that a good practical inference must involve a desire?[80] Why not allow what Brandom calls 'material inferences', like the following:

> (A) Only opening my umbrella will keep me dry, so I shall open my umbrella
>
> (B) I am a bank employee going to work, so I shall wear a necktie
>
> (C) Repeating the gossip would harm someone, to no purpose, so I shall not repeat the gossip[81]

are perfectly sound practical inferences just as they stand? I think the answer can be seen when we recognize the right requirements for being able to understand and rationalize behavior from a third-person perspective. In order to see someone else's action as reasonable, it is not necessary that I see the considerations on which he acts as reason for me to act, or reasons for anyone else other than him to act. So, for example, in order to understand and rationalize the actions of an agent who actually goes through the inferences (A)-(C), I need not be able to see the premises as reasons that anyone would have for acting in the relevant ways. I can imagine myself in the same situations as the

[80] I will argue later that desires strictly speaking do not provide reasons, but that only *beliefs* about desires do. That is, I don't think that a desire-based account like mine commits me to a Humean view about *reasons* any more than it commits me to a Humean view of motivation. All the cases I will be dealing with in the remainder of this chapter should be read as assuming that the agents in question believe that they have the desires in question, and that they are reasonable in holding those beliefs.

[81] Brandom (1998), p.469.

agent, yet still think that the premises do not give me reason to act as the conclusions dictate: perhaps I like getting wet from walking in the rain; perhaps I would hate working at a bank, and this hatred would compel me to disregard dress expectations; perhaps I get great pleasure from doing harm to others, even when doing so provides no other benefit to me. And yet, in such cases, I could still see the premises as good reasons *for the agent*. I can recognize that someone is rational to act on these premises even if I don't see them as reasons for action. In fact, a great deal of practical reasoning seems to have just this feature.

Suppose we were dealing with a very strong form of the view that desires are not necessary for good practical inference: suppose, that is, that someone held the view that *all* practical inferences functioned in the same way that (A)-(C) do. On such a view, seeing a fact *R* as a reason is just being disposed to act in certain ways when one believes that it obtains. It seems to me that this sort of view would be unable to accommodate the phenomenon I have been describing. If I don't have the disposition in question, then I don't see *R* as a reason for action; how, on this view, could I come to see it as a reason for others to act? The obvious response would be: I can see *R* as a reason for others by recognizing that they have a disposition to act on it in certain ways. The problem with this response is that I can recognize this disposition in others without seeing this disposition as rational in any way, and hence without seeing it as providing that person with a reason for acting the way they do. What, according to the view in question, distinguishes the cases where I see the disposition as providing reasons from those where I don't see the disposition in this way? It cannot be that I recognize the disposition as providing reasons when and only when I also have the disposition: as we saw, I can see

the cases (A)-(C) as providing someone else with reasons for action even when I don't have the relevant dispositions to act.

I think the easiest way to recognize that a certain fact can provide reasons for some people but not for others is to posit a desire, goal, commitment, or the like which facilitates the inference from certain facts to certain intentions and actions. Mark Schroeder provides a good example:

> "Tonight there is going to be a party, and everyone is invited. There will be good food, drinks, friendly chat, music – and dancing. Ronnie and Bradley, like everyone else, have been invited to the party. But while Ronnie loves to dance, Bradley can't stand it. Not only does he not like dancing, he prefers to stay away from where it is going on, lest he come under pressure to be shown up in his awkward maneuvers by those with fewer left feet than he. So while the fact that there will be dancing at the party is a reason for Ronnie to go, it is *not* a reason for Bradley to go. Far from it; the fact that there will be dancing at the party is a reason for Bradley to stay away. Ronnie's and Bradley's reasons therefore differ – each has a reason that the other does not."[82]

It is, of course, easy to see that Ronnie and Bradley have different reasons in this case. And as Schroeder says, it is easy to see in this case *why* their reasons are different: "It is because of what they *like*, *care about*, or *want*."[83] What is most important for my purposes, however, is that Ronnie and Bradley, if they are rational, should be able to see why each of them has different reasons. Ronnie will be able to see that the fact that there will be dancing at the party gives Bradley a reason *not* to go, and Bradley will be able to see that it provides Ronnie with a reason *to* go. Positing desires in this way – and viewing good practical inferences as inferences about the satisfaction of one's desires – allows us

[82] Schroeder (2007), p.1. In the next chapter I will deny that desires, strictly speaking, provide reasons for action. Rather, I will claim, it is our beliefs about our desires which provide reasons for action, and they can play this role even when they are false and we do not actually have the desires we think we do. I assume that in Schroeder's example Ronnie and Bradley believe that they have the desires Schroeder discusses.
[83] Ibid.

to explain how we can acknowledge that different people have different reasons, in a way that the material-inference view does not.

(It might be thought that on this score the desire-based view does not have any advantage over Velleman's autonomy view: Velleman could just as easily say that the reasons one has are determined by the elements in one's subjective motivational set. The difference is that, as we saw, Velleman's unrestricted account gives agents too many reasons, since it allows that any action we would predict that we would do would thereby make sense to us. In order to provide a more plausible account of what 'makes sense' to the agent, Velleman needs to restrict himself to certain elements in the subjective motivational set, and his autonomy view provides no principled way to pull off this restriction. Since my view is not that the view of action is to do what we believe we will do, I do not need to start down the road of saying what kinds of predictions make the most sense to agents.)

Here we begin to see the kind of relativity of reasons that comes along with the desire-based view, and which Velleman explicitly wanted to avoid. If the satisfaction of desire is the constitutive aim of action, then an action counts as good or successful to the extent that it satisfies these desires. If there is no guarantee that all agents will have the same desires, then there is no guarantee that there will be actions which it would be good for all agents to perform in the same circumstances. That is, even if all agents in that circumstance reason their way to the correct answer about what the best thing for them to do is – what action, of those available, would best satisfy the constitutive aim of action – they will not necessarily be reasoning their way to the same belief. For some agents, the best thing to do will be to Φ, for some it will be to Ψ, and so on. On my view, this fact

does not automatically entail that all such agents will have *reason* to do different things[84], but we will see in the next chapter that a difference in what we have reason to do will often accompany a difference in what would best satisfy the constitutive aim of action.

Setiya has recently raised a problem for the kind of argument I have just given, that we need to posit desires in order to mediate the inference from beliefs to intentions or actions. The problem is that we might face a regress: someone who takes my position needs to explain why we are justified in positing a desire to mediate the disposition to act on the basis of a belief, but we are not justified in doing so for other practical dispositions, for instance, the disposition to form a desire or intention to Φ on the basis of the desire to Ψ together with the belief that by Φ-ing one will Ψ. Positing the latter will, of course lead to a regress: if we must posit the desire to form the intention to perform the means to one's ends, then we must also posit a desire to explain the inference from the input belief and desire, plus the desire to form the desire to perform the means to one's ends, and so on. At some point we need to posit a brute disposition to move from one mental state to the other, and to count this as a practical inference in its own right. What we need, then, is an independent reason to posit a desire to mediate the move from a belief to an intention or action, which will not *also* justify positing a desire to mediate the inference from the belief *plus* the desire to the intention or action.

[84] To anticipate: on my view, *A* has a reason to Φ iff. *A* has a reason to believe that Φ-ing would best satisfy the constitutive aim of action. On this view, the following scenario is possible: *A* and *B* are in the same circumstance, trying to figure out what to do. Φ-ing is what would best satisfy the constitutive aim of action for *A*, since it would best satisfy his desires; Ψ-ing is what would best satisfy the constitutive aim of action for *B*, since it would best satisfy his desires. However, *A* does not believe that he has the desires which would be satisfied by Φ-ing, but he does have a reasonably-held belief that he has the desires that *B* has (we could even stipulate that *A* in fact has those desires). *A* forms the reasonably-held belief that Ψ-ing would best satisfy his desires, unaware that in fact Φ-ing is what would best satisfy this aim. In such a case, I say, *A* and *B* both have reason to Ψ.

This is another point at which the function account shows its strength. I have already argued that desires are required to explain the difference of reasons across persons, and to explain how we can recognize the reasons that others have but which we lack. However, desires *aren't* required to explain the inference from desires and beliefs to intentions on the function account, because the aim of satisfying one's desires is not just another among one's ends. The move from desires and beliefs to intentions doesn't require a further desire, because the inference is a good one simply in virtue of its being an inference which is conducive to the satisfaction of the function of action. A simple inference from a belief to an intention does not have this feature, so we have a principled reason to draw the line where the desire-based account draws it.

CHAPTER 4: REASONS FOR BELIEF AND ACTION

Having argued that the constitutive aim of belief is truth, and that the constitutive aim of action the satisfaction of desire, I can now move on to discuss how these aims can be used to provide accounts of theoretical and practical rationality. This is, of course, a large project, but I hope that this chapter will give some idea of how it can be done. In particular, I will argue for a unified account of theoretical and practical reasoning, one which does not make a sharp distinction between the two varieties of reasoning. In Section 4.1 I outline my view of the aims of theoretical and practical reasoning: namely, the production of true beliefs about what propositions are most likely to be, and what actions are most likely to satisfy one's desires, respectively; reasoning aims at the production of true beliefs about which belief or action is most likely to fulfill the constitutive of aim of those kinds. I outline a process-reliabilist view of correct belief-forming methods and use it to explain the conditions under which we have reason to hold the beliefs which I claim reasoning aims at. Finally, at the end of Section 4.1 I defend the view that, although the constitutive aim of action is the satisfaction of desires, desires themselves do not provide an agent with reasons for action. Rather, only an agent's rationally-held beliefs about what desires she has – beliefs which may be mistaken – can provide her with reasons for action. I thus defend a cognitivist theory of practical reasoning, that is, one which takes beliefs as inputs and produces beliefs as outputs. With the basic view of rationality and reasons in place, I go on to defend it against a competitor theory which has been defended by Michael Smith, and which I call the 'Ideal Observer Theory' or 'IOT'. In Section 4.2 I outline this theory, according to which the question of what reasons an agent has must be answered by reference to the actions or

attitudes of a suitably idealized version of the agent in a counterfactual world. Finally, in Section 4.3 I argue that the IOT is false. I move through the three criteria that Smith provides for being a 'suitably idealized version of the agent', and argue that (a) it is possible for an agent to have a reason for action where the IOT says that she doesn't, and (b) it is possible for an agent to fail to have a reason for action where the IOT says that she does have a reason.

4.1. A UNIFIED ACCOUNT OF THEORETICAL AND PRACTICAL REASON

When we engage in theoretical reasoning, we are trying to determine which, of the available beliefs on the subject we are considering, is the most likely to be true. When we engage in practical reasoning, we are trying to determine which, of the available actions in our circumstances, is the most likely to satisfy our desires. This difference in the objects of theoretical and practical reasoning has sometimes been taken to motivate a further distinction between theoretical and practical rationality, namely, that theoretical rationality is a cognitive matter, whereas practical reasoning is, at least in part, a non-cognitive matter. Theoretical reasoning leads us to form beliefs, which represent the world as being a certain way, while practical reasoning leads to form intentions, desires, and the like, which are typically thought to be non-cognitive states: they represent a way the world might be as something desirable or to-be-brought-about, rather than as the way the world actually is.[85] On my view, by contrast, theoretical and practical rationality are both cognitive: they both lead us to a certain kind of belief.

On the function account of goodness properties, an object of kind *K* is a good *K* if and only if it has some property *F* which allows it to adequately fulfill its function *qua K*.

[85] See, for example, the opening paragraph of Bratman (1991).

The better a *K* fulfills the constitutive function of *K*s, the better a *K* it is. A *K* that perfectly fulfills its function *qua K* is the best kind of *K* there is. In the case of belief, the constitutive aim is truth: beliefs by nature have as their function the accurate representation of reality. If a belief is true, then it is a good belief *qua* belief. In the case of action, the constitutive aim is the satisfaction of desires: actions by nature have as their function the satisfaction of the agent's desires. If an action performs this function, then it is a good action *qua* action. I have argued that these functions of belief and action can serve as setting the standards for theoretical and practical reason, respectively. What this means is that when we are engaged in theoretical reasoning we are trying to come to true beliefs; when we are faced with a choice between different possible beliefs we could hold, we want to know which one of those beliefs is the best one *qua* belief, which is to say, which one is true. In practical reasoning we are trying to determine which action, of those available to us, is the best one *qua* action, which is to say, which one will best satisfy our desires.

What differentiates theoretical and practical reasoning, if they are both cognitive enterprises aimed at the formation of true beliefs? There are, I think, two related features which allow us to distinguish the two kinds of rationality, although the distinction will not be the sharp one implied by a non-cognitive view of practical reason. First, each variety of reasoning is aimed at forming beliefs with a certain kind of content. In the theoretical case, the belief is second-order: it is a belief about beliefs:

> Of all the beliefs I could possibly form in my circumstances, the belief that *p* is the belief which is most likely to be true; the belief that *p* is the belief

which would come closest to achieving the minimum level of evidential support required in my circumstances.

By contrast, practical reasoning is not aimed at the formation of second-order beliefs; rather, it is aimed the formation of first-order beliefs about actions:

Of all the actions I could possibly perform in my circumstances, the action Φ is the action which is most likely to satisfy my desires.

This first difference is what generates the second difference between the two kinds of reasoning, which is a difference in the disposition which is required for one's reasoning to be effective. We can distinguish between two kinds of rationality:

Inferential Rationality: This consists in being able to make the right inferential moves of the sort I have been describing, and to properly form beliefs about what is most likely to be true, and what is most likely to satisfy one's desires.

Calibrational Rationality: This consists in having one's first-order beliefs, or one's actions and intentions, conform to the conclusions of one's inferential reasoning.

There is some intuitive reason to distinguish these. For instance, David Owens points out, it seems that a believer or an agent can suffer from different forms of irrationality.[86] For instance, I might suffer from an inability to reliably determine what beliefs are most likely to be true, perhaps because I am bad at balancing evidence, or I am bad at following logic inference-rules. But in addition to this, I might suffer from an inability to believe what I think is most likely to be true: I might think that theory T is the most likely to be true, and yet for some reason retain my belief in theory T+. The same holds true of

[86] Owens (2000), p.104. I discuss Owens' view in more detail in Chapter 5.

my beliefs about what action is most likely to satisfy my desires, and my actual intentions and actions; I might be bad at figuring out what action will most likely satisfy my desires, or I might fail to act as I think would be best for me to act. In both cases, failures of calibrational rationality are a kind of *akrasia.* What distinguishes the two is what it is that must be properly 'calibrated'. In the theoretical case, my first-order beliefs must be calibrated to my second-order beliefs; in the practical case, my actions and intentions must be calibrated to my beliefs about actions.[87] (In Section 4.1.4 I will argue that the disposition to do what you think will best satisfy your desires is a disposition that is constitutive of rational agency.)

If this view of the goal of reasoning is correct, then both theoretical and practical reasoning are aimed at forming true beliefs, although in the theoretical case these will be second-order beliefs: beliefs about beliefs. That is, in theoretical reasoning we want to know which first-order belief is true, or at least which one is most likely to be true; in practical reasoning we want to know which action best satisfies our desires, or at least which one is most likely to do so. Both kinds of reasoning, then, are aimed at forming a

[87] This view of calibration, I think, suggests that the difference in the objects of inferential rationality in the theoretical and practical spheres is as I suggest that it is. It might be thought that practical reasoning *also* leads to second-order beliefs: beliefs about which belief to the effect that 'Φ is the action which will most likely satisfy my desires' is most likely to be true. But if calibrational rationality in the practical case consists in getting one's actions or intentions to accord with one's first-order beliefs of this sort, rather than getting those first-order beliefs to accord with one's second-order beliefs, then it seems that we have reason to model inferential rationality in the practical sphere as I have done. (Thanks to Duncan MacIntosh for raising this issue.

That being said, adopting the second-order belief model of practical reasoning would not be devastating to my view. At worst, I think, it would imply that there is an extra step involved in practical reasoning. Although the circuitous route from second-order beliefs to first-order beliefs to intentions or actions is *longer* than I think is required, I see no reason to completely rule this out *a priori*. It might even provide for a greater unity between theoretical and practical reason, which might balance out the cost of treating practical reason in this circuitous way.

true belief. This view already suggests that both kinds of reasoning are cognitive, but in Sections 4.1.3 and 4.1.4 I will argue that it is not only the conclusion of a practical inference which is a cognitive state, but that the materials upon which practical inferences operate are also cognitive states: practical reasoning, like theoretical reasoning, involves inferring certain propositions from other propositions. Before presenting that argument, however, I will flesh out my conception of what good reasoning actually involves, and the conditions under which one has a reason for belief.

4.1.1. THE RELIABILIST CONCEPTION OF PROPER BELIEF-FORMATION

It might be thought that the function account of the goodness-conditions for beliefs and actions implies an implausible conception of rationality. Since 'rational', 'reasonable' and the like are evaluative terms, it might seem that the function account implies that a perfectly rational belief is just a belief that fulfills the function of beliefs, and that a perfectly rational action is just an action that fulfills the function of actions. That is, it might seem that the function account implies that a perfectly rational belief is just a belief that is true, and that a perfectly rational action is just an action which best satisfies one's desires. (I suspect that this idea is what motivates Owens' skepticism – which I mentioned in passing in Chapter 2 – that the truth-aim account of belief and the resulting account of rationality could account for the fact that it is sometimes rational to believe a contradiction. Since contradictions are necessarily false, if being a rational belief just consisted in being true, and being an irrational belief just consisted in being false, then it could not be the case that anyone ever has a reason to believe a contradictory proposition.

As Owens rightfully says, however, there are conditions under which one can believe a contradiction without thereby failing to be rational.[88])

This sort of view is obviously too strong, since I can come to a false belief, or perform an action which doesn't satisfy my desires, without having failed to be rational in any way: a belief can be perfectly rationally formed even if it is false, and an action can be perfectly rationally performed even it is doesn't satisfy the agent's desires. I think the view of reasoning that was sketched at the beginning of this section can avoid falling into such an implausible account of rationality, however, by shifting the locus of rationality from the belief or action considered in isolation, to the reasoning process which led to it. That is, roughly, a belief is rationally held if the second-order belief that this belief is the most likely, of those available, to be true, was properly formed (although here we need to remember the role that the norms governing inquiry play in determining how much evidence one needs to have in order for a proposition to count as 'sufficiently likely to be true'). An action is rationally undertaken if the belief that this action is the most likely, of those available, to satisfy one's desires, is rationally held by the agent. (This description is rough, because it needs to be supplemented (a) by an account of proper belief formation, and (b) by a clause which allows a belief or action to be rational even when the relevant belief taken towards it – that it is likely to be true, or to satisfy one's desires, respectively – is not explicitly formed.)

So, what is it for a belief to be properly formed? In this work, I will be adopting a process-reliabilist conception of proper belief-formation which is inspired by the work of Alvin Goldman. This approach seems to me to fit well with the function account of goodness-properties: a belief-forming method is to be evaluated in terms of its reliability,

[88] Owens (2003) pp.287-288.

where reliability consists, not just in producing a high number of beliefs, but in producing a high number of *good* beliefs, which is to say, a high number of *true* beliefs. Just as a technique for making coffee gets evaluated by reference to how reliably it produces good coffee, a belief-forming method gets evaluated by reference to how reliably it forms true beliefs.

Following Goldman, we can distinguish two kinds of reliability: 'conditional' and 'unconditional'; these sorts of reliability correspond, respectively, to 'belief-dependent' and 'belief-independent' processes. The former sorts of processes – such as memory and reasoning procedures – take beliefs as their inputs. The latter sorts of processes – such as perception and rational intuition – do not take beliefs as their inputs: rather, they take in such things as experiential data. This distinction is crucial, since a belief-dependent process should not be evaluated *simply* as to how often its output beliefs are true. On this point, Goldman writes: "A reasoning procedure cannot be expected to produce true belief if it is applied to false premises. And memory cannot be expected to yield a true belief if the original belief it attempts to retain is false."[89] A process which does not take beliefs as inputs is reliable if it produces a sufficiently high number of true beliefs, while a process which takes beliefs as inputs is reliable if it produces a sufficiently high number of true beliefs when given true beliefs as inputs.

We should also note that a subject can form a rational belief even if she reasons her way to it via a conditionally *un*reliable process. Suppose that two people disagree as to what the correct modal logic is: one of them insists on adopting S5, while the other thinks that S4 is the correct logic. At most one of these logics can be correct: if S5 is correct, then the S4 theorist will fail to draw inferences which are actually truth-

[89] Goldman (1976), p.340.

preserving (namely, inferences from '◊□*p*' to '□*p*'); if S4 is correct, then the S5 theorist will constantly be drawing inferences which aren't truth-preserving. However, it seems that as long as each theorist has adopted their logic because they have reached the belief that it is the correct logic via reliable belief-forming processes, the S4 theorist will be rational in refraining from drawing certain inferences, while the S5 theorist will be rational in drawing them. Even if S5 is the correct logic, the S4 theorist would be irrational to reason from '◊□*p*' to '□*p*' if she reasonably believes the inference to be a bad one. Likewise, even if S4 is the correct logic, the S5 theorist would be irrational to refraining from inferring '□*p*' from '◊□*p*' if she reasonably believes the inference to be truth-preserving.

Finally, we should take note of what Goldman calls '*ex ante*' justification, where a subject does not actually believe a proposition *p*, but has some sort of justification for *p* available to him; he is justified in believing *p* without having a justified belief that *p*. Michael Smith's example of 'Blanking John' makes this idea vivid: John is engaged in a philosophical debate, and is unable in the moment to come up with a satisfying answer to a difficult question. Later on in the day, however, the answer comes to mind, and on reflection the answer seems completely obvious. Perhaps the answer relies on a style of argument that John had already used elsewhere in the discussion, and he needed only to apply it to the case at hand.[90] In a sense, John already had the justification for the answer. Likewise, we can imagine cases where someone has evidence for *p*, but he fails to use it, perhaps because he is distracted. Nevertheless, it seems we can say that he has a reason to believe *p*, even though he has not put this reason into practice and formed the belief.[91]

[90] Smith (2003), p.20.
[91] Goldman (1976), pp.344-345.

4.1.2 WHEN DO YOU HAVE A REASON FOR BELIEF OR ACTION?

The adoption of the reliabilist conception of justification of the kind I have been sketching suggests the following principles regarding when a subject, S, has reason to believe a proposition, *p*:

> S has reason to believe *p* iff. either:
>
> (i) S believes *p*, and this belief was formed by an unconditionally reliable method; or
>
> (ii) S believes *p*, and this belief was derived from beliefs which are reasonable in the sense of (i), via conditionally reliable methods; or
>
> (iii) S believes *p*, and this belief was derived from beliefs which are reasonable in the sense of (ii), via conditionally reliable methods; or
>
> (iv) S believes *p*, and this belief was derived by methods which are either conditionally unreliable or unconditionally unreliable, but S believes these methods to be either conditionally reliable or unconditionally reliable, and *this* belief satisfies one of (i)-(iii); or
>
> (v) S does not believe *p*, but *p* is derivable, by a belief-dependent method which is either conditionally reliable or such that S has a belief that it is conditionally reliable which satisfies one of (i)-(iii), from propositions which S does believe, and these beliefs themselves satisfy one of (i)-(iv).

Note that on this view of reasons one does not automatically have a reason to believe a proposition *p* if one has derived *p* from other propositions via conditionally reliable methods. Rather, one only has a reason to believe *p* in this sort of case if one has a reason to believe the propositions from which it is derived. Consider inferences via logically

valid inference-rules. On my view, such rules are reason-*transmitting* rather than reason-*generating*; just as *modus ponens* only preserves truth, rather than ensuring that whatever propositions we derive via *modus ponens* will be true, regardless of the truth of the premises, so it only preserves reasons, rather than ensuring that we have a reason to believe whatever propositions we derive via *modus ponens*, regardless of whether or not we have a reason to believe the premises of our inference. This seems to me a correct account: I only have a reasonable belief in the conclusion of a logically valid argument if I have reason to believe the premises.

This conception of reasons stands in contrast to another conception, one which has proven influential, particularly in the literature on practical reason. Before turning to discuss this opposing view, however, it will be helpful to deal with an objection to my own conception, an objection I have already hinted at. It will no doubt have been noticed that only beliefs have been mentioned in principles (i)-(v), but practical reasoning is often thought to deal with desires in addition to beliefs. Even if desires are treated as *analogous* to beliefs in some way, theories of rationality tend to differentiate theoretical and practical reasoning precisely by claiming that theoretical rationality deals with cognitive states such as belief, while practical reasoning stems from non-cognitive states such as desires. My account makes no such distinction. It is my view, to be defended in the next subsection, that although desires may provide motivation for action, they do not provide reasons for action; only *beliefs about desires* provide reasons.

4.1.3. DESIRES ARE NOT CONSTITUTIVE OF REASONS

In order to make clear what I mean when I say that desires are not constitutive of reasons, it will be helpful to briefly examine how Michael Smith spells out the view that desires are constitutive of one's *motivating* reasons: those elements of one's mental economy which explain why one acts as one does, although without necessarily justifying the action:

> *The Constitutive View:* "R at t constitutes a motivating reason of agent A to Φ iff there is some Ψ such that R at consists of an appropriately related desire of A to Ψ and a belief that were she to Φ she would Ψ."[92]
>
> *The Non-Constitutive View:* "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 she to Φ she would Ψ."[93]

It is one thing to say, as the non-constitutive view does, that in order for someone to have a motivating reason to Φ they must have an appropriate desire coupled with a means-ends belief; it is another thing to say that the belief and desire themselves constitute one's motivating reason – that they explain why one does what one does. Thomas Nagel, for example, adopts the view that having a motivating reason requires having an appropriate desire, but he does not accept that the desire always plays the causal-explanatory role that the constitutive view attributes to it. He writes:

> "The claim that a desire underlies every act is true…only in the sense that *whatever* may be the motivation for someone's intentional pursuit of a goal, it becomes in virtue of his pursuit *ipso facto* appropriate to ascribe to him a desire for that goal… That I have the appropriate desire simply *follows* from the fact that these considerations motivate me; if the likelihood that an act will promote my future happiness motivates me to perform it now, then it is

[92] Smith (1994), p.92.
[93] Ibid., p.93.

> appropriate to ascribe to me a desire for my own future happiness. But nothing follows about the role of the desire as a condition contributing to the motivational efficacy of those considerations. It is a necessary condition of their efficacy to be sure, but only a logically necessary condition. It is not necessary either as a contributing influence, or as a causal condition."[94]

The difference between the constitutive and non-constitutive views is captured by their logical form: the latter only posits the desire/belief pair as a necessary condition of having a reason, while the former posits it as both necessary and sufficient.

In principle, this distinction can also be made with regard to *justifying* reasons, those features of our mental economy which can be used, not only to give a causal explanation of why we act the way we do, but to show that our actions are justified or rational, when they are. On the non-constitutive view, one must have the appropriate desire/belief pair in order to have a justifying reason, but the pair is not necessarily constitutive of one's reason; on the constitutive view, the belief/desire pair is both necessary and sufficient, since having the pair *just is* having a reason for acting. I will not deal with these two views separately, in the way it might be appropriate to deal with the two views of motivating reasons. Rather, I will simply argue that each conditional:

> *Necessity:* Agent A at t has a justifying reason to Φ only if there is some Ψ such that, at t, A desires to Ψ and believes that were she to Φ she would Ψ;
>
> *Sufficiency:* If there is some Ψ such that, at t, agent A desires to Ψ and believes that were she to Φ she would Ψ, then A has a justifying reason to Φ;

is false. Having the relevant belief/desire pair is neither necessary nor sufficient for having a reason to act: one can hold the belief, and possess the desire, yet still fail to have a justifying reason to act; and one can hold the belief while lacking the desire, yet still have a justifying reason to act. As the reader will probably have noticed, my focus is on

[94] Nagel (1970), pp.29-30.

the presence or absence of the desire, and what effect this has on an agent's reasons. I will argue by example that, assuming the agent has the relevant means-ends belief, it is not the presence or absence of the appropriate desire which determines whether or not one has a reason for action; rather, on my view, it is the presence or absence of a belief to the effect that one has the appropriate desire – in particular, a belief which is rational in the sense sketched so far in this chapter – which determines whether or not an agent has a reason to act. After presenting the examples I will deal with a potential objection that the cognitivist view of practical reasoning which results from my position is in tension with my desire-based account of the success-conditions on action. I hope that my response to this objection will not only assuage the worries about a tension in my view, but also bring to light considerations which add to its plausibility.

The examples I will use to defend my view come from Michael Smith's argument against the phenomenological conception of desires, according to which desires have distinctive 'feels' to them, so that we are always in a position to know what we desire. First, consider a case where a person has a desire, but is unaware of it:

> "Suppose each day on his way to work John buys a newspaper at a certain newspaper stand. However, he has to go out of his way to do so, and for no apparently good reason. The newspaper he buys is on sale at other newspaper stands on his direct route to work, there is no difference in the price or condition of the newspapers bought at the two stands, and so on. There is, however, the following difference. Behind the counter of the stand where John buys his newspaper, there are mirrors so placed that anyone who buys a newspaper there cannot help but look at himself. Let's suppose, however, that if it were suggested to John that the reason he buys his newspaper at that stand is that he wants to look at his own reflection, he would vehemently deny it. And it wouldn't seem to John as if he were concealing anything in doing so."[95]

[95] Smith (1994), p.106.

It seems in this case that John's desire to see his reflection is what *motivates* his action, but does it give him a reason to act as he does? I don't think that it does. He cannot reflect on his mental economy and pinpoint the desire which his action satisfies. We, from our external perspective, might be able to recognize that desire, and hence be able to attribute to him a *motivating* reason; we can provide a causal explanation of his behavior. But this does mean that we can rationalize his behavior. We cannot justify his behavior on the grounds that it satisfies his desire, because he is completely unaware that it does so. If John were to be told about his desire, and form the rational belief that his walking to this particular newsstand better satisfies his desires than walking to another newsstand, *then* he will have a reason to continue going there, and be able to justify future trips to himself. But that will not allow him to see his earlier ventures as justified, any more than I can see my previously held beliefs as justified because I have recently acquired evidence for them that I did not previously possess.

Second, consider a case where a person doesn't have a desire, but thinks that they do:

> "Suppose John professes that one of his fundamental desires is to be a great musician. However, his mother has always drummed into him the value of music. She is a fanatic with great hopes for her son's career as a musician, hopes so great that she would be extremely disappointed if he were even less than an excellent musician, let alone if he were to give up music altogether. Moreover, John admits that he has a very great desire not to upset her, though he would, if asked, deny that this in any way explains his efforts at pursuing excellence in music. However, now suppose John's mother dies and, upon her death, he finds all of his interest in music vanishes. He gives up his career as a musician and pursues some other quite different career."[96]

In this case, we can imagine John's practical reasoning being based on his belief that he desires to be a great musician. He might, for example, decide to practice the piano for

[96] Ibid.

two hours a day – even though he does not enjoy practicing in the least – on the grounds that doing so will be a good way to pursue his goal of being a great musician. Does John, in this case, have a reason to practice the piano for two hours a day? I think the answer is 'yes', provided his belief that he has the desire has been well formed. If he makes his plans before having exposed himself to a great deal of the work involved in learning an instrument, and thus before having to face the question of why he is so miserable when practicing, and why he keeps it up, then his belief might be quite well-justified. If so, then his belief does give him a reason to practice, despite the fact that he does not actually have the relevant desire, and so the action he has chosen to undergo will not, in fact, meet the description 'most likely to satisfy my desire to Ψ'. 'My desire to Ψ' is a non-referring term in this case, and so there is nothing for John's action to satisfy.

Common or garden variety examples of desires which are present without being believed to present, or absent while believed to be present, can be difficult to imagine, because we seem to be fairly reliable at discerning what it is that we desire. John's case, in both instances, seems to be fairly unusual. Nonetheless, if the two examples presented do the work I claim they do, then we have some reason to think that a belief/desire pair is neither necessary nor sufficient for having a justifying reason for action. Whether or not one has a justifying reason for action, I think, is not dependent upon what desires one has, but upon what desires one is reasonable in thinking that one has. If John is presented with evidence that the only explanation for his going to his preferred newsstand is that he wants to look at his reflection, and if John forms the belief that things are as the evidence tells him, then this belief might actually provide him with a reason for going there each day. If he is presented with evidence that he cannot really want to be a musician, and that

the only explanation for his pursuing a career in music is his desire not to disappoint his mother, and if John forms the belief that things are as the evidence tells him, then this belief might give him a reason for action. The supposition that it is reasonable beliefs *about* desires which contribute to one's justifying reasons, rather than desires themselves, explains the conditions under which John does, and the conditions under which he does not, have a justifying reason to act as he does.

4.1.4. RELIEVING THE TENSION

The view of the role that desires play in practical reasoning that I have just argued for can seem to be in tension with my desire-based account of the success-conditions of action.[97] In this section I consider two objections to my view which bring out this tension.

The first objection asks what work there is left for desire to do on my view. If I don't want to commit myself to a Humean view of motivation (see Chapter 3), and I don't want to commit myself to a Humean view of reasons, why adopt the desire-based account of the function of action at all? This objection can be put in a more precise – and perhaps even more troubling – way. One of the major advantages to which desire-based accounts of reasons can lay claim is that they can provide a plausible account of how it is that we can be motivated by our reasons. Here, of course, we return to an issue which was raised in the last two chapters: a theory of rationality needs to explain how it is that we can be brought to care about theoretical and practical norms, how we can be moved to revise our beliefs and intentions in accordance with them. If to have a justifying reason to Φ is just to have a desire to Ψ plus the belief that by Φ-ing one will Ψ, then we can

[97] My thanks to Greg Scherkoske for raising this objection, and for using the understated phrasing I have adopted here, of a 'tension' in my view.

explain the motivational force of this practical reason in terms of the motivating force of the desire to Ψ. However, on my view, there need be no such desire for a reason to be present, and so this move is not open to me; I cannot explain the motivational force of one's practical reasons in terms of desires that don't exist.

I think, however, that there is something I can appeal to in order to explain the motivational force of practical reasons: the disposition to do what one believes will best satisfy one's desires. On my view, reasoning about what one should do is reasoning about which, of the available actions, will best satisfy one's desires. If I conclude that Φ-ing is the action which satisfies this description, then I will be disposed to Φ, and this disposition is a rational disposition, just as the disposition to believe those propositions that I think are most likely to be true is a rational disposition. It seems perfectly reasonable to allow that the move from the belief

> Φ-ing, is, of the actions available to me in my circumstances, the action that will most likely satisfy my desires

to the act of Φ-ing – or at least the intention to Φ – is a rational one even if this belief is false, provided I formed it rationally. The most common instance of this kind of failure will probably be the sort of case where my belief about how the action relates to the satisfaction of certain desires is false: I might miscalculate, and conclude that Φ-ing is most likely to satisfy my desire *D*, when in fact Ψ-ing is much more likely to do so. However, my view allows the move to be rational when (1) is false for a different reason, namely that I do not, in fact, have the desire I think I do. I might be exactly right in my calculation that Φ-ing is most likely to satisfy *D*, but be mistaken in thinking that *D* is one of my desires.

The key point to be kept in mind when considering this argument is that the disposition in question is not simply a disposition to do what will best satisfy one's desires. It is not plausible that rational beings have such a disposition, since we are often ignorant about what will best satisfy our desires. In such cases of ignorance, performing the action that will best satisfy one's desires may actually be *ir*rational: if I don't have a rational belief that Φ-ing will satisfy my desires, then I don't have a reason to Φ. Rather, a rational disposition to have is the disposition to do what one rationally *believes* will satisfy one's desires; the rational disposition to have is one which operates on beliefs rather than desires. Crucially, the belief on which the disposition operates can be false in one of two ways: either the means/ends belief is false, or the belief about one's desires is false. My view about the role of desire in practical reason can explain the motivational force of one's reasons in terms of this disposition: to have a reason to Φ is to have a reasonable belief that Φ-ing will best satisfy one's desires, and rational agents are disposed to do what they rationally believe will best satisfy their desires.

It might be objected that this move – to posit a disposition to do what one believes will best satisfy one's desires – threatens to be *ad hoc*.[98] After all, a similar move can be made in the context of *any* account of the goodness-conditions of action: for any such condition *c*, can't we explain the motivational force of practical norms – norms which would be based on *c* – by positing a disposition to do those actions which satisfy *c*? It is true that we can make this move with any account of the goodness-conditions of action, but I think the disposition I want to posit has some claim to be more than an *ad hoc* element in my theory.

[98] Both Greg Scherkoske and Duncan MacIntosh have raised this objection.

My argument for this claim is derived from some of James Dreier's recent work on means-end reasoning. He argues that the 'means-end rule':

> (M/E) "If you desire to Ψ and believe that by Φ-ing, you will Ψ, then you have a reason to Φ."[99]

has the status of a categorical imperative. Of course, on my view this principle is not quite right as stated, since it qualifies the 'means' component by reference to the agent's beliefs – you don't have a reason to do what will help you to Ψ, but only what you *believe* will help you to Ψ – but does not so qualify the 'ends' component – you have a reason to what you believe will help you to Ψ if you have a desire to Ψ, regardless of whether or not you *believe* that you have this desire, and of whether this belief is rational. It is a quick fix, however, to develop a modified means-end rule:

> (M/E*) If you rationally believe that you desire to Ψ and rationally believe that by Φ-ing, you will Ψ, then you have a reason to Φ.

Although this rule is slightly different than the one that Dreier defends, his argument that (M/E) counts as a categorical imperative still applies to (M/E*). He asks us to imagine someone named Ann who doesn't reason in accordance with the rule:

> "We tell her that she ought to take a prep course for the LSATs. She asks why. We point out that she wants to raise her chances of getting into a competitive law school, and she can raise her chances by taking the prep course. She admits as much, but she still isn't motivated to take the prep course. So we cite a rule, [(M/E*)]. Now suppose that Ann agrees that this rule does indeed instruct her to take the prep course (or at least it tells her that she has some reason to take it), given what she believes and desires, but she shrugs and doesn't accept the rule."[100]

---

[99] Dreier (2001), p.38.
[100] Ibid., pp.38-39.

I think Dreier is right when he says that we "must now conclude that there is something wrong with Ann."[101] She doesn't seem to be responsive, in this case, to basic means-end reasoning. This lack of responsiveness on Ann's part makes itself most apparent when we ask ourselves what it is that we need to add to her mental economy in order to help her reach the right conclusion. We can represent her mental economy like this:

(1) I believe that I desire to Ψ.

(2) I believe that by Φ-ing I will Ψ.

(3) (M/E*) says that if (1) and (2) are true, then I have a reason to Φ.

Ann fails to draw the conclusion that she has reason to Φ, so what is she missing? Dreier considers the possibility of adding a desire to satisfy (M/E*), plus the relevant means/end belief:

(4) I believe that I desire to satisfy (M/E*).

(5) I believe that by Φ-ing I will satisfy (M/E*).

The problem, of course, is that (4) and (5) are analogous to (1) and (2), respectively. Ann's whole problem is that she seems to be unable to connect her beliefs about her desires and her means/ends beliefs in order to motivate action. That is, if she is unmoved by (3), then there is no reason to expect her to be moved by (6):

(6) (M/E*) says that if (4) and (5) are true, then I ought to Φ.

If that is so, then whatever it is that Ann lacks, it cannot simply be a desire to satisfy (M/E*). As Dreier puts it,

> "Once you accept the (M/E) rule, what you need to get you to accept other rules is one or another desire. But no desire will get you to accept the (M/E) rule itself. (Compare modus ponens. Once you have modus ponens, what you

[101] Ibid., p.39.

need to get you to accept other rules is a belief in some conditional. But a belief in a conditional won't get you modus ponens itself.)"[102]

Dreier's argument concerns the explicit acceptance of inference rules, but I think the point holds as well in the context of dispositions. To see this, we can simply modify the example, so that Ann accepts:

(1) I believe that I desire to Ψ.

(2) I believe that by Φ-ing I will Ψ.

but has no motivation to Φ. As before, I think we have reason to conclude that there is something wrong with Ann, but there at least two possible answers to the question 'What is wrong her?' – and which answer we choose depends on how we flesh out the example. Suppose, first, that we are dealing with an isolated failure to be motivated to satisfy the desires that she believes she has: Ann is generally moved by considerations about her desires and what actions will satisfy them, but in this case she isn't. We can always explain particular failures of this sort without describing Ann as being *generally* irrational. For instance, she might be depressed, and her depression saps her beliefs of the motivational force that they would ordinarily have for her.

However, if Ann consistently demonstrates a failure to be motivated by beliefs sharing the schemas of (1) and (2), then I think we have reason to question whether she is actually a rational agent. To see why, consider the following, Davidsonian thought-experiment. Suppose you are in the position of a radical interpreter, trying to interpret the behavior of what seems to you to be a rational being. You attribute to this being a set of beliefs like the following:

I have desire D_1 By Φ-ing, I will satisfy D_1

[102] Ibid., p.41.

I have desire D_2 By Ψ-ing, I will satisfy D_2

I have desire D_3 By Ω-ing, I will satisfy D_3

The question is, what sorts of behavior should undermine your confidence in this attribution? As I said, individual failures to be motivated to act so as to satisfy the desires that one believes one has will not undermine your confidence, since we know from our own experience that motivation can fail to occur even when we sincerely hold beliefs of these kinds, as a result of fear, depression, and the like. However, if the being you are interpreting demonstrates a general failure to be motivated in the way you might expect – that is, motivated to Φ, Ψ, Ω, and so on – then I think you ought to revise your interpretation. If its behavior cannot be fitted into the framework of means-end reasoning, then you would have a hard time making sense of it. If it believes that it wants some cheese, and believes that there is some in the kitchen, then why hasn't it gotten up to get some? If it wants to be alert during the departmental meeting tomorrow, and believes that it can bring this about by going to bed early, then why hasn't it gone to bed? Individual failures of this sort of disposition can be explained while maintaining your original attribution of beliefs, but the more failures there are, the less plausible your attribution becomes. How can such a being hold the beliefs you have attributed to it, and yet systematically fail to act so as to satisfy the desires it believes itself to have? If this thought-experiment is convincing, then we have reason to think that the disposition to do what you think will best satisfy your desires is a disposition that is constitutive of rational agency. In order to be interpreted as a rational agent with both means-end-beliefs and beliefs about its desires, a being must not fail to be generally disposed to do what it believes will best satisfy its desires. If this is right, then the disposition is not simply an

ad hoc posit within my theory, but rather something which must be posited by *any* theory of rationality.

(It is important to note, for recall at the end of the chapter, that the same point holds for the disposition to believe what seems most likely to one to be true. If we attribute to a being a set of beliefs like the following:

p is the most likely proposition to be true

q is the most likely proposition to be true

r is the most likely proposition to be true

then our interpretation of this being can only withstand so many failures for the being to believe what it apparently thinks is most likely to be true. The occasional discordance – say the being appears to believe that *p* is the most likely proposition to be true, while apparently failing to believe *p* itself – can be explained without calling into question our interpretation. Just as there is such a thing as akratic action or motivation, so there is such a thing as akratic belief. But if the creature we are interpreting seems to fail *in general* to believe what it thinks is true, then we ought to revise our interpretation, because such a state would not make sense. The two dispositions I have discussed in this section, therefore, have a claim to being constitutive of rationality. They are dispositions that fall under the heading of calibrational rationality, as this term was introduced at the beginning of Section 4.1.)

The view I have just defended is a version of what Wallace calls 'meta-internalism'. Wallace distinguishes between 'internalism' and 'meta-internalism' as follows: both views adopt what he calls the 'empiricist' conception of motivation, according to which motivational states are not under our direct control, but rather are

states with regard to which we are passive. I cannot will myself into desiring to Φ except indirectly, perhaps by training myself to become the sort of person who would want to Φ, or by focusing my attention on certain facts which I think will help motivate me to Φ. The difference between the two views is how they account for what Wallace calls 'the motivation requirement' and 'the guidance condition' on practical reasons:

> *The Motivation Requirement:* "[I]f agent A has reason r to perform action x, and A is properly aware that r obtains, then A must be motivated to do x, on pain of irrationality."[103]
>
> *The Guidance Condition:* If agent A has reason r to perform action x, then A must be able to recognize and grasp r in "a way that directly gives rise to action."[104]

It can easily seem as though these two conditions are equivalent, but Wallace insists that there is a difference between them. The guidance condition ensures, he thinks, that reasons for action are properly connected to a view of rational beings as agents:

> "When we say that a given consideration is a reason for A to x, we are apparently suggesting that it is a requirement of reason that A be moved to do x. Rational requirements on action, however, are connected with our conception of persons as capable of deliberative agency. They are not merely classificatory norms that stipulate what it is for people to be good in some respect (defining an ideal of rational virtue, so to speak). They are norms of reasoning, which can be grasped and applied in a way that directly gives rise to action."[105]

That is, the motivation requirement only ensures that certain causal forces are at play when one has a reason, while the guidance condition is meant to ensure that agents can be affected by their reasons in a more reflective way. That is, they can reflect on their

---

[103] Wallace (1999), p.218.
[104] Ibid., p.219.
[105] Ibid., pp.218-219.

reasons and guide their own behavior in accordance with them, rather than simply being moved by them. Wallace thinks that this sort of agency – whereby one controls one's own behavior, rather than being subject to the push and pull of one's desires, goals, and the like – is required for rationality.

Internalism attempts to accommodate the motivation and guidance conditions by connecting an agent's reasons to her antecedent motivations. For instance, the view that one has a reason to Φ iff one has the appropriate belief/desire pair attempts to explain the motivational force of one's reasons in terms of the desire-component of the belief/desire pair. If an agent is already motivated to go to law school, then this motivation can explain how she can be motivated by her recognition that she has a reason to write the LSAT: her recognition of this reason just consists in (something along the lines of) her recognition that writing the LSAT is necessary condition for getting something she wants.[106] Meta-internalism does not explain the motivational force of reasons in this way. Rather, meta-internalist views posit "a disposition or desire, subjection to which is constitutive of our being rational or (alternatively) of our being agents, and that it is this abstract disposition or desire that ultimately both makes possible and explains action in accordance with our conception of our reasons."[107] My view has this characteristic. I do not explain the motivational force of reasons in terms of one's antecedent desires, since I allow that the desires to which an agent refers when she ascribes a reason to herself may actually be illusory; I can hardly explain an agent's motivation in terms of a mental state that she only believes herself to be in. Rather, I explain the motivational force of reasons in terms of a standing disposition to do what one believes will satisfy one's desires.

[106] For Wallace's objections to internalism, see ibid., pp.220-227.
[107] Ibid. p.227.

Wallace has recently argued that meta-internalism fails to explain how it is that we can be motivated by our reasons, because it fails to satisfy the guidance condition. His objection to meta-internalism stems from a thought-experiment: he asks us to imagine a teacher who believes that he ought to grade the stack of papers on his desk (here Wallace is concerned with the standing disposition to do what one believes one ought to do, but the point should still apply to my account[108]) but who instead decides to watch television. Wallace writes:

> "The distinctive claim of the meta-internalist is that the operations of practical reason can be traced back to the causal effects of the basal desire or tendency that makes us (rational) agents in the first place, such as the desire to do what we ought. In the case under consideration, however, this basal disposition is by definition not strong enough to motivate the agent to act rationally…If we interpret the basal disposition to do what one ought as an ordinary psychological state that competes with other states of the agent's desire for causal influence, then we must conclude that the agent's desire to watch the game was stronger than the disposition to do what they ought. It follows that, in a quite straightforward sense, the *akratic* agent was incapable of acting rationally under the circumstances. This in turn calls in question the important assumption that the agent really ought to have graded the papers in the first place. For how can it be required rationally that one do what one lacks the basic capacity to do? Doesn't ought imply, in this sense, can?"[109]

In this scenario, Wallace argues, the disposition to do what one ought to do – or at least what one believes one ought to do – gets overridden by the teacher's desire to watch television, or his desire not to grade papers, or what have you. But if that is true, then the teacher *cannot* grade the papers: the desire is stronger than the disposition, and so the disposition could not have outweighed the desire. This is especially true if we are meant

---

[108] The two dispositions are not necessarily identical, even on my account of the truth conditions of reasons-claims, since beliefs generate intensional contexts. Even if 'I rationally ought to Φ' is logically equivalent to 'Φ-ing is that action, of those available to me, which will best satisfy my desires', the dispositions to act on these beliefs will not be identical if the subject does not know or believe that the two thoughts are logically equivalent.

[109] Wallace (1999), p.232.

to assume a causal determinist view – which seems plausible, since Wallace goes on to claim that we can only accommodate both the motivation and guidance requirements by positing a faculty of volition, or 'the will'. If causal determinism is true, then if the course of events which led to the teacher's decision at time *t* makes it the case that the desire to avoid grading papers is stronger than the disposition to do what one believes one ought to do at *t*, then that course of events also determines the teacher's decision to watch television. In a straightforward sense, the teacher could not have done otherwise, and so Wallace claims that the meta-internalist cannot justify the claim that the teacher *ought* to have graded the papers.

I think that the best response available to the meta-internalist is to bite the bullet, and insist that 'ought' does not imply 'can' in the sense that Wallace thinks it does. To flesh out this response in full would require a detailed foray into the literature on freedom and responsibility; I will simply sketch the response here.

I will begin my response with two negative arguments designed to weaken the force of Wallace's objection by undermining the plausibility of his 'volitionist' alternative. As I suggested above, positing a faculty of volition of the kind that Wallace wants us to posit seems to commit us to a kind of indeterministic view of human action. He writes:

> "Agents who are equipped with what we might call a will – that is, with the capacity for self-determination independently of the desires and dispositions to which they are merely subject – can retain the capacity to comply with reasons that they knowingly act against. When they act *akractically*, for instance, it is not merely true of them that they would have done what they believe they ought had they been subject to a different configuration of desires and dispositions. It is true, more strongly, that they could have done what they ought, holding fixed the desires and the dispositions to which they were subject when they chose to do otherwise."[110]

[110] Ibid. p.238.

We are not meant to judge the teacher's action on the basis of nearby worlds whose causal history is similar to ours without being identical – worlds in which, for instance, the desire to avoid grading papers is not as strong as it is in the actual world. Rather, we are meant to keep the causal and psychological history of the teacher the same as in the actual world, while maintain that he could have done otherwise: these points clearly imply a kind of indeterminism. So, at time *t* when the teacher is deciding what to do, there are two open possible worlds indexed to time *t*+1: in one world, the teacher grades the papers, while in the other world he watches television.[111]

The first difficulty I have with this view is that it leaves us with a puzzle about the explanation of action. Remember that we are meant to be considering alternative possibilities which extend from the actual world – call it w_0 – which is indexed to time *t*, the time at which the teacher deliberates about what to do. Everything in the teacher's

[111] Wallace seems to think that the view developed in "Three Conceptions of Rational Agency" is consistent with his earlier view, developed in *Responsibility and the Moral Sentiments* (Wallace (1994)), which is a compatibilist view; or at the very least, he seems to think that the views are in the same spirit. He writes: "[The] capacity for self-determination is among the general powers of reflective self-control discussed in my book *Responsibility and the Moral Sentiments*. As such, it can assume a place in a compatibilist interpretation of agency and responsibility." (Wallace (1999), p.238, n.31.)

Unless, as seems perfectly possible, there is a subtle point in Wallace's work that I have failed to appreciate, I don't see how the volitionist conception can comfortably fit into a compatibilist framework. In *Responsibility and the Moral Sentiments* Wallace argues that "the conditions of responsibility do not involve freedom of the will, but primarily include the possession of certain rational powers: the power to grasp and apply moral reasons, and the power to control one's behavior by the light of such reasons. The 'can' that matters in moral responsibility is thus not the 'can' of alternative possibilities, or strong freedom of the will, but rather then 'can' of general rational power." (Wallace (1994), pp.7-8.) But in "Three Conceptions of Rational Agency" he seems to be concerned with a power to do otherwise than one did in one's circumstances, which seems to require alternative possibilities. Part of his argument against meta-internalism is that *akratic* actions could not have been avoided, and that this undermines the agent's responsibility and the authority of practical norms. But the only available alternative seems to be that the agent *could* have done otherwise, in a sense that requires alternative possibilities.

mental economy – his beliefs and the degrees of credence he gives to them, his desires and their relative strengths, etc. – is held fixed. From here, we are meant to posit two possible worlds – call them w_1 and w_2 – each indexed to time t+1, the time at which he acts. In w_1 the teacher grades the papers, while in w_2 he watches television. The question is, what explanation is available for why the teacher does what he does in each world? In w_1, what can we say by way of explanation for the teacher's decision to grade the papers? We cannot appeal to his grasp of the fact that he ought to grade the papers, since by hypothesis he also grasps this fact in w_2. The same argument applies to w_2: we cannot explain the teacher's decision to watch television by appealing to the strength of his desire, since by hypothesis he has the same desire with the same strength in w_1.

The problem becomes particularly salient if we keep in mind that volition is being described in terms of *powers* and *capabilities*. (See, for instance, the quotations in n.111.) Powers and capabilities are typically treated as *dispositions*, in the same way that meta-internalism treats the capability of being motivated by one's beliefs about one's reasons as a disposition.[112] But if that is so, then the faculty of volition should behave like any other disposition: that is, whenever it fails to be manifested or exercised at a time *t*, there are no possible worlds with a causal history identical to that of the actual world, but in which the disposition is manifested or exercised at *t*. This is, after all, the root of the problem Wallace raises for meta-internalism: if a disposition is outweighed by one of the agent's desires at *t*, then the course of action which would result from the disposition is a metaphysical impossibility at *t*. The teacher has the capacity to reflect on his reasons and act upon them at *t*, but if that capacity is not exercised, then there should be no possible

[112] For some recent discussions of powers and dispositions, see Mumford (1998) and Molnar (2003).

world with a causal history identical to that of the actual world, but in which the teacher exercises his volition.

However, Wallace seems to want to say that the action of grading the papers *is* metaphysically possible, and that leaves us with a problem when we try to explain the teacher's behavior in the alternative possible worlds. In w_0 the teacher has the disposition to control his behavior in light of his reasons, but this disposition does not get exercised in w_2, since it is overridden by the strength of his desire not to grade the papers. But by hypothesis, the strength of his desire is exactly the same in w_1, wherein he grades the papers. What explains the fact that the disposition gets exercised in one world, but is overridden in the other? The strength of the desire is the same in both worlds, the credence that the teacher attaches to his beliefs is the same in both worlds, and all of his dispositions are the same in both worlds. There seems to be nothing that the teacher can appeal to in order to explain his behavior in w_1, since anything he might appeal to will be exactly the same in w_2, wherein he fails to grade the papers; and the point holds vice-versa.

My second difficulty with Wallace's position is that it seems open to a parity-of-reasoning argument. His argument is that, unless our motivational states are under our control in some direct way, we cannot be held accountable for *akrasia* and the like, in which case practical norms lose their authority, since 'ought' implies 'can' in the sense that I have been discussing. But it seems we could run the same argument in the theoretical case to show that our beliefs must be under our control in some direct way. Consider the following example:

> A philosophy student has been doing a lot of reading in meta-ethics. After several years of reading various arguments for various positions, and assessing their merits, he comes to the conclusion that a moral error theory is the account of the semantics and metaphysics of morals which is most likely to be true: he concludes that a cognitivist view of moral semantics is more plausible than a non-cognitivist view, and that a metaphysical account which does not posit moral properties of any kind is more likely than an account which does posit such properties. However, although he holds the corresponding second-order belief, namely that he ought to believe the error theory, his first-order moral beliefs are not responsive to this second-order belief. Perhaps because of a deep emotional commitment to the truth of at least the majority of his moral opinions, he is unable to abandon those opinions in light of the error theory.

We may suppose that the student has the standing disposition to believe what he thinks he ought to believe, or what he thinks is most likely to be true, but that this disposition is overridden by other elements of his mental economy, such as his emotions. Now, if Wallace is right that 'ought' implies 'can' in the sense he insists upon, then a determinist view would lead us to conclude that the student is not being irrational in failing to believe the error theory. But Wallace's alternative account has very little plausibility in this case: it doesn't seem to be true that we have a faculty of volition through which we can directly motivate our beliefs. The student might re-read the arguments in favor of the error theory in the hopes that his emotions will eventually be overridden by his evidence, but he cannot simply will himself into believing the error theory. If we are going to account for

the authority of theoretical norms, then we are going to have to do it without supposing that an agent's beliefs are under his direct control. By parity of reasoning, if we are going to account for the authority of practical norms, then we are going to have to do it without supposing that an agent's motivational states are under his direct control.

Hopefully the arguments just presented should make Wallace's position seem less plausible: it poses difficulties when we try to explain an agent's behavior, and the reasoning which leads to it could also be used to defend the view that our beliefs are under our direct control, which is false.

What alternative is there to Wallace's volitionist view? How can we explain why the teacher is being irrational in choosing to watch television, if there was no possibility of him acting otherwise? I think Wallace is right to say that "conditions of responsibility should be understood in terms of general rational powers rather than in terms of freedom of the will,"[113] and that we need an approach which does not define the powers in question in such a way as to require alternative possibilities at the time of the agent's deliberation. The remainder of my response to Wallace consists in sketching such an approach.

The first thing to note is that the disposition we are concerned with is a quite general one: it is the disposition to do what one believes will best satisfy one's desires. This disposition covers a wide variety of cases, and a wide variety of particular beliefs. Just as an object can still be water-soluble – have the disposition to dissolve in water – even when it fails to dissolve in a particular case, so an agent can have the disposition to do what she believes will best satisfy her desires even if she fails to do so in a particular case. Why is this point relevant to the argument? Because it provides us with a sense of

[113] Wallace (1994), p.8.

'could' which might be able to ground the teacher's responsibility for doing something that, in Wallace's stronger sense of 'could', the teacher couldn't have avoided doing. If we are to interpret the teacher as a rational agent in the first place, then we must assume that he has the general disposition to do what he thinks he ought to do. That is, although it was physically impossible for the teacher, in his circumstances, to grade the papers, that is not because the general kind of activity that is 'doing what you believe you ought to do' was beyond his abilities. Presumably, if it is true that such a disposition is constitutive of rational agency, then he is capable of this kind of activity, and engages in it more often than not. Because he has this general disposition, he is the kind of person we would ordinarily expect to grade the papers.

Likewise, if I am right that the disposition to do what you believe will best satisfy your desires is a disposition constitutive of rational agency, then we can modify Wallace's example to make the same point. If the teacher believes that grading the papers is what will best satisfy his desires – perhaps because he wants to avoid getting fired – then we would expect him to grade the papers, since he is generally disposed to do what he thinks will best satisfy his desires. This is a fair expectation on our part, and the fairness of the expectation might be able to explain the fairness of criticizing him for watching television. We think it would be better for him to grade the papers, since that is what he believes will best satisfy his desires, and satisfying the agent's desires is the constitutive aim of action. If he remains unresponsive to his beliefs, then his action is not responsive to those facts which make grading the papers the act most likely to satisfy the constitutive aim. If the constitutive aim of action grounds the evaluative properties of action in the way that I have argued, then the act of watching television is a bad act. If the

teacher were not generally rational, then criticism might be unwarranted, since we could not have expected him to do otherwise. But since by hypothesis the teacher is generally rational, we can expect him to do what he thinks will best satisfy his desires, and hence we can expect him to grade the papers. His failure to do so is a failure to meet a fair expectation of ours, and hence warrants criticism whether or not he could have avoided the failure.

As I said, this response is sketchy, at best. The crucial point I am trying to motivate is that what is relevant to our judgments of responsibility is not simply whether it was possible for the agent to do otherwise than they did. A determinist view rules out such a possibility from the start. Rather, what is important is whether the *kind* of action the agent failed to perform is a *kind* of action that we ordinarily expect the agent to perform. Of course, any action can be described in a multitude of ways, and will fall under many different kinds, but I think we have a good enough of sense of what kinds are appropriate for normative judgments and what kinds are not. The kind 'Φ-ing because one believes one ought to, even though one has an overwhelming desire not to Φ' is inadmissible, since 'overwhelming' just means that the desire overrides the disposition; we could not reasonably expect this behavior from anyone. However, the kind 'Φ-ing because one believes one ought to, even though the agent has a strong desire not to Φ' is admissible, since we can reasonably expect people to do what they think they ought to do, even when they have a strong desire not to do it. In the teacher's case, it is not physically possible for him to grade the papers, but the action falls under a kind which (a) he is normally able to perform and (b) we can reasonably expect anyone who satisfies (a) to perform. Because the teacher has a general capacity to do what he thinks he should do

despite his contrary desires, he can be criticized for failing to do so in the case we are considering.

Of course, anyone sympathetic to Wallace's argument will not be satisfied with my reply. Even if my reply is completely filled out, it might be thought, it misses the point of the argument entirely. All I have done is to argue that the teacher is generally susceptible to the causal force of his beliefs about reasons; but since it was not possible for him to moved by the belief that he ought to grade to the papers, there was nothing he could do to get himself to grade the papers. He may generally be susceptible to the causal force of his beliefs about his reasons, but he cannot be criticized for any individual failure simply on that basis, since such failures are beyond his control. If he is going to be held responsible for not being motivated, then being motivated must be something that is under his control. Hopefully the arguments I presented against the volitionist view are sufficient to undermine my reader's confidence in the sorts of intuitions that Wallace appeals to in pressing this sort of argument. The volitionist view does a poor job of helping us explain an agent's behavior in the two open possible worlds at *t*, since there seems to be nothing to appeal to which would explain how the agent ended up in one world instead of another. Moreover, it seems to motivate a volitionist conception of belief which is manifestly implausible. By contrast, the view I have sketched seems to work well in the case of akratic belief. Consider the example I raised earlier, of the student who fails to adopt a moral error theory which he believes to be the most probable account of the semantics and metaphysics of morality. It may be physically impossible for the student to believe the error theory at the particular moment we are considering, but there is a sense in which he could do otherwise: namely, he has the general disposition to

believe what he thinks is most likely to be true. He is thus the sort of being we would expect to form the belief that the error theory is false, and his failure to do so is a rational failure on his part. He can be criticized for his failure, not because it was somehow within his power to believe otherwise than he in fact does, but because he is generally responsive to his reasons for belief.

4.2. THE IDEAL OBSERVER THEORY OF REASONS

On the view of rationality being developed here, a rational inference, whether practical or theoretical, is aimed at arriving at a true belief: in the theoretical case, it is aimed at arriving at a true belief about which, of the available beliefs about a certain issue, is most likely to be true; in the practical case, it is aimed at arriving at a true belief about which, of the available actions within a certain context, is most likely to satisfy one's desires. I have argued that this view does not commit us to an implausibly strong claim which might initially be thought to follow from my account of the goodness-conditions for beliefs and actions: namely, that one always has a reason to believe what is true, or to do what will best satisfy our desires. The reason that strong claim is implausible is that it seems quite clear that one can form a false belief or perform an action which doesn't satisfy one's desires – that is, one can 'get it wrong' in both theoretical and practical spheres – without thereby being irrational. What matters for the rationality of belief or action is not simply whether the agent has actually met the success-conditions for belief or action; rather, what matters is the process by which the belief or action was arrived at. The function account of the goodness of beliefs and desires can accommodate this point by construing the rationality of a belief or action as a function of the process by which it

is formed. If the process is sufficiently reliable in producing beliefs or actions which meet the relevant success-conditions, then any belief or action produced by this process will be rational.

Although it seems obvious that we do not always have reason to believe the truth or perform the action which will best satisfy our desires, there is a popular view in the literature on practical reason which comes close to claiming precisely this. I will call the account in question the Ideal Observer Theory – or simply IOT. The view is inspired by that presented by Bernard Williams in his "Internal and External Reasons", although it goes beyond Williams' view in certain ways.[114] In "Internal and External Reasons, Williams considers the example of a man who believes that the glass in front of him is filled with gin, and has the desire to mix himself a gin and tonic. He thus reasons his way to the action of mixing the contents of the glass with tonic and drinking it. In actual fact, the glass is filled with petrol, not gin, so that mixing it with tonic and drinking it will not actually satisfy the man's desire. Because of this fact, of which the man is himself wholly ignorant, Williams thinks that "it is just very odd to say that he has a reason to drink this stuff, and natural to say that he has no reason to drink it, although he thinks that he has."[115] He uses this idea to motivate the following restriction on when an element in an agent's subjective motivational set – which Williams calls '*S*' – provides the agent with an internal reason for action:

[114] Thanks to Greg Scherkoske for setting me straight on this particular issue.
[115] Williams (1979), p.78.

> A member of *S*, *D*, will not give *A* a reason for Φ-ing if either the existence of *D* is dependent on a false belief, or *A*'s belief in the relevance of Φ-ing to the satisfaction of *D* is false.[116]

The petrol case is probably intended to be an instance of the second kind of failure: the agent has a desire for a gin and tonic, and his belief that mixing the stuff in the glass with tonic will help him satisfy this desire is false. We can fairly easily think of cases corresponding to the first kind of failure: perhaps the agent's desire for a gin and tonic is based on the erroneous belief that it has less alcohol in it than a glass of beer. In such cases as these, Williams thinks, the agent does not in fact have a reason to mix the stuff in the glass with tonic and drink it, although we can provide an explanation of his action which "displays him as, relative to his false belief, acting rationally."[117]

As I indicated above, Williams is not the only person to adopt this sort of view of reasons for action. Kieran Setiya, who is otherwise admirably sensitive in his development of a virtue theory of rationality to the differences between what a virtuous person would do and what a non-perfectly virtuous person *ought* to do, still explicitly aims to accommodate Williams' argument when presenting his own view of practical reasons: "The fact that *p* is a reason for A to Φ just in case A has a collection of psychological states, C, such that the disposition to be moved to Φ by C-and-the-belief-that-*p* is a good disposition of practical thought, and C contains no false beliefs."[118]

---

[116] Ibid., p.79
[117] Ibid.
[118] Setiya (2007), p.12. He goes on to say "If the glass contains petrol, the fact that I am thirsty is no reason to drink from it, at all; there is no good reason to drink what is in the glass. The inclination to say otherwise turns on the fact that I have a collection of psychological states – including the belief that the glass contains water – such that the disposition to be moved to drink by them, together with the belief that I am thirsty, is a good disposition of practical thought. What the example shows is that *good practical*

Both Williams and Setiya think that a decision to Φ which is based on a false belief does not count as a decision which has been reached on the basis of genuine reasons. Michael Smith, whose view I will be most concerned with in this chapter, has presented the strongest view of this sort of which I am aware. In *The Moral Problem*, Smith argues that what one has reason to do depends on the actions or attitudes of an idealized version of oneself. He presents the view like this:

> "someone has a reason to Φ in circumstances C if and only if she would desire that she Φs in circumstances C if she were fully rational, where in order to be fully rational an agent must satisfy the following three conditions:
>
> (i) the agent must have no [relevant[119]] false beliefs
>
> (ii) the agent must have all relevant true beliefs
>
> (iii) the agent must deliberate correctly"[120]

There are at least two ways of developing this view of reasons, which Smith has distinguished: the Example Model and the Advice Model. He phrases the distinction in terms of possible worlds. On both views, what I have reason to do in the actual world depends on the judgment of a fully rational version of myself existing in a counterfactual world. However, they differ as to whether the fully rational agent's judgment is made regarding what *he* would do in the world he inhabits – Smith calls this the 'evaluating world': the world from which the judgment is being made – or whether it is made

---

thought corresponds to reasons only when it involves no false beliefs." (Ibid., emphasis added.)

[119] Smith just says that the agent must have *no* false beliefs, but as Greg Scherkoske has pointed out to me in conversation, this is a stronger condition than he needs. Surely, for instance, a false opinion on a difficult and abstract philosophical problem, such as how to accurately represent the logic of conditionals, or whether there is a coherent activity of quantifying over *everything*, need not impugn the agent's rationality if that belief is not relevant and plays no role in their decision.

[120] Smith (1994), p.156.

regarding what he thinks *I* should do in the world that *I* inhabit – the 'evaluated world'. Smith formulates the two models as follows:

> *The Advice Model:* "[T]he desirability of the agent's Φ-ing in the evaluated world depends on whether her fully rational self in the evaluating world would desire that she Φs in the evaluated world."[121]
>
> *The Example Model:* "[T]he desirability of the agent's Φ-ing in the evaluated world depends on whether her fully rational self in the evaluating world would desire to Φ *in the evaluating world*."[122]

I will deal with both of these views separately, but we should note the similarity: what one has reason to do depends on facts about a suitably idealized counterpart[123] - hence the title 'Ideal Observer Theory'.[124]

4.3. OBJECTIONS TO THE IDEAL OBSERVER THEORY

4.3.1. NO ANALOGUE IN THE THEORETICAL SPHERE

My first objection to the IOT begins from the recognition that, although it has been advocated as an account of the conditions under which one has reasons for action, it has no plausibility when construed as an account of when one has reasons for belief. I will begin by showing why I think the IOT doesn't apply to reasons for belief, and then argue

---

[121] Smith (1995), p.18.
[122] Ibid.
[123] I will be using the term 'counterpart' to denote the occupant of the evaluating world, but my use of this term should not be read as committing me to any particular view about identity across possible worlds. I use the term 'counterpart', not because anything I say will hinge on the acceptance of a counterpart theory of identity across worlds, but because it is simply more elegant than alternative ways of speaking.
[124] Astute readers will have noticed a switch from talk of 'what one has reason to do' to 'what it is desirable for one to do' between the two quotations I have given. As I will presently argue, this is not an innocent switch.

that if we don't accept the IOT as an account of reasons for belief then we ought not to accept it as an account of reasons for action, either.

My focus in this section will be on the first two of Smith's criteria for full rationality, that one have all relevant true beliefs and no relevant false beliefs. These criteria render the IOT implausible as account of reasons for belief.

Let us consider the Example Model first: it is simply not true that I have a reason to believe what a fully informed, correctly deliberating version of me would believe in my circumstances. Suppose, for instance, that I have a rational belief that propositions *p* and *q* are true, and I know that (*p* and *q*) implies *r*. The inference is an instance of modus ponens, so by hypothesis I am deliberating correctly if I derive *r* from (*p* and *q*). On the Example Model, I only have a reason to believe *r* – on the basis of following a modus ponens argument from (*p* and *q*) to *r* – if I would form that belief if I had full information at the time of the inference and also lacked any false beliefs. That is, we need to consider the activity of a fully informed version of me in a counterfactual world. We can easily describe a world in which *both* conditions on full information are met in such a way as to prevent my counterpart from drawing the inference that *r* is true:

> In the counterfactual world, I am privy to two pieces of information. First, I know that (*p* and *q*), when combined with other true propositions to which the actual me has no access, actually entails a contradiction. Therefore, I know that at least one of those conjuncts must be false. That is, the possession of all relevant true beliefs causes me to fail to draw the inference, since it also causes me to abandon a false belief, namely that both *p* and *q* are true. We can also suppose that some of the true beliefs which I only know in the

counterfactual world allow me to infer that it is *p* in particular which is false.

I thus do not have the false belief that *p* is true.

A fully informed version of me would not hold the premises that I actually hold, since he holds certain true beliefs which undermine those premises, and which I do not hold. On the Example Model, then, I do not have a reason to believe *r* in the actual world.

But it seems quite clear – as was stated in Section 4.1.2 – that I can have a reason to believe *r* even if I have derived it from false premises, provided that my belief in the premises is reasonable. The mere fact that *p* is false, for instance, does not in and of itself render my belief in *r* unreasonable. This point is clearest in cases of testimonial knowledge. Take, for example, my practice of deferring to the opinion of the scientific community when it comes to such questions as the truth of evolutionary theory, the existence of global warming, etc. If it turned out that the beliefs I formed were false, it would not follow that I had no reason to believe the propositions I derived from them. I would have every reason to believe those propositions, because I derived them from beliefs that I gained through a rational, reliable process, namely, deferring to the scientific community on scientific issues. If, in the case in question, my belief in *p* is based on deference to the scientific community, and my counterpart's true beliefs which show that *p* is false are beliefs about the correct scientific theory, then it seems that I have every reason to believe that *p* is true. To claim otherwise would be to allow the generation of uncountably many cases of supposedly unreasonable beliefs, which count as unreasonable only because the community deferred to made a mistake.

Moving on to the Advice Model, it is not immediately clear how to apply it in the theoretical case. The model is originally introduced in order to deal with cases where the

imperfect version of me has some behavioral disposition which the fully rational counterpart would lack: Because the model is concerned with dispositions to behave in certain ways, I think it is motivated more by Smith's third criterion than by the two with which I am concerned here. (I will say more about the third criterion in Section 4.3.3.)

However, I think we can construct test cases for the Advice Model based on what has already been said about the Example Model. Consider, first, the example where I believe that (*p* and *q*) is true, and I know that (*p* and *q*) implies *r*, so I conclude that *r* is true. My counterpart knows that (*p* and *q*) entails a contradiction, and that *p* is false. The question now is not what my counterpart believes, but what he wants me to believe.

It seems to me that there are two ways the IOT theorist could go here. First, he could maintain that my counterpart would want me to draw the conclusion that *r* is true. My counterpart knows that this inference will be based on a false belief, but he recognizes (a) that if I don't reject either *p* or *q* then my belief-set will be quite blatantly inconsistent if I don't derive *r*, and (b) that from my place of ignorance I have no way to know that (*p* and *q*) implies a contradiction, and so no way to know that my belief-set would contain fewer falsehoods without it, let alone which of the conjuncts I ought to abandon. The problem with this proposal is that it is structurally unlike the applications of the Advice Model which are meant to apply in practical cases. Consider the petrol case: we are presumably meant to conclude that my counterpart would not want me to drink the contents of the glass, even though I might have every reason to think that it contains gin, and no reason to suspect that it contains petrol. In order for the Advice Model to apply in the theoretical sphere, we require a much stronger analogy between theoretical and practical cases.

The second proposal retains the analogy, and says that my counterpart would want me to refrain from drawing the conclusion *r* from my beliefs. Even supposing that this actually is what my counterpart would want, it seems to me clear that I nonetheless have no reason to fail to draw the conclusion *r*. By hypothesis, my belief that (*p* and *q*) is perfectly rationally formed, and I have no way of realizing that this belief actually implies a contradiction. If I have a rational belief, and I know that this belief logically implies *r*, then *prima facie* I have reason to believe that *r* is true. Of course, that proposition might itself have evidence against it which is available to me, and that might provide stronger reason for me to reject my belief that (*p* and *q*), but such considerations have not played a role in the example. If we assume that I have no such evidence, then it seems that I have no reason whatsoever to refrain from drawing the conclusion that *r* is true; failure to draw this conclusion would result in an obvious inconsistency in my belief-set, and that seems to be a paradigm case of *ir*rationality.

Likewise, consider the case where I have accepted the truth of a scientific theory, T, on the basis of deference to scientific community. Suppose my counterpart knows that T is false and that another theory, T+, is true. If I am currently examining the question of whether a proposition *p* is true, and I derive *p* from my belief in T – for the sake of simplicity, we can assume that I know that the inference is logically valid – do I have a reason to believe *p* or not, according to the Advice Model?

Again, we can go one of two ways. According to the first proposal, my counterpart recognizes that my belief in T is rationally held, and that failure to draw the conclusion that *p* would result in an obvious inconsistency. (I am currently attending to the question of whether *p* is true, so my failure to conclude *p* would not be like my failure

to believe the answers to questions that don't concern me and which never occurred to me.) On this basis, my counterpart forms the desire that I draw the conclusion *p*. As I argued above, this proposal is entirely disanalogous to the way in which the Advice Model is designed to operate in practical cases, and so has little motivation if we adopt the standard method of applying the model to practical questions. According to the second proposal, which retains the analogy, I do not have a reason to believe *p*, since my counterpart would not want me to draw conclusions from false premises. Again, this proposal seems to me to give the entirely wrong answer. My belief in T is formed on the basis of deference to the scientific community, which seems to be a perfectly rational belief-forming process. Or at the very least, if this isn't a rational process, it is not irrational simply because the scientific community might make a mistake. I am not a scientist, and hence I am in no position to investigate the question of which scientific theory is best in anything approaching a reliable way. The scientific community will be far more reliable at coming up with the right answer than I will, and so it is perfectly rational for me to accept their conclusions. But if my belief in T is rational, and I know that *p* follows logically from T, then if I am actually considering the question whether *p* is true – if I am actually trying to figure out the answer to this question – then it seems I have every reason to believe that *p*, and a good reason not to refrain from forming this belief, since doing so would generate an obvious inconsistency in my belief-set.

If what I have argued so far in this section is correct, then the IOT does not apply in the theoretical sphere. It is not true that I have a reason to believe *p* only if my fully rational counterpart would either believe *p* in my situation, or want me to believe it. I can have a reason to believe *p* even when neither of those conditions is met, and the reason

for this on which I have focused in this section is that it is not simply the truth or falsity of one's premises which determine whether one has a reason to believe a conclusion derived from them. Rather, what matters is how those premises were arrived at.

This point about the relevance of the rationality of one's premises to the rationality of one's conclusion, if taken to heart, should motivate a re-examination of Williams' petrol case. On reflection, the case seems to be under-described; we haven't been given enough information about the circumstances, and about precisely how the belief that there is gin in the glass was formed. Suppose we imagine that Jim, fresh from his return from a trip to the Amazon – and an unpleasant confrontation with a solider named Pedro – walks over to his liquor cabinet with the intention of mixing himself a gin and tonic. Unbeknownst to him, a burglar entered his home while he was away and replaced the gin with petrol. Under these circumstances, Jim has every reason to believe that he is pouring gin, and not petrol. In fact, it is difficult to think of a *rational* process that would lead him to the belief that he is pouring himself a glass of petrol. Jim might actually believe that – despite the lack of any trace – it is highly likely that someone broke into his house while he was away and replaced his gin with petrol, but such a line of thought is indicative of paranoia, not rationality. Since his belief was rationally formed, Jim has a reason to drink what's in the glass, even though doing so will not, in fact, satisfy his desire. Suppose, however, that we imagine that Sherlock Holmes is in the process of confronting Moriarty, and the villain agrees to go quietly, although he would like to offer the detective a gin and tonic before they leave. In this situation, it seems that Holmes would not be rational in forming the belief that there is actually gin in the glass – he might not be rational in forming the belief that it contains specifically *petrol*, but it

seems reasonable for him to assume that Moriarty is not planning on giving him a refreshing drink – and so he has no reason to drink from it.

4.3.2. PLANNING

My second objection to the IOT is derived from the recent literature on planning in practical reason. There is, of course, some contention as to how exactly plans figure in practical reasoning, but hopefully what I say in this section will be sufficiently uncontroversial. To anticipate the conclusion: I will argue that it can be rational to Φ because one has formed a plan to Φ, even where this plan was formed – as plans often are – on the basis of incomplete information, and one would have formed a different plan, perhaps a plan to Ψ, had one been fully informed. That is, even when my counterpart would Ψ, or would advise me to Ψ, it can still be true that I have a reason to Φ, because I formed a plan to Φ which has certain features.

Before setting out my argument, I should say a little bit about what I take plans to involve. They are much like intentions, but the word 'plan' usually applies to something relatively long term. Michael Bratman points out two features which distinguish plans from intentions:[125]

(1) Plans are usually partial, in the sense that they do not commit one to all points of detail which would be required for a 'complete' plan. On this point, Bratman writes: "Suppose I decide this morning to go to a concert tonight. I do not settle all at once on a complex plan for the evening. Rather, I decide now to go to a concert, and leave till later deliberation about which concert to go to, how to get tickets, how to get to the concert in

[125] Bratman (1987), p.29.

ways consistent with my other plans, and what to do during intermission."[126] This sort of incompleteness is typical of the plans that we form in our everyday lives. We do not settle on every point of detail when we form our plans. Rather, we choose a goal, and figure out how to achieve it as we go. Some plans will be more detailed than others, of course. I may settle on having a paper written by a certain date, and set dates for the completion of all the intermediate steps, but leave open exactly how to achieve those steps (whose arguments I have the time and space to deal with adequately, how much time per day to leave aside for editing, and if things aren't going well, whether I can afford to stay awake all night to finish). But there will still be some degree of incompleteness involved in any plan I come up with.

(2) Plans have a hierarchical structure. What this means is that smaller, short-term plans are embedded within long-term plans. We can hold fixed certain plans while re-evaluating others. For example, I may plan to get my PhD, but re-evaluate my plans as to whether I will do so immediately after getting my master's degree, or take a year or two off. Once a plan is settled on with respect to that issue, I can re-evaluate my plans as to which universities I want to attend, and this change in plans can result in the need to form new plans regarding my applications to those universities.

There are several advantages to incorporating plans into one's practical reasoning. Settling on a plan allows one to put an end to a potentially endless process of deliberation. One can always gather more evidence, ask more people for advice, re-evaluate the strength of one's desires, etc. If we never settled on a plan, and were always deliberating, we would be far less likely to get many things done. For instance, I can't deliberate forever about what schools to apply to, because I will miss the deadlines; if I'm

[126] Ibid., p.29.

always re-evaluating my career path, that activity will leave less time and energy to actually do my job well. Given that the function of action is to satisfy our desires, it is rational to incorporate plans into one's decision-making policies. Given the limitations of time and cognitive capacities we all have to endure, it can be rational to form a plan on the basis of inconclusive reasoning and stick to it. If one forms one's plans in a rational way – not deciding what to do on a whim, or on *very little* evidence – then a policy of plan formation and follow-through will be a reliable way of satisfying one's desires.

Moreover, settling on a plan has the advantage that we choose a goal while we are in fairly leisurely circumstances, rather than putting ourselves in the position of having to decide what to do later, when circumstances will not allow the requisite time or perhaps the requisite lack of stress. If I form a plan to Φ when time *t* comes around, then as long as the circumstances at *t* are basically as I imagined them to be when I formed the plan, I don't have to deliberate at *t*, when the time available to me may be much shorter than is required to make a well-reasoned decision.[127]

Crucially, in order for plans to be effective, they must be fairly resilient. It is no good to form a plan if I am just going to re-open the question of what to do later. Part of the point of a plan, as I have said, is that it allows us to put an end to a process of deliberation; if we re-open the question too easily, then we have not effectively put an end to our deliberation. Of course, as Richard Holton points out, this resilience is not indefeasible: the point is that we shouldn't re-evaluate our plans too easily, not that we shouldn't re-evaluate them at all. As Holton puts it,

> "sometimes things will change so radically from what was expected that it will be rational to reconsider the intention. However, provided things do not change radically, it will be rational to go ahead with the intention without

[127] Bratman (1983), p.204; Bratman (1987), pp.29-30.

> reconsidering…[B]y and large, not reconsidering is beneficial. It enables economy of effort (I consider once, and then do not waste scarce time and effort in further consideration); and it provides coordination advantages (having fixed an intention, my other actions and the actions of others can be coordinated around it)."[128]

For example, getting an 'A' rather than an 'A+' on a small assignment is not a reason for a student to re-evaluate their plan of finishing their degree; discovering that I need to leave my house ten minutes earlier than expected in order to make it to the concert on time, and hence will have to make a small, quick meal rather than a large one, is not a reason to re-evaluate my plan of going to the concert. Only if the situation changes drastically in some way will it be rational for me to re-evaluate my plans.

Of course, there is at least one important feature that plans share with intentions: namely, that they inherit their rationality from the process by which they were formed. A long-term plan which is hastily formed, on the basis of very little information, despite the fact that the agent has plenty of time to consider what she wants to do, is not a rational plan, and such a plan should be more open to re-evaluating than a plan which was formed over a longer period of time, on the basis of a great deal of information.

What has all of this got to do with the IOT? Consider the Example Model first. The Example Model says that what I have reason to do in circumstances C is what my fully rational counterpart would do in C, where full rationality includes having full information. However, plans are made, for the most part, on the basis of *in*complete information; part of their whole point is that they allow us to settle on a course of action without having to constantly re-evaluate what we are going to do on the basis of incoming information. The decision to perform a certain action, where this decision is

[128] Holton (2004), p.515.

part of a plan, does not cease to be rational simply because it is made on the basis of incomplete information. Consider the following, borrowed, example:

> A meteor is on a collision course with Earth, and if its hits it will destroy the town of Springfield. The town has pulled together the money to fire a rocket at the meteor in order to destroy it before it can pose any immediate danger. Everyone decides, at time *t*, to remain in town until after the rocket has been fired at time *t*+1, reasoning that if the rocket misses its target, they can always escape via the only bridge out of town days before the collision. Due to a malfunction in the rocket's design, it does miss the meteor, and actually destroys the only bridge out of Springfield. Everyone is now trapped in the town.

If I am one of the residents of Springfield, it seems that I have every reason to decide, at *t*, to wait in town until after the rocket is fired, and only head for the bridge if the rocket misses its target. My fully informed counterpart, presumably, knows that the rocket will miss its target, since the question of whether it will miss is of course highly relevant to my decision. But the fact that I would know this if I was fully informed, and hence would head for the bridge at *t*, does not give me, as I actually am, in my state of ignorance, a reason to head for the bridge. What makes the situation I have described so funny is that no one could have reasonably predicted, not only that the rocket would miss, but that it would end up destroying the bridge.

Why, it will be asked, shouldn't we say that the residents of Springfield *do* have a reason to head for the bridge, but that they are unaware of this reason? The answer, I think, is that the processes by which the residents would reach the decision that their best

bet is to head for the bridge do not fall within the scope of what Goldman calls *ex ante* justification. That is, there is no evidence that these people currently have which could lead them to that conclusion if they applied any of their available reasoning processes. If I already have a certain well formed belief from which it is possible for me to reason via a simple logical inference to the proposition *p*, then it is certainly plausible to say that I already have a reason to believe *p*, even though I am not currently aware of it. By contrast, the residents of Springfield do not have any such beliefs, and in order to reach the conclusion that they should head for the bridge they would need to acquire new evidence. But if they would need to acquire new evidence to reach this conclusion, then they don't currently have a reason to head for the bridge.

Now consider the Advice Model. Presumably, what my fully-informed counterpart wants me to do is what he would advise me to do – hence the title, 'Advice Model'. Since he would advise me to head for the bridge (after all, he presumably doesn't want me to become trapped in a city that faces almost certain destruction), the Advice Model entails that I have reason to head for the bridge. But if the argument against the Example Model is right, then it should also apply to the Advice Model in this case. The fact that my fully-informed counterpart would advise me to head for the bridge does not entail that I have a reason to do so when I am in my state of ignorance.

The basic idea behind the counter-example can be used to generate many more. Plans have an important place in our practices of practical reasoning, and the activity of doing what I have planned to do does not fail to be rational simply because I would do something different if I were fully informed, or because my fully rational counterpart would want me to do something different. We are beings who often have to make plans

based on incomplete information, and who often have to remain resolute about our plans even if we might decide to change them if we had more information than we actually do. These activities, I have argued, are rational activities, and so a plan to Φ can give me a reason to Φ, regardless of what my counterpart has to say on the matter.

4.3.3. THE CORRECT DELIBERATION CRITERION

My first two arguments against the IOT focused on Smith's first two criteria of full rationality: that one possess all relevant true beliefs, and that one lack all relevant false beliefs. I have argued that if these two criteria are necessary conditions for fully rationality, then it is not true that the actions or attitudes of my fully rational counterpart determine whether or not I have a reason for action in my circumstances. It might be thought that even if my arguments work – and so the IOT cannot be correct if the first two criteria for full rationality are retained – we might still be able to adopt the IOT by restricting the definition of full rationality to Smith's third criterion: that one deliberates correctly. That is, we might claim that I have reason to Φ in my circumstances iff I would Φ if I were deliberating correctly. Regardless of what my beliefs are, or whether they are true or false, I have a reason to Φ only if there is a "sound deliberative route" from my beliefs to the decision to Φ.[129] In this section I will argue that even this version of the IOT is false.

As usual, consider the Example Model first. On this view, an agent has a reason to Φ in circumstances C iff her correctly deliberating counterpart Φs in C. In order to focus on the correct deliberation criterion, we need to keep in mind two restrictions on our description of the agent's counterpart:

[129] Williams (2001), p.91.

(1) There is no suggestion that the counterpart possesses any relevant true beliefs that the agent lacks, or that the counterpart lacks any relevant false beliefs that agent possesses. The difference between the agent and counterpart is not one of beliefs, but of how they go about deliberating in C. Therefore, we should stipulate that their belief-sets are identical: the counterpart has all and only those beliefs that the agent has. This first restriction implies a second…

(2) The counterpart cannot possess any general dispositions or patterns of deliberation that the agent lacks. Rather, any difference in deliberation in C – where the agent reasons incorrectly and the counterpart reasons correctly – can only be a difference in whether the general dispositions and patterns of deliberation that the agent and the counterpart share are active in C. This is because a difference in general patterns of reasoning suggests a difference in beliefs, and that is a difference we want to avoid positing. Consider an example from the theoretical sphere: if my counterpart has a tendency to infer according to the S5 principles while I do not, this might plausibly be thought to show that he believes such principles to be true while I do not.

By holding to these restrictions, we get a clearer picture of how an agent's reasons for action relate to the actions of her correctly deliberating counterpart. Consider the example just given: If I and my counterpart both believe that the S5 principles hold, then the Example Model has an explanation of why I have a reason to infer '$\Box p$' from '$\Diamond\Box p$' in C even if I do not in fact draw the inference: my counterpart reasons from the latter proposition to the former, and this is an instance of correct theoretical reasoning.

It seems the same sort of explanation would be adopted in the practical sphere: suppose that it is a principle of practical rationality that one rationally ought to do what one believes one morally ought to do. On this assumption, if I believe that it is morally required of me to give to charity, then I act irrationally if I don't give to charity. If I and my counterpart both hold this belief, then the Example Model has an explanation of why I have a reason to make the inference from 'I am morally required to give to charity' to the act of giving to charity: my counterpart makes this inference, and this is an instance of correct practical reasoning.

So, on the Example Model one has reason to believe *p* at time *t* iff one's counterpart, who has all the same beliefs but who reasons correctly at *t*, believes *p*. So, when considering a deduction, a subject has reason to believe the conclusion iff he would reason to the conclusion from those premises if he was deliberating correctly. These biconditionals are false, however, and for a reason which relates to a point I made earlier in this chapter: inference does not *generate* reasons, but only *transmits* them; the mere fact that one has reached one's conclusion via logically valid inference rules does not provide one with a reason to believe that conclusion. If one is to be justified in believing the conclusion *on the basis of one's deduction*, then one must have reason to believe the premises of the deduction. Of course, one might already have additional evidence for one's conclusion, so that one is not entirely without reason to believe it if one fails to have reason to believe the premises of the deduction. But in such a case, one's reasons are not a function of one's deductive inference: a deductive inference only provides reason for believing its conclusion when one has reason to believe its premises.

To see why this point is relevant to the Example Model, suppose that I am in a situation where I believe a proposition *p* for no good reason, and where *p* logically entails *q*. The Example Model says that I have a reason to believe *q* iff my correctly deliberating counterpart reasons his way from *p* to *q*, and – since the inference from *p* to *q* is by hypothesis a logical inference – we can be confident that my counterpart will reason his way from *p* to *q*, and therefore the Example Model says that in this case I have a reason to believe *q*. The problem is that even if there is a logically valid inference to be made from *p* to *q*, and my counterpart – who otherwise has all and only the beliefs that I do – draws it, that will not provide my counterpart with a reason to believe *q*, since he does not have a reason to believe *p*. But if my counterpart doesn't have a reason to believe *p*, then how can the fact that he comes to believe it in my circumstances provide me with a reason to believe it? Part of the intuitive appeal of the IOT, I think, is that the fully rational counterpart is plausibly thought to have a justified belief – or perhaps knowledge – in the proposition in question – the counterpart's behavior can legitimately guide mine, because he forms a justified belief, or perhaps knowledge; I ought to follow his lead because I am more likely to get the right answer by doing so. But if the counterpart holds his premises without good reason, then it is not true that his belief in *q* is justified or knowledgeable. If that is the case, then there is no reason to suppose that I ought to believe whatever my counterpart believes. What the Example Model amounts to in the theoretical sphere is: a subject has a reason to believe *p* iff there is a sound deliberative route from the subject's belief-set to the belief that *p*. But this claim, as I have argued, is false; one's sound inferences only provide reasons for believing their conclusions when the subject has reasons to believe their premises.

This argument also applies, I think, in the case of practical reasoning. Take the example given earlier, of the inference from 'I am morally required to give to charity' to the act of giving to charity, or the intention to give to charity. It seems to me that this move only provides one with a reason to give to charity if one's belief that it is morally required is rationally held. By hypothesis, the inference from 'Φ-ing is morally required' to the act of Φ-ing is a good one, and so my counterpart will make the inference. But the inference does not give the counterpart a reason to Φ if my moral belief is not rationally held, since his belief-set is identical to mine. If, for instance, I believe that giving to charity is morally required because I believe that it follows from a divine command theory that I do not have a good reason to adopt, then my moral belief is unjustified, and therefore cannot provide the counterpart with a reason for action. But again, if my counterpart has no reason to Φ, then the fact that he Φs cannot provide me with a reason to do the same.

It might be thought that we can avoid this objection by supposing, not just that the counterpart deliberates correctly at the time *t* – where *t* is the time-index for the proposition that the agent has a reason for action – but that the counterpart has *always* been deliberating correctly, and so even though her set of beliefs is identical to the agent's, the counterpart has a reason for everything she believes. But this is surely hopeless: I cannot justify my belief that *q*, which I have derived from my belief that *p*, on the grounds that I *would* have a reason to believe it if I had a reason to believe *p*, when in fact I have no such reason.

Turning to the Advice Model, I think the exact same reasoning applies. As before, we hold fixed the subject's mental economy while bestowing his counterpart with perfect

reasoning. Consider again the case where the subject is performing a logically valid deduction, but has no justification for the premises. Will the counterpart want the subject to believe the conclusion of the deduction, *p*? He holds all of the same premises that the subject does, and he knows that the conclusion follows by the deduction the subject is performing, so presumably he will want the subject to believe *p*. But in this case, the counterpart lacks a reason for believing *p*, since he lacks reason for believing the premises. If that is so, then why should the agent's reasons for action depend on the counterpart's desires? Part of what makes the Advice Model seem intuitive, I think, is the idea that a fully informed, correctly-deliberating counterpart's belief about what it would actually be best for the agent to do is reasonably held, or perhaps even a case of knowledge. Without Smith's first two criteria, however, this appearance completely vanishes: the counterpart has no reason to believe his premises, and so his desire for me to believe *p* – since this desire is presumably grounded in the fact that he has deduced *p* from his belief-set – cannot give me a reason to believe it.

4.3.4. A RESPONSE ON BEHALF OF THE IOT THEORIST 1: EXPLAINING THE AUTHORITY OF REASONS

I have been arguing that IOT does not give a correct account of when a subject has a reason for action. I have argued in two ways: first, by example – I have tried to present cases where intuitively the agent has a reason to Φ but the IOT says that they don't, or where intuitively the agent doesn't have a reason to Φ but the IOT says that they do; and second, by analogy with theoretical rationality – I have tried to show how and why the IOT fails to apply to reasons for belief, and motivate the idea that the same considerations show that it fails to apply to reasons for action.

However, there is something to be said on behalf of the IOT. In his defense of the Advice Model, Smith poses the following argument for its appeal: that adopting the Advice Model of reasons – together with Smith's account of what it means for one's counterpart to be fully rational – helps us to explain the normative authority of our beliefs about what reasons we have:

> "When I believe that it would be desirable to Φ in circumstances C, the internalism requirement tells us that my belief has the following content: that I would desire that I Φ in C if I were fully rational. [Note: This should be understood as saying that my *counterpart* desires that I Φ.] But now, if indeed I do believe this, and if I believe that I am in circumstances C, then surely the only rational thing for me to desire is to Φ. For a psychology that includes both the belief that I would desire that I Φ in C if I were fully rational – that is, the belief that I would have that desire if my desires formed a maximally coherent and unified set – *and* the desire that I Φ in C is itself a more coherent and unified psychology than one that includes the belief that I would desire to Φ in C if I were fully rational and yet *lacks* the desire to Φ in C. Coherence and unity are thus on the side of a *match* between the content of our evaluative beliefs and our desires."[130]

That is, when I believe that I have a reason to do something, I commit myself to the view that a counterpart of myself who is fully informed about all relevant information, and who deliberates correctly, would want me to do that thing. I cannot rationally fail, on this view, to have a desire to perform that action, since such a failure leads to a disconnect between the actions which I believe are the best ones for me to perform and the actions which I actually perform.

I think Smith is right that his analysis can help explain the normative authority of reasons, and this poses a problem for my view. I have been arguing against the IOT, but those arguments might be taken in stride by a supporter of the IOT, on the grounds that it is an explanatorily useful theory. However, I have argued in the last two chapters that the

[130] Smith (1995), p.36.

function account can explain the normative authority of epistemic and practical norms, and I think the explanation can be extended to the case of reasons.

To believe oneself to have a reason to believe *p*, on my view, is to believe that a belief in *p* satisfies, or would satisfy, one of the five principles I set out at the end of Section 4.1.2: it is to commit oneself to the view that a belief in *p* is the result of reliable processes of belief-formation. Of course, one can believe that both *p* and not-*p* are propositions which one could come to believe, in one's circumstances, via reliable belief-forming processes, and hence one can think one has a reason to believe *p and* a reason to believe not-*p*. For the purposes of this argument, however, we can leave aside such cases. Smith's example is one in which the subject has come to a definite answer as to what to do, so we can restrict ourselves to examples of this kind. In cases of very high levels of credence, what one believes when one believes that *p* is the proposition that ought to be believed, is that *p* is the proposition, of those which are competing in one's mind for credence, which is the most likely to be true. Since the aim of belief is truth, one would expect that rational beings would develop in such a way that their first-order beliefs are responsive to beliefs of this kind. A failure to form the belief that *p* is a failure to have one's first-order beliefs be governed by what one thinks is likely to be true, which is a failure to form one's first-order beliefs in such a way that they will likely (from the subject's point of view) satisfy the constitutive aim of belief. The same kind of argument can be made for practical reasons: a failure to act on one's beliefs about what one ought to do is a failure to act in such a way that one's actions will likely (from the subject's point of view) satisfy the constitutive aim of action. Since these constitutive aims are not simply ends or goals in the agent's cognitive economy, but instead are built into the

nature of what it is for an action or belief to be good *qua* action or belief, this kind of failure can rightly be considered a failure of rational-belief formation, or of rational action. Someone who fails the believe *p* when she thinks that *p* is the proposition most likely to be true – assuming that in addition to this *p* also meets the minimal evidential criterion for rational belief in her context of inquiry – demonstrates a failure of calibrational rationality, as this term was introduced in the beginning of Section 4.1. Likewise with someone who fails to do what seems most likely to satisfy her desires. The dispositions to believe and act in accordance with what one believes is most likely to be true, or to satisfy one's desires, are constitutive features of rationality.[131] If what I have said in this section is right, then their status as such – which I defended earlier, in Section 4.1.4 – can be explained by the functional account, without any appeal to an ideal observer.

4.3.5. A RESPONSE ON BEHALF OF THE IOT THEORIST 2: A DISANALOGY BETWEEN THEORETICAL AND PRACTICAL RATIONALITY

It might be objected that the IOT theorist can take many of the arguments made in this chapter in stride, by claiming that I have simply pointed out a disanalogy between theoretical and practical rationality: we cannot determine what a subject has reason to believe by considering what their fully rational counterpart would believe or want, but we *can* determine what a subject has reason to do by considering what their fully rational

[131] As I stated earlier, these dispositions only have to hold in general for one to count as rational. The occasional failure to believe what seems likely to be true, or to do what seems most likely to satisfy one's desires, can be attributed to *akrasia* or some such thing, without threatening one's status as a rational agent. However, if one fails sufficiently often to behave as these dispositions suggest that one would, then one's status as a rational agent comes into question. We cannot interpret a creature as a rational believer or agent unless it possesses the dispositions to believe what it thinks is true, and to do what it thinks will satisfy its desires.

counterpart would believe or want. I think, however, that this move will seem less plausible to anyone convinced by my re-evaluation of Williams' petrol case. That case *was* under-described before we started including information about how the agent's belief that there is gin in the glass was formed. In some cases such a belief is perfectly reasonable, and so is the resulting act of mixing the petrol with tonic and drinking it; since one has every reason to think one is mixing oneself a gin and tonic in such cases, one has every reason to drink the petrol. In other cases one will have reason to think that the glass is not filled with gin, and this will have the result that one would be unreasonable to drink its contents. The IOT theorist needs to explain why their evaluation of the petrol case, and of similar cases, withstands my analysis if she is going to maintain that there is a disanalogy between theoretical and practical reasoning.

CHAPTER 5: CONCLUSION

I have argued for a unified account of theoretical and practical reasoning. On my view, the rationality or irrationality of an inference leading to a belief or to an action is explained in terms of the rationality or irrationality of the corresponding inference to a second-order belief, or a belief about action, respectively. The inference to a belief that *p* is rational iff the inference to a second-order belief, that a belief that *p* is most likely to be true, is a rational inference; likewise, the inference to an action Φ, or to an intention to Φ, is rational iff the inference to the belief that Φ is the action most likely to satisfy one's desires, is a rational inference. The latter sort of inference, on my view, is explanatorily fundamental, because it makes explicit the role that the constitutive functions of belief and action play in determining the quality of beliefs and actions.

Because the explanatorily fundamental kind of reasoning involves making inferences between propositions, or between beliefs, on my view both theoretical and practical reasoning are cognitive matters. Non-cognitive attitudes like desires and goals do not give one reasons for action; only one's beliefs about one's desires and goals provide reasons for action. Because beliefs are cognitive states, practical reasoning can be represented, just like theoretical reasoning, as the drawing of inferential connections between propositions.

The account has been based on a functional theory – although an admittedly underdeveloped one – of evaluative properties. There can only be such a thing as a good belief or a good action, on this view, if the kinds 'belief' and 'action' incorporate functions into their essences. By determining what beliefs and actions are *for*, we can determine what makes for a good or bad belief or action. It is these functions which

determine the content of the propositions which serve as the conclusions of rational inferences: we want to know what proposition is most likely to be true, because the function of belief is the accurate representation of reality; we want to know what action will best satisfy our desires, because the function of action is to satisfy our desires. Likewise, the processes of belief-formation which we use in reasoning are evaluated in functional terms: a belief-forming method has the production of good (read: true) beliefs as its function, and so is evaluated by how reliably it fulfills this function.

The view I have offered here is, of course, only a sketch at this point, one that I hope to flesh out over the coming years. I would like to use the remaining pages to give the reader an idea of how I would now proceed to develop the theory, and what problems seem to be most salient. A warning in advance: the following pages are meant to provide nothing more than a sketch of how the project might be continued.

First of all, there is an issue about whether my account actually lands me in a skeptical view of practical reason. This worry takes two forms. The first form is that, by treating both theoretical and practical reasoning as involving inferences between propositions, there is nothing practical about so-called practical reasoning. If all it involves is figuring out which action will satisfy such-and-such a description, how can it be practical in its import? I have hinted at a solution in the last chapter: namely that what makes these inferences practical is that it is a condition on rational agency that one has a standing disposition to do the action which one believes satisfies the description 'most likely to satisfy one's desires'. We can distinguish between two kinds of rationality on this view:

Inferential Rationality: This consists in being able to make the right inferential moves of the sort I have been describing, and to properly form beliefs about what is most likely to be true, and what is most likely to satisfy one's desires.

Calibrational Rationality: This consists in having one's first-order beliefs, or one's actions and intentions, conform to the conclusions of one's inferential reasoning.

Being unable, in general, to reach the right conclusions about what belief or action is most likely to fulfill the constitutive aim of its kind, is a failure of inferential rationality. Being unable, in general, to believe what one thinks is most likely to be true, or to do what one thinks is most likely to satisfy one's desires, is a failure of calibrational rationality. One's practical inferences have practical import because it is a condition on rational agency that one generally does what one thinks will best fulfill the constitutive aim of action.

(This distinction allows me to accommodate the distinction between theoretical and practical reasoning drawn by Harman – namely that theoretical reasoning concerns the revision of beliefs while practical reasoning concerns the revision of intentions[132] – without having the distinction lead a view according to which theoretical reason is a cognitive matter while practical reason is a non-cognitive matter. The difference Harman points to can be accommodated as a difference between the attitudes which are relevant for calibrational rationality, while theoretical and practical reason both remain cognitive at the level of inferential rationality.)

[132] See Harman (1986), pp.1-2, p.113.

While each of the dispositions that I have posited as being conditions on rationality seem to me to have some plausibility, there are still some puzzles in the vicinity which would need to be worked out in order for us to have a complete account. For one thing, we need to know what the exact rational status is of such dispositions. David Owens suggests that dispositions of this sort are rational *independent* of whether one actually reached one's inferential conclusion for good reasons. He writes that three kinds of failure are possible in practical reasoning: (1) the agent is wrong about her interests are; (2) the agent does not deliberate correctly on the basis of true beliefs about her interests; and (3) the agent suffers from some form of *akrasia*, and does not do or intend to do what she thinks she should. He writes:

> "It is crucial to appreciate that [the third] element of irrationality is quite independent of any which might afflict the first two stages of practical deliberation. Suppose I am under some illusion about what I want, or else that I deliberate badly and judge wrongly what I should do to get it. There is an element of irrationality in all this. But if I then fail to do that which I have wrongly determined it would be best for me to do, I am guilty of an *extra* bit of irrationality: I lack the self-control to which every rational agent should aspire. My views about what I ought to do have a right to control my actions which is independent of the cogency of those views: my practical judgments have an *intrinsic authority* over what I do…"[133]

Of course, I would take issue with the claim that being mistaken about one's interests, or deliberating incorrectly, is necessarily a kind of irrationality, but those issues are not quite as pressing as how my view could accommodate this claim to intrinsic authority, or whether it even ought to. If we model calibrational rationality in the way that I have modeled conditionally reliable belief-forming methods – which would certainly seem the natural thing to do in the case of *theoretical* calibrational rationality, since it consists in forming certain first-order beliefs on the basis of certain second-order beliefs – then this

[133] Owens (2000), p.104.

intrinsic authority is hard to account for. On my view, conditionally reliable methods do not generate reasons, but only transmit them. If we suppose that one's second-order belief – call it *p* – was formed in such a way that the agent has no reason to believe *p*, then even if she does not suffer from theoretical *akrasia*, and hence forms the first-order belief – call it *q* – she will not have a reason to believe *q*. There were no reasons there to be transmitted. But if that is so, then how can it be irrational for her to fail to believe *q*? She would not have formed a reasonable belief anyway, so she has not really failed to gain anything in the way of rational beliefs. A similar problem might arise in the practical case: if one's belief about what action would best satisfy one's desires is not reasonably held, then one might not, on my view, have a reason to perform that action. In that case, how could *akrasia* be irrational?

Answering these questions would probably require an in-depth discussion of the relationship between rationality and reasons. We would need to ask, "When we say that a subject has attitudes that she is rationally required to have, does that entail that she has those attitudes for reasons?"[134] If this entailment doesn't hold, then there might be a way of treating calibrational rationality as independent of inferential rationality, since the worries about reasons-transmission would not undermine the claim that one is rationally required to have certain beliefs or perform certain actions, given one's other attitudes. The difficulty here is reconciling this approach with my own account of reasons, sketched in the last chapter, which seems to assume that rational methods *do* always provide one with reasons. By contrast, we could argue that calibrational rationality is not independent of inferential rationality. The dispositions in question are conditionally reliable, because they reliably get one to have beliefs which fulfill the aim of belief, or perform actions

[134] Smith (2007), p.279.

which fulfill the aim of action, provided one's beliefs about what will fulfill these aims are rationally formed. This might involve biting the bullet, and claiming that someone who holds an irrational belief about what is most likely to be true does *not* suffer from an additional form of irrationality if she fails, in that instance, to form the corresponding first-order belief.

The second form of the worry about practical reason is that my account seems to offer very little in the way of criticism of one's desires, goals, etc. I have not undertaken an explanation of how one's desires might be rationally ordered: do desires for certain things, like one's health or well-being, automatically take precedence over other desires? How should we understand prudence? That is, under what conditions is it rationally to give precedence to the desires one knows one will have in the future, but which one doesn't have right now? At the moment, I don't know quite how to answer these questions. The only obvious restriction that might be placed on desires and goals, given what I have said in this work, is that they be consistent, in the sense of being directed at non-contradictory propositions describing a way the world might be.[135] If I have an obviously inconsistent goal, then this goal can plausibly be thought to be irrational. One way we might actually demonstrate that conclusion within my account is by arguing along the following lines: If the proposition *p* is inconsistent, then there can be no action which brings it about that *p*. Therefore, a desire that *p* be the case is a desire that it is impossible to satisfy. Therefore, there is no action which can satisfy this desire, and *a fortiori* no good action which can satisfy this desire. I think one way to handle the rationality of irrationality of desires like this would be to provide an account according to

[135] For a clear, helpful discussion of consistency-reasoning in the practical sphere, see O'Neill (1989a).

which their rationality or irrationality is derived from the rationality or irrationality of the belief that an action might satisfy them. This seems to me to be promising, particularly because it allows that a goal or desire can be rational despite being directed at an inconsistent proposition, on the grounds that the agent is not irrational in thinking that the proposition is *not* inconsistent. For example, Lois Lane's goal of wooing Superman while avoiding all contact with Clark Kent – while possibly ill thought-out – is not proven irrational by the fact that she cannot possibly do what she wants to do. If she is fully rational in believing that Clark Kent and Superman are different people, then the fact that they are identical does not undermine the rationality of her goal.[136]

This brings us back to an issue which was raised in Chapter 2: namely the rationality or irrationality of believing contradictory propositions. As I suggested in Chapter 4, Owens' skepticism about the truth-aim theory's ability to accommodate the rationality of contradictory beliefs stems from his thinking that the truth-aim theory is committed to the view that beliefs are rational only if they are true. I argued that the truth-aim theory can be used to ground a reliabilist view of proper belief formation, and hence avoid this problem. I think that what I said in Chapter 2 can be used to solve another problem about rational beliefs, one which stems from the question of the normative role of logic. In his *Change in View*, Harman argues that the rules of deductive logic do not have "some sort of special relevance to the theory of reasoning."[137] One way he does this is by arguing against what he calls the Logical Closure Principle:

---

[136] Of course, it is questionable whether Lois *is* rational in holding this belief, since eyeglasses should by all rights be a terrible disguise. Readers sufficiently troubled by this aspect of the example – or who feel like taking sides in the Marvel/DC dispute – should feel free to substitute Mary-Jane Watson, Peter Parker, and Spiderman into it.
[137] Harman (1986), p.11.

Logical Closure Principle: "One's beliefs should be 'closed under logical implication.' In other words there is something wrong with one's beliefs if there is a proposition logically implied by them which one does not already believe. In that case one should either add the implied proposition to one's beliefs or give up one of the implying beliefs."[138]

Harman argues that this principle is false. One is not rationally required to believe everything which follows logically from one's antecedent beliefs, because many of these logical consequences will be entirely trivial. For instance, the belief that *p* logically implies all of the following:

$p \vee q$

$p \vee (q \vee r)$

$\neg(\neg p \wedge q)$

and so on, for any substitution instance of *q*, *r*, and the like. Surely we aren't required to form all of these beliefs on pain of irrationality. As Harman puts it, it would be "worse than pointless" to form these beliefs.[139] To capture this fact, Harman proposes a new principle, called 'Clutter Avoidance':

Clutter Avoidance: "One should not clutter one's mind with trivialities."[140]

This principle would be hard to explain on Owens' interpretation of the truth-aim theory. As long as *p* is true, one is guaranteed to form true beliefs by deriving its logical consequences, so why does the Clutter Avoidance principle hold if the aim of belief is truth? The answer is that the principle does not govern the formation of belief, after the process of inquiry has ended; rather, the principle governs inquiry. The logical

[138] Ibid., p.12.
[139] Ibid.
[140] Ibid.

consequences of any belief *p* are likely to be completely trivial, and therefore unhelpful in one's investigations. Because of this, it is irrational to spend one's time deriving these consequences, unless one has reason to think that some of these consequences *will* be helpful. For instance, if I want to know whether I should accept Leibniz's Law of the Non-Identity of Discernibles:

$$\Box(\forall x)(\forall y)(\forall X)((Xx \wedge \neg Xy) \supset x \neq y)$$

then it might be helpful for me to go about deriving the logical consequences of its negation, particularly by trying to derive a contradiction, which of course I can:

$$\Diamond(\exists x)(\exists y)(\exists X)((Xx \wedge \neg Xy) \wedge x = y)$$

The rationality of investigating all of the logical consequences of one's beliefs depends entirely on whether it is rational to think such consequences will be anything more than trivial. In some cases they are not trivial at all: they can actually be used in a *reductio ad absurdum* argument. But for the most part, the logical consequences of one's beliefs will pretty clearly be trivial, and one only wastes one's time and energy by deriving them.

Finally, I would like to briefly discuss an issue raised by my approach to the role of desire in practical reasoning. On my view, it is not necessary for one to have a desire *D* in order for one to have a reason to Φ, since one might have a reasonable belief that Φ-ing will satisfy one's desire *D*, even though one doesn't actually have the desire *D*. One can still be motivated to Φ, on this view, as long as one has the standing disposition to do what one believes will best satisfy one's desires. I argued at the end of Chapter 3 that the function account allows us to explain the rationality of this disposition *without* positing a desire to do what one believes will best satisfy one's desires. This set of claims puts me in direct opposition to Thomas Nagel, who writes:

> "The claim that a desire underlies every act is true…only in the sense that *whatever* may be the motivation for someone's intentional pursuit of a goal, it becomes in virtue of his pursuit *ipso facto* appropriate to ascribe to him a desire for that goal… That I have the appropriate desire simply *follows* from the fact that these considerations motivate me; if the likelihood that an act will promote my future happiness motivates me to perform it now, then it is appropriate to ascribe to me a desire for my own future happiness. But nothing follows about the role of the desire as a condition contributing to the motivational efficacy of those considerations. It is a necessary condition of their efficacy to be sure, but only a logically necessary condition. It is not necessary either as a contributing influence, or as a causal condition."[141]

On my view, it isn't always appropriate to ascribe to someone a desire whenever they are motivated to act. Nagel's claim can seem perfectly obviously true, however, so I take on a burden of showing it to be false. I have a few thoughts on how this might be pulled off:

First, we should note that there is a potential worry about epiphenomenalism in the vicinity. It seems that Nagel is making his point far too strongly in claiming that a desire is a 'logical' condition for the motivational efficacy of beliefs, since there is no logical system – at least none I am aware of – that has this consequence. The positing of a desire is far more substantive than the phrase 'logical condition' suggests; to posit a desire is to posit a distinct mental state, and this is no triviality. But if that is so, then we might wonder how we could possibly be justified in positing a mental state while giving it no causal work to do in an explanation of an agent's behavior. On Nagel's view, my belief 'I ought to Φ' can cause *both* my desire to Φ *and* my actually Φ-ing, without the desire having to do any causal work. But now the desire is a mere epiphenomenon – it is itself caused but does not cause anything in turn – and Nagel must explain why he is justified in positing it in the first place.

There might be a way of defending Nagel: there might be a way of explaining how a desire can be a necessary condition on motivation without itself being a cause. We

[141] Nagel (1970), pp.29-30.

might treat desires as functional states or dispositions, along the lines of Smith (1994).[142] This can be of help to Nagel, since there is a debate to be had over whether functional states or dispositions can feature in an informative causal explanation of anything. Stephen Mumford presents this objection quite forcefully, though without endorsing it, when he writes:

> "To be fragile means nothing more than to break if dropped; to be soluble means nothing more than to dissolve if immersed in water. A similar analysis can be given for any disposition: to say that something possesses a disposition is just to say that the appropriate response will follow upon the appropriate stimulus. But if solubility means just 'dissolves in water', then any explanation of why a substance dissolves in water in terms of it being soluble will be nothing more than a trivial analytic explanation, which is no explanation at all. If fragile means nothing more than 'breaks when dropped', then it is no explanation of why something breaks when dropped."[143]

To see how this works in Nagel's example, suppose I form the belief 'I ought to Φ', and this belief causes in me a desire to Φ, which is to say, a disposition to Φ in certain circumstances – call them *C* – such as those where I believe I ought to do so, and where the action is available to me. If the desire to Φ is just a disposition to Φ in *C*, then we cannot answer the question 'Why did I Φ when placed in circumstances *C*?' by responding that I had a desire to Φ.

There are several ways we could go in responding to this version of the argument. First, we could accept that desires are dispositions, and argue that dispositions can figure in genuine causal explanations.[144] Second, we could argue that desires are not simply dispositions or functional states, and argue that it is the distinguishing features of desires which allow them to play the particular causal role that they do, if any. This seems to me a promising strategy, since it would allow us to provide a more substantive account of

[142] See pp.111-116.
[143] Mumford (1998), p.134.
[144] This is Mumford's strategy. See ibid., ch. 6.

what desires are, one which helps us distinguish between the many different things that can cause a person's behavior, such as compulsions, reflexes, etc. If desires are not simply dispositions to behave in certain situations – both because there are other things which play this functional role, and because desires are distinguished from the other members of this set by certain features[145] - then the triviality objection rests on a false conception of desire. Thus, my account leaves some interesting work to be done in the philosophy of mind on the nature of desires and other mental states, but I think that even without having presented such work here, we have some reason to be skeptical of Nagel's claim that we can always posit a desire when an agent is motivated.

[145] The most likely such features would be phenomenological and epistemological: the 'feel' that many desires seem to have, and the first-person access that we seem to have to many of them (although not infallible access, if my arguments are sound, and one can mistaken about what one desires). Michael Smith explicitly sets himself up in opposition to these sorts of arguments (see Smith (1994), pp.104-116). For a reply, see Miller (2003), pp.277-279.

BIBLIOGRAPHY

Altham, J. E. J. (1986) "The Legacy of Emotivism," in MacDonald and Wright (1986), pp.275-288.

Brandom, Robert. (1998) "Action, Norms, and Practical Reasoning," in Millgram (2001), pp.465-480.

Bratman, Michael. (1983) "Taking Plans Seriously", in Millgram (2001), pp.203-220.

———————. (1987) *Intention, Plans, and Practical Reason*. Cambridge: Harvard University Press.

———————. (1991) "Cognitivism about Practical Reason", in Bratman (1999), pp.250-264.

———————. (1999) *Faces of Intention: Selected Essays on Intention and Agency*. New York: Cambridge University Press.

Cahn, Steven M. and Joram G. Haber. (1995) (eds.) *Twentieth Century Ethical Theory*. Upper-Saddle River: Prentice Hall.

Dancy, Jonathan. (2004) "Discussion – On Knowing What One is Doing", *Philosophical Studies* 121, pp.239-247.

Dreier, James. (2001) "Humean Doubts about Categorical Imperatives", in Millgram (2001), pp.27-48.

Field, Hartry. (2009) "What is the Normative Role of Logic?" *Proceedings of the Aristotelian Society Supplementary Volume,* 83. pp.251-268.

Gamut, L. T. F. (1991) *Logic, Language, and Meaning, Volume 2: Intensional Logic and Logical Grammar*. Chicago: The University of Chicago Press.

Geach, P. T. (1956) "Good and Evil", in Cahn and Haber (1995), pp.300-306.

Goldman, Alvin I. (1976) "What is Justified Belief?" in Sosa *et al*. (2008) pp.333-347.

Harman, Gilbert. (1986) *Change in View*. Cambridge: The MIT Press.

Holton, Richard. (2004) "Rational Resolve", *The Philosophical Review*, 113: 4, pp.507-535.

Lewis, David. (1986) *On the Plurality of Worlds*. Malden: Blackwell Publishing Ltd.

———————. (1988) "What Experience Teaches", in Lewis (1999), pp.262-290.

———————. (1999) *Papers in Metaphysics and Epistemology*. New York: Cambridge University Press.

MacDonald, Graham and Wright, Crispin. (1986) (eds.) *Fact, Science and Morality: Essays on A. J. Ayer's Language, Truth and Logic*. Oxford: Basil Blackwell.

McDowell, John. (1978) "Are Moral Requirements Hypothetical Imperatives?" in McDowell (1998), pp.77-94.

———————. (1998) *Mind, Value, and Reality*. Cambridge: Harvard University Press.

Miller, Alexander. (2003) *An Introduction to Contemporary Metaethics*. Malden: Polity Press.

Millgram, Elijah. (2001) (ed.) *Varieties of Practical Reasoning*. Cambridge: The MIT Press.

Molnar, George. (2003) *Powers: A Study in Metaphysics*. Oxford: Oxford University Press.

Mumford, Stephen. (1998) *Dispositions*. New York: Oxford University Press.

Nagel, Thomas. (1970) *The Possibility of Altruism*. Princeton: Princeton University Press.

O'Neill, Onora. (1989a) "Consistency in Action", in O'Neill (1989b), pp. 81-104. Reprinted in Millgram (2001), pp.301-330.

———————. (1989b) *Constructions of Reason: Explorations of Kant's Practical Philosophy.* New York: Cambridge University Press.

Owens, David. (2000) *Reason without Freedom: The Problem of Epistemic Normativity*. New York: Routledge.

———————. (2003) "Does Belief Have an Aim?" *Philosophical Studies* 115, pp.283-305.

Pettit, Phillip and Smith, Michael. (1990) "Backgrounding Desire," *Philosophical Review* 99, pp.565-592.

Railton, Peter. (1997) "On the Hypothetical and Non-Hypothetical in Reasoning about Belief and Action," in Railton (2003), pp.293-321.

———————. (2003) *Facts, Values and Norms: Essays Toward a Morality of Consequence*. New York: Cambridge University Press.

Schroeder, Mark. (2007) *Slaves of the Passions*. New York: Oxford University Press.

Setiya, Kieran. (2007) *Reasons without Rationalism*. Princeton: Princeton University Press.

Smith, Michael. (1994) *The Moral Problem*. Malden: Blackwell Publishing, Ltd.

———————. (1995) "Internal Reasons", in Smith (2004), pp.17-42.

———————. (2003) "Rational Capacities, or: How to Distinguish Recklessness, Weakness, and Compulsion", in Stroud and Tappolet (2003) pp.17-38.

———————. (2004) *Ethics and the A Priori: Selected Essays on Moral Psychology and Meta-Ethics*. New York: Cambridge University Press.

———————. (2007) "Is There a Nexus Between Reasons and Rationality?" in Tenenbaum (2007), pp.279-298.

Sosa, Ernest. (1999) "How to Defeat Opposition to Moore," *Philosophical Perspectives* 13, pp.141-153.

————————. (2007) *A Virtue Epistemology: Apt Belief and Reflective Knowledge, Volume 1*. New York: Oxford University Press.

Sosa, Ernest, Jaegwon Kim, Jeremy Fantl and Matthew McGrath. (2008) (eds.) *Epistemology: An Anthology*, *2nd Edition*. Malden: Blackwell Publishing, Inc.

Stalnaker, Robert. (2008) *Our Knowledge of the Internal World*. New York: Oxford University Press.

Stroud, Sarah and Tappolet, Christine. (2003) (eds.) *Weakness of Will and Practical Irrationality*.

Tenenbaum, Sergio. (2007) (ed.) *Moral Psychology* (*Poznań Studies in the Philosophy of the Sciences and the Humanities*, vol. 94). New York: Rodopi.

Thomson, Judith Jarvis. (2008) *Normativity*. Chicago: Open Court.

Velleman, David J. (1996) "The Possibility of Practical Reason," *Ethics* 106 (July 1996), pp.694-726.

———————. (2000) *The Possibility of Practical Reason*. Ann Arbor: Oxford University Press.

———————. (2004) "Replies to Discussion on *The Possibility of Practical Reason*," *Philosophical Studies*, 121, pp.277-298.

Wallace, R. Jay. (1994) *Responsibility and the Moral Sentiments*. Cambridge: Harvard University Press.

———————. (1999) "Three Conceptions of Rational Agency", *Ethical Theory and Moral Practice* 2:3, pp.217-242.

Watson, Gary. (1975) "Free Agency", in Watson (1982), pp.96-110.

———————. (1982) (ed.) *Free Will*. (1982) Oxford: Oxford University Press.

Williams, Bernard. (1973a) "Deciding to Believe," in Williams (1973b), pp.136-151.

———————. (1973b) *Problems of the Self*. Cambridge: Cambridge University Press.

———————. (1979) "Internal and External Reasons," in Millgram (2001), pp.77-89.

———————. (2001) "Postscript: Some Further Notes on Internal and External Reasons," in Millgram (2001), pp.91-97.

Zagzebski, Linda. (1994) "The Inescapability of Gettier Problems," in Sosa *et al*. (2008) pp.207-212.